ILTA
INTERNATIONAL LANGUAGE TESTING ASSOCIATION

สถาบันภาษา
LANGUAGE INSTITUTE
Chulalongkorn University

คณะครุศาสตร์
FACULTY OF EDUCATION
Chulalongkorn University

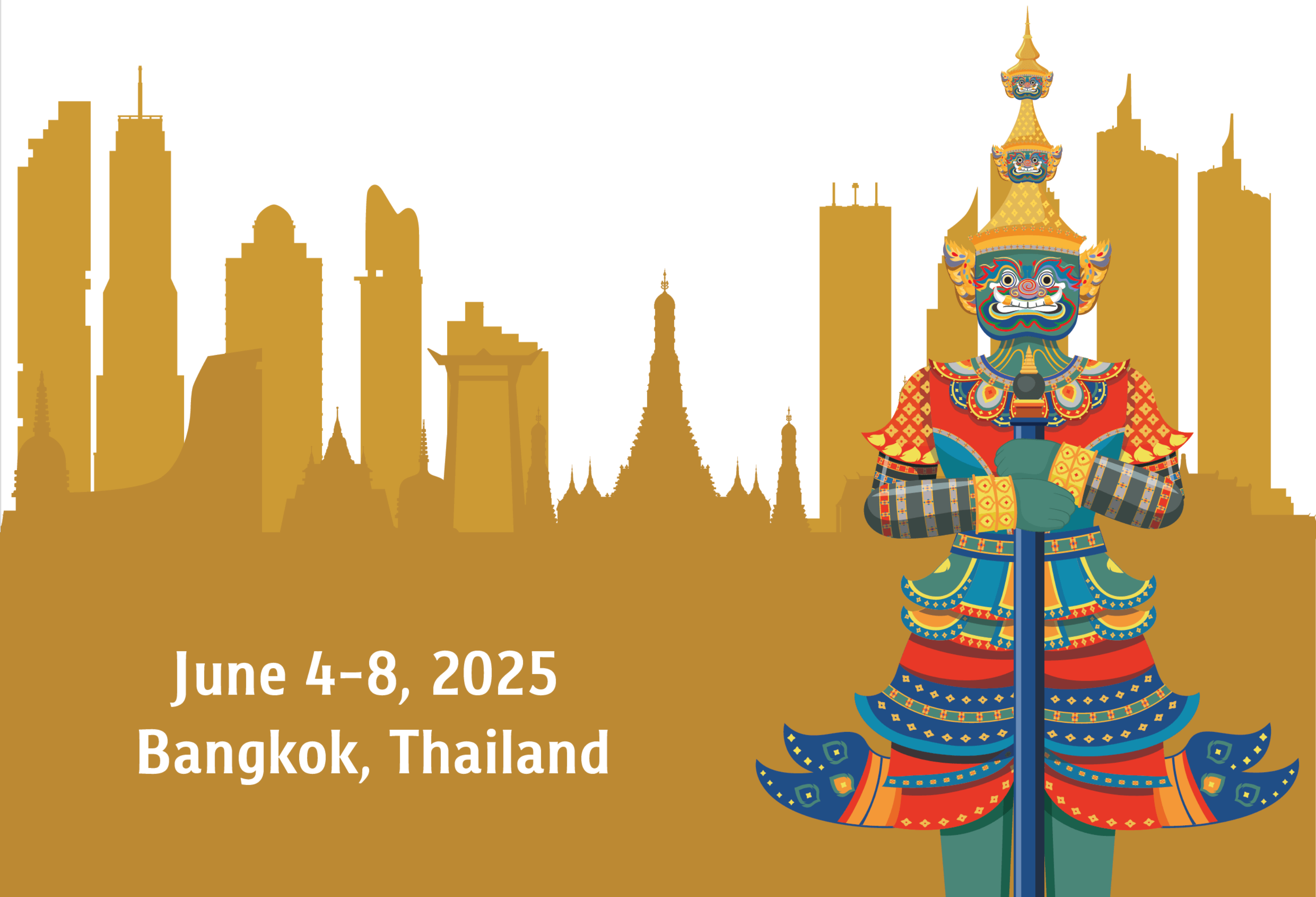
ENGLISH AS AN INTERNATIONAL LANGUAGE
Chulalongkorn University

INDIANA UNIVERSITY

LTRC 2025

46th Language Testing Research Colloquium

Language Assessment in Multicultural Contexts: West Meets East



June 4-8, 2025
Bangkok, Thailand

Pearson | PTE

The secure, accurate & trusted English test

CA, AU, NZ
ACCEPTED WORLDWIDE
UK & USA

Attend our sessions and join discussions

PTE, a global leader in language assessment, is trusted by governments, universities, and professional associations worldwide for fair and reliable testing.

Backed by cutting-edge research, we ensure accuracy with AI-driven scoring and expert verification, industry-leading security, and assessments tailored for academic success.

Learn more about our research
pearsonpte.com/research

Pearson LTRC Session

Alternative Approaches to Reading Item Difficulty Calibration: Perspectives from Text Complexity

Presenters: Dr Ying Zheng, University of Southampton and David Booth, Pearson.

Pearson LTRC Session

A study in contrasts: Investigating human and automated scorer evaluation of textual changes

Presenter: Sarah Hughes, Pearson.

Pearson LTRC Session

Further evidence around intelligibility as an aspect of the L2 speaking construct

Presenter: Dr William Bonk, Pearson.

PLATINUM SPONSORS

Chula
Chulalongkorn University

duolingo
english test

CAMBRIDGE

Pearson | PTE

IELTS

英検 Eiken Foundation of Japan

GOLD SPONSORS

LTTC GEPT BESTEP

GOETHE INSTITUT

BRITISH COUNCIL

SILVER SPONSORS

M UNIVERSITY OF MICHIGAN | MICHIGAN LANGUAGE ASSESSMENT

BRONZE SPONSORS

WIDA™
UNIVERSITY OF WISCONSIN–MADISON

Lancaster University

Conference Schedule Overview

Date: Wednesday, 04/June/2025		
8:00am - 9:00am	**Registration** Location: **Foyer in front of Poonsapaya** (Building 2, Third Floor)	
9:00am - 4:00pm	Location: **Ampai** (Building 1, Ground Floor) **Workshop A - Language Test Design and Alignment to the CEFR** Viphavee Vongpumivitch and Napat Jitpaisarnwattana	Location: **Phramingkwan** (Building 2, Second Floor) **Workshop B - Publishing Your Research in the AI Era: Sharing the Road Traveled with Budding English as a Foreign or Second Language Scholars**) Qin Xie

Date: Thursday, 05/June/2025				
8:00am - 9:00am	**Registration** Location: **Foyer in front of Poonsapaya** (Building 2, Third Floor)			
9:00am - 4:00pm	Location: **Ampai** Building 1, Ground Floor **Workshop A - Language Test Design and Alignment to the CEFR** Viphavee Vongpumivitch and Napat Jitpaisarnwattana	Location: **Room 405** Building 3, Fourth Floor **Workshop C - An Applied Introduction to No-code Regression and Machine Learning for Language Assessment** Vahid Aryadoust	Location: **Phramingkwan** Buiding 2, Second Floor **Workshop D - Impact Evaluation: Global Perspectives and Best Practices** Micheline Chalhoub-Deville, Hanan Khalifa, and Eunice Jang	Location: **Room 409** Building 3, Fourth Floor **ILTA Executive Meeting**

Date: Thursday, 05/June/2025	
3:30pm - 5:30pm	**Registration** Location: **Foyer in front of Poonsapaya** (Building 2, Third Floor)
4:00pm - 5:00pm	**Newcomers' Session** Location: **Ampai** (Building 1, Ground Floor)
5:00pm - 6:30pm	**Opening Symposium** Location: **Poonsapaya** (Building 2, Third Floor) **East Meets West: A Multifaceted Interaction of Constructs and Contexts in Language Testing and Assessment** ***Chair(s):*** Liying Cheng ***Discussant(s):*** Micheline Chalhoub-Deville *Presentations of the Symposium* **The Testing Circle: East to West and Back Again** Barry O'Sullivan **Context Validity of Language Frameworks: A Comparative Study of the CEFR and the CSE Oral Proficiency Scales** Yan Jin, Lin Zhang **A Historical Review of the Target Construct Reflected in the Design of the National University Entrance Examination in Japan** Yasuyo Sawaki **Consequences: Constructs in context** Liying Cheng **Impact Frameworks Evolution & Usage in Multicultural and Multilingual Contexts** Hanan Khalifa **Integrating Teaching, Learning, and Assessing: The SBA Approach** Antony Kunnan
6:30pm - 8:00pm	**Welcome Reception (Sponsored by Duolingo English Test)** Location: **Sumon** (Building 1, Ground Floor)

Date: Friday, 06/June/2025						
8:00am **-** **8:45am**	**Registration** Location: **Foyer in front of Poonsapaya** (Building 2, Third Floor)					
8:45am **-** **9:30am**	**Welcome and Opening Remarks** Location: **Poonsapaya** (Building 2, Third Floor)					
9:30am **-** **10:30am**	**Samuel J. Messick Memorial Lecture** Location: **Poonsapaya** (Building 2, Third Floor) **A Personal Odyssey through Language Assessment and the Fields of Validity** Professor Lyle F. Bachman, University of California, Los Angeles (UCLA)					
10:30am **-** **11:00am**	**Coffee Break (Sponsored by Chulalongkorn University)** Location: **Sumon** (Building 1, Ground Floor)					
11:00am **-** **12:30pm**	**Research Papers** Location: **Ampai** Building 1, Ground Floor **Promoting Multiculturalism in Arabic: A Pan-Arab Initiative for Language Preservation and Literacy Enhancement** Hanan Khalifa, Jing Wei, Alistair Van Moere	**Research Papers** Location: **Phramingkwan** Buiding 2, Second Floor **L1 Intuition vs. Three Popular LLMs: Can LLMs Mark L2–L1 Meaning Recall Tests?** Aaron Olaf Batty, Jeffrey Stewart, Laurence Anthony, Christopher Nicklin, Keita Nakamura, Kanako Tomaru, Stuart McLean	**Research Papers** Location: **Room 401** Building 3, Fourth Floor **Cognitive Processes in Intertextual Summary Tasks: A Study of Multilingual Asian Test-Takers** Nathaniel Ingram Owen, Haiyan Xu, Oliver Bigland	**Research Papers** Location: **Room 405** Building 3, Fourth Floor **Gender Representation in IELTS and NMET Reading Texts: A Comparative Analysis Using SFL and Social Actor Network Frameworks** Xiaoqin Huang, Xiangdong Gu	**Research Papers** Location: **Poonsapaya** Building 2, Third Floor **Survive and Thrive in a Third Space: When Chinese Students Come To NZ** Qiuxian Chen, Gavin Brown, Yue Wang	**Research Papers** Location: **Duangduen** Building 3, Ground Floor **Construct Comparability of TOEFL iBT and Duolingo English Test** Sara Cushing

Date: Friday, 06/June/2025						
	Translanguaging in Listening Assessment: Inclusion of L1 Responses in L2 Recall Tasks Shelly Xueting Ye, Matthew Wallace	**An Agile Approach to Utilizing AI Technology to Support Young EFL Students' Writing Skills** Mikyung Kim Wolf, Michael Suhan	**Evaluating the Integration of Listening Comprehension in Paired Oral Communication Tests** Leyla Karatay	**Assessing Second Language Pragmatic Competence for Intercultural Communication: Test Localisation Targeting UK Pre-Sessional Students** Shishi Zhang	**One Step Further: Understanding The Role of a Tailor-Made Genai-Powered Chatbot in Affecting L2 Learner Interaction and Engagement** Shangchao Min, Yuhong Gao, Jie Zhang	**Distinct Listening Biotypes and Their Application in Test Validation: A Neuroimaging Study** Vahid Aryadoust
	Interactive Videos in An L2 Listening Test: How do They Affect Test Scores? Shanshan He		**Language Assessment Literacy Inventory Development: Understanding Learner Perspectives in Multilingual Contexts** Jiyoon Lee, Yuko Butler	**Assessing Language for Professional Registration: Does the IELTS Academic Capture Australian Teachers' Communicative Demands?** Xiaoxiao Kong	**Exploring Cultural and Pragmatic Challenges in AI-mediated Speaking Assessments** Yasin Karatay, Jing Xu, Leyla Karatay	**Metacognitive Awareness and Its Relationship with L2 Listening in a Model of Language Proficiency Using a Meta-Analytic Structural Equation modeling** Yo In'nami, Mike W.-L. Cheung, Rie Koizumi, Matthew P. Wallace
12:30pm - 1:30pm	**Lunch** Location: **Sumon** (Building 1, Ground Floor)					
1:30pm - 3:00pm	**Research Papers** Location: **Ampai** Building 1, Ground Floor **Leveraging Generative AI for Interactive Assessment in Multicultural Contexts** Inyoung Na	**Research Papers** Location: **Phramingkwan** Building 2, Second Floor **Investigating Factors Influencing HKDSE English Writing Scores: A Five-Facet Rasch Analysis** Meng-Hsun Lee, Kuan-Yu Jin	**Research Papers** Location: **Room 401** Building 3, Fourth Floor **From Text to Speech: Exploring Content-Related Features in an Integrated Listening-Into-Speaking EAP Task** Nahal Khabbazbashi, Fumiyo Nakatsuhara, Chihiro Inoue, Johnathan Jone	**Research Papers** Location: **Room 405** Building 3, Fourth Floor **Language Assessment Literacy: Development of Pre-Service English Teachers' Perceptions and Practices Before and After Doing the Practicum** Punchalee Wasanasomsithi, Benjawan Plengkham	**Symposium** Location: **Poonsapaya** Building 2, Third Floor **Evolving Glocalization in Language Testing in Asia: Reflection and Implications** ***Chair(s):*** Jessica Row Whei Wu ***Discussant(s):*** Lynda Brigid Taylor	**Symposium** Location: **Duangduen** Building 3, Ground Floor **The Promise and Perils of Investigating Writing Assessment in Other Languages through the Lens of English** ***Chair(s):*** Atta Gebril, Beverly Baker ***Discussant(s):*** Slobodanka Dimova

Date: Friday, 06/June/2025						
	Can LLMs Generate Human-Like Responses for Training Fairer AES Systems? Burak Senel	**The Effects of Automated Writing Evaluation (AWE) on EFL Students' IELTS Writing** Napat Jitpaisarnwattana, Nick Saville	**A Mixed-Methods Investigation of Test-Takers' Performances on Integrated Speaking Tasks: Considering Task Design, Modality of Input and Proficiency Levels** Sathena Chan[1], Lyn May	**Understanding Chinese Lexical Inferencing: Assessment and Impacts of Word-internal and Word-external Abilities** Yuxin Peng[1], Stanley Haomin Zhang, Cecilia Guanfang Zhao	*Presentations of the Symposium* **Translation Assessment in Large-scale Language Testing: A Case Study of the College English Test** Yan Jin **Glocal Tests and Test Glocalization: The Case of TEPS** Yong-Won Lee **Evolution of English Proficiency Testing: Journey from GEPT to BESTEP** Rachel Yifen Wu, Anita Chunwen Lin **The Evolution and Devolution of Language Tests in Vietnam - An Ecological Review** Quynh Thi Ngoc Nguyen	*Presentations of the Symposium* **Chinese Character Matters!: An Examination of Linguistic Accuracy in Writing Performances on the HSK test** Xun Yan, Jiani Lin **Lexical and Syntactic Analyses Procedures Applied to the French Language: What Works and What's Lost in Translation?** Randy Appel, Angel Aria, Beverly Baker, Guillaume Loignon **The Applicability of Hyland's Metadiscourse Model to L1 Arabic Writing: A Critical Cross-Linguistic Study** Abdelhamid M. Ahmed, Lameya M. Rezk **Beyond EFL – The Use of Theoretical Models for Integrated Writing Assessment in a German Context** Sonja Zimmermann
	Are We Still Measuring the Intended Construct Through Computerized Dynamic Assessment? Insights from Confirmatory Factor Analysis Meng-Hsun Lee		**An Actor-Network Approach to Diversity in the Interpretation and Use of Concordances for Tests of English for Academic Purposes** Anthony Green, Leda Lampropoulou	**Proctoring Language Assessments in Multicultural Contexts** Alina A von Davier, Will Belzak, Rose Hastings, Basim Baig		

Date: Friday, 06/June/2025						
3:00pm - 3:30pm	**Coffee Break (Sponsored by LTTC)** Location: **Sumon** (Building 1, Ground Floor)					
3:30pm - 5:00pm	**Research Papers** Location: **Ampai** Building 1, Ground Floor **Comparing Different Standard Setting Methods for Aligning Local English Reading and Listening Comprehension Test Scores with the CEFR** Supong Tangkiengsirisin, Sun-Young Shin, Suchada Sanonguthai	**Research Papers** Location: **Phramingkwan** Building 2, Second Floor **Uncovering K12 Students' Engagement Strategies with ChatGPT's Feedback in Reading Assessments: A Lag Sequential Analysis** Ziqi Chen, Wei Wei, Jiawei Zhang	**Research Papers** Location: **Room 401** Building 3, Fourth Floor **A Longitudinal Investigation of the Relationship of Language and Academic Success of International Students** Donald Bruce Russell	**Research Papers** Location: **Room 405** Building 3, Fourth Floor **Examining Product and Process Features of Two TOEFL iBT Opinion Writing Tasks: A Validation Study** Huiying Cai, Ching-Ni Hsieh	**Research Papers** Location: **Poonsapaya** Building 2, Third Floor **Unequal Trajectories? An Examination of the Role of Socio-Economic Status in Shaping Language Proficiency Development in a Mexican Higher Education Context** Nahal Khabbazbashi, Parvaneh Tavakoli, Edgar Emmanuell Garcia-Ponce, Gareth McCray	**Research Papers** Location: **Duangduen** Building 3, Ground Floor **Alternative Approaches to Reading Item Difficulty Calibration: Perspectives from Text Complexity** Ying Zheng, David Booth
	Can Expert Reviewers Accurately Identify and Explain the Reasons for DIF In Language Assessments? Jacqueline Church, Will Belzak, Yigal Attali, Yena Park	**English Exit Exam in Thai Higher Education: Test Characteristics and Teachers' Reflections on Their Practice** Anchana Rukthong, Punjaporn Pojanapunya, Somruedee Khongput	**Simulated Real-Life Tasks as a Tool to Investigate the Extrapolation Inference of Language Assessments for Professional Purposes** Fatima Montero	**Young Learner Effect in Assessment: An Interactional Multimodal Analysis** Gordon Blaine West, Jason Kemp, Shea Head	**Investigating Presentation-Mode Effects on L2 Graph-Based Writing: An Eye-Tracking Study** Daniel Yu-Sheng Chang	**Integrating Learning with Assessment: Evaluating the SBA Approach with Rcts** Thi Ngoc Quynh Nguyen, Antony Kunnan, Do Thu Hoa
	Understanding the Language Assessment Literacy Needs of Junior High EFL Teachers in China: A Validation Study Zhengqing Luo, Dunlai Lin	**Implementing Learning Oriented Assessment at an Institutional Level: Key Considerations and Challenges for its Validity** Angeliki Salamoura	**Working Memory among Chinese Learners in Compulsory Education and Its Relationship with English Achievement** Yinjie Tang		**Impact and Fairness of Retaking Single IELTS Test Components: Global Perspectives and Local Impacts in Asia** Hye-won Lee, Emma Bruce, Jan Langeslag, Tony Clark, Reza Tasviri	**I Can Still Manipulate a Human Interlocutor: Test Taker Perceptions of Taking an Interactive Speaking Test with an Avatar-Enabled Spoken Dialogue System** Reza Neiriz

Date: Friday, 06/June/2025	
5:00pm - 6:30pm	**Works-in-Progress** Location: **Sumon** (Building 1, Ground Floor) 1. **AI-Powered Listening Tests—How Reliable are They?** Junyan Guo 2. **Workshopping LAL: Engaging with University Admissions and Recruitment Staff in Language Assessment Literacy Development** Daniel M. K. Lam, Angela Gayton 3. **Developing In-House English Proficiency Tests for Thai Universities: A Practical Approach to Aligning with CEFR** Worasuda Wattanawong, Siriphan Suwannalai, Dusadee Rangseechatchawan 4. **How Test Takers' Attention Distribution on Source Text Affects Their Performance in the Continuation Task: An Eye-Tracking Study** Yang Zhang, Lin Shi, Lianzhen He 5. **Using a GenAI-based Conversational Agent to Assess Second Language Learners' Interactional Competence** Zhuohan Hou, Shangchao Min 6. **The Development and Initial Validation of a Standardized Test of English for University Admission Aligned with a National Reformed Curriculum** Thao Thi Phuong Nguyen, Chi Thi Nguyen, Yen Thi Quynh Nguyen, Hoa Quynh Nguyen, Quynh Thi Ngoc Nguyen, Duong Thuy Le 7. **Feeding the Machine: A Comparison between Analytical and Holistic Scoring to Inform an Automated Essay Scoring System** Joni Kruijsbergen, Fauve De Backer, Orphée De Clercq, Goedele Vandomme 8. **Exploring Raters' Scoring Processes in Assessment of English-Chinese Consecutive Interpreting: A Qualitative Study Based on Retrospective Verbalization** Mengting Jiang 9. **Effects of Response Language on Test-Takers' Performance and Cognitive Processes in L2 Listening Recall Tasks** Phuong Nguyen, Ahmet Dursun 10. **Exploring Task Designs Suitable for High-Stakes Spoken Japanese Language Assessment** Fumiyo Nakatsuhara, Chihiro Inoue, Atsuko Osumi, Yumi Horikawa 11. **Prototyping an Information-Based Academic Writing Assessment** Chengyuan Yu

Date: Friday, 06/June/2025	
	12. The Predictive Power of Lexical Richness Indices in Chinese EFL Learners' Performance in L2 Speaking Tasks Lin Shi, Yang Zhang, Lianzhen He **13. Test Scores and Speech Samples: Extrapolating from a Computer-Delivered Test of Speaking for University Admission to Group Oral Discussions** Yujia Zhou, Masashi Negishi, Asako Yoshitomi **14. Investigating Self-Identified Language Assessment Literacy Needs of Teachers in North America, Europe, and Asia** Benjamin Kremmel, Luke Harding **15. Investigating the Feasibility of ChatGPT for Generating Passages and Items in Different Types of EFL Reading Tasks** Lin Shi, Yuhong Gao, Lianzhen He **16. Exploring Assessment Literacy among EMI University Teachers in STEM Fields: A Mixed-Methods Study in Taiwan** I-Chun Vera Hsiao **17. Understanding Multilingual Communicative Competence: Exploring Possibilities for Language Assessment in Higher Education** Slobodanka Dimova **18. Grammatical Complexity in Thai EFL English-major Students' Writing** Pong-ampai Kongcharoen, Xinyu Zhao **19. Rating Quality across Different Presentation Modes in L2 Writing Assessments: A Comparison of Human Raters and AI-Generated Scoring** Daniel Yu-Sheng Chang, Pu Pu
6:30pm - 8:30pm	**Networking Dinner** Location: **Meet in front of Sumon** (Building 1, Ground Floor)

Date: Saturday, 07/June/2025						
8:30am - 10:30am	**Research Papers** Location: **Ampai** Building 1, Ground Floor **Disentangling Multimodality in Speaking Assessment: The Interplay of Nonverbal Behavior, Affect, and Language in Estimates of Second Language Ability** John Dylan Burton	**Research Papers** Location: **Phramingkwan** Building 2, Second Floor **Culturally Tailored Assessments: Investigating the Role of Personalized Images in Writing Tasks** Andrew Runge, Geoffrey T. LaFlair, Jacqueline Church	**Research Papers** Location: **Room 401** Building 3, Fourth Floor **Evaluating the Logic of a Policy-Driven National Language Test in Indonesia: The Test of Indonesian Proficiency (UKBI)** Rahmad Adi Wijaya	**Research Papers** Location: **Room 405** Building 3, Fourth Floor **Differential Item Functioning in Audio-Visual English Listening Comprehension Assessments Among Young Learners** Sun-Young Shin, Senyung Lee	**8:30am-10:00am** **Symposium** Location: **Poonsapaya** Building 2, Third Floor **Assembling, Adapting, Adopting: The Development and Implementation of Standards and Frameworks for Young Learners** ***Chair(s):*** Barry O'Sullivan ***Discussant(s):*** Barry O'Sullivan *Presentations of the Symposium* **Principled localisation of the CEFR: an example of the CEFR-J** Masashi Negishi **Linking Materials to the CEFR to Develop a Standardised Approach to Assessment in a Global Context** Carolyn Westbrook, Johnathan Cruise	**8:30am-10:00am** **Symposium** Location: **Duangduen** Building 3, Ground Floor **Assessing Internationally Mobile Healthcare Professionals: Redrawing the Boundaries of Language Testing** ***Chair(s):*** Gad Lim, Peter Kim ***Discussant(s):*** Lynda Taylor *Presentations of the Symposium* **Modelling Communication Challenges of Aged Care Workers from Multilingual and Multicultural Backgrounds** Ute Knoch, Philipa Mackey, Sally O'Hagan, Ivy Chen **The Social Impact of Tests: A Multi-Stakeholder Evaluation of Introducing OET in the UK** Brigita Seguis
	The Impact of an AI-interviewer's Relationship-Building Dialogue Strategies on Language Performance Fuma Kurata, Masaki Eguchi, Máo Saeki, Shungo Suzuki, Yoichi Matsuyama	**Testing Extended Time Accommodations: Differential Effects on Language Test Performance** Ramsey Lee Cardwell, William Belzak, Jill Burstein, Ruisong Li	**Participants as Co-Researchers: A Co-Analysis Approach to Language Assessment for Immigration** Coral Yiwei Qin	**Assessing Business English Competence: The Role of Linguaskill Business Test in Multicultural Contexts** Aynur Ismayilli Karakoc		
	Advances in the Assessment of Interactional Competence: A Systematic Literature Review Anh Nguyen, Noriko Iwashita	**Developing and Validating a Self-Assessment Tool for Measuring Vietnamese Teacher Competence in Multiple-Choice Test Item Writing for Large-Scale English Reading and Listening Skill Tests** Phuong Viet Ha Ngo, Shelley Gillis, Cuc Nguyen	**A Study in Contrasts: Investigating Human and Automated Scorer Evaluation of Textual Changes** Sarah R. Hughes	**Charting the Landscape of K-12 Educators' Views on Automated Writing Scoring & Feedback** Jason A. Kemp, Lynn Shafer Willner		

Date: Saturday, 07/June/2025						
	Predicting Washback of a Speaking Assessment in the Japanese University Entrance Exam Context David Allen	**Examining the Willingness to Communicate (WTC) Scale in Advanced Learners of Languages Other Than English (LOTE)** Troy L Cox	**Investigating the Construct of the Continuation Task and Test-Takers' Cognitive Processes: An Eye-tracking Study** Yang Zhang, Lin Shi, Lianzhen He	**Expanding the Interactive Academic Listening Construct: Addressing the Gap Between Academic Lectures and Proficiency Tests** Burak Senel, Ananda Senel	**Developing a Cognitive Diagnostic Computerized Adaptive Test (CD-CAT) based on China's Standards of English Language Ability (CSE)** Lianzhen He **Proficiency Level Descriptors in U.S. Contexts: From Standards to Assessment to Instruction** Margo Gottlieb, Lynn Shafer Willner	**From Preparation to Practice: The Role of OET, IELTS, and PTE in Preparing Nurses for Workplace Communication** Jason Fan, Ute Knoch, Michael Davey, Ivy Chen, Sally O'Hagan, David Wei Dai **Navigating Language Proficiency Testing Policies: Challenges for Migrant Healthcare Professionals in Canada's Health System** Eunice Jang, Maryam Wagner
10:30am - 11:00am	**Coffee Break (Sponsored by Goethe-Institut)** Location: **Sumon** (Building 1, Ground Floor)					
11:00am - 12:00pm	**Alan Davies Lecture** Location: **Poonsapaya** (Building 2, Third Floor) **Assessment of Social Language Use: From Pragmatics to Interactional Competence** Professor Carsten Roever, University of Melbourne					
12:00pm - 1:30pm	**Lunch (Sponsored by Eiken Foundation of Japan)** Location: **Sumon** (Building 1, Ground Floor)					
12:30pm - 2:00pm	**ILTA Annual Business Meeting** Location: **Ampai** (Building 1, Ground Floor)					

Date: Saturday, 07/June/2025

1:30pm
-
3:00pm

Posters
Location: **Room 104** (Building 1, Ground Floor)

1. **Exploring the Appropriateness of the IELTS Academic for Australian Teacher Registration: Insights from Domain Insiders**
   Xiaoxiao Kong

2. **Development of an English as a Second Language Proficiency Test for Spanish-speaking migrant children in Trinidad & Tobago**
   Romulo Guedez Fernandez

3. **Inclusive and Equitable Test Development: Stakeholder Involvement of the Jewish Community in Antwerp (Flanders, Belgium)**
   Mieke De Latter, Fauve De Backer

4. **Teacher-Student Collaborative Assessment in the Production-Oriented Approach to Improve English Writing Proficiency and Perceived Self-Efficacy of Chinese Undergraduate Students**
   Xi Li, Punchalee Wasanasomsithi

5. **Impact of Corpus-Assisted Self-Assessment on Enhancing Speaking Proficiency and Reducing Anxiety in EFL Learners**
   Pei-Ju Hsiung, Po Han Wu, Wei-Ting Wu

6. **Training Needs of Middle School EFL Teachers on Language Assessment Literacy: A Study Based on Quantitative Ethnography**
   Hui Liu, Xiaomei Ma

7. **Language Assessment Practices from Public School Teachers in Chile: Balancing Contextual Factors**
   Salomé Villa Larenas

8. **Redesigning a Kindergarten to Grade 12 ELP Assessment Score Report: Gathering Evidence from Multiple Perspectives**
   Ahyoung Alicia Kim, Jason A. Kemp, Fujiuju Daisy Chang, Kerry Pusey, Fabiana MacMillan

9. **Assessing Listening Skills in Vietnamese: A Case of Track Separation in Beginning Vietnamese**
   Hanh Nguyen

10. **Raters' Professional Background as a Potential Source of Distinct Behaviors: A Mixed-Methods Investigation to Interpreting Quality Assessment**
    Xini Liao, Jinjie Wu, Mingwei Pa

11. **Empowering Teachers and Students with AI: Developing a Vocabulary Assessment for Korean EFL Learners on the CAFA Platform**
    Jung-Hee Byun, Kyung-Il Yoon

Date: Saturday, 07/June/2025

12. **The Use of AI to Generate Picture Prompts for Story Writing Tasks**
    Haeun Hannah Kim

13. **Investigating the Impact of Languages Other Than English Subjects in the National Matriculation Test: A Poststructuralist Perspective**
    Chenyang Zhang

14. **Mapping Language Needs Through a Proficiency Framework: A Case Study**
    Troy L Cox

15. **A Comparison of Writing Assessment between Japanese High School Teachers and Aptis English Test Professional raters**
    Chiho Young-Johnson

16. **My B1 is not Necessarily your B1 - Evidence from Germany**
    Hella Klemmert, Juliane Braeutigam

17. **Analyzing Written Proficiency Levels in Portuguese as an Additional Language: Corpus-driven Results from the CorCel Corpus**
    Elisa Marchioro Stumpf, Juliana Roquele Schoffen, Deise Amaral, Isadora Dahmer Hanauer

18. **Rubric Co-construction in Language for Specific Purposes Assessments**
    Qiaona Yu

19. **Enhancing Reliability in Writing Assessment: Investigating the Use of Customised ChatGPT 4.0 and Human Raters**
    Turgay Han, Özgür Şahan, Doğan Saltaş

20. **Adapting to AI: A Qualitative Study of Teachers' Formative Assessment Practices and Perceptions of Generative AI**
    Angelie Ignacio

21. **Stepping Stones or Stumbling Blocks: Oral Proficiency Level Descriptors and their Effects on Rater Confidence**
    Birgitte Grande, Clayton D. Leishman

22. **Enhancing Students' Feedback Literacy Using Generative Artificial Intelligence (GAI)**
    Limei Zhang

23. **Washback of Multilingual Assessment**
    Karin Vogt, Dina Tsagari, Lucilla Lopriore

24. **The Use of Automated Feedback in Turkish EFL Students' Writing Classes**
    Turgay Han, Elif Sari

Date: Saturday, 07/June/2025						
	25. **The Impact of Exposure to Different English Accents on Chinese Children's Learning of EFL --- Listening, Speaking, and Attitude** Zhuohan Chen 26. **Rating Performance and Quality of Novice and Expert Raters of Integrated Writing: Findings from a Mixed Methods Study in a German Higher Education Context** Valeriia Koval, Ximena Delgado Osorio, Claudia Harsch, Johannes Hartig 27. **Towards Effective CLIL Assessment: Understanding Classroom Questioning Practices in Asian EFL Settings** Wenhsien Yang 28. **Dynamic Assessment of Integrated Argumentative Writing: Diagnosing and Promoting Multilingual Writers' Development in the ZPD** Lu Yu 29. **Proposing and Applying a Conceptual Model of Test Fairness in a Local Context in China** Juan Zhang, Lianzhen He					
3:00pm - 3:30pm	**Coffee Break** Location: **Sumon** (Building 1, Ground Floor)					
3:30pm - 5:00pm	**Research Papers** Location: **Ampai** Building 1, Ground Floor **The Role of Prompt Engineering in Ensuring the Consistency Between Instructor and LLM Checklist Ratings on Written Summary Content** Yasuyo Sawaki, Yutaka Ishii[2], Hiroaki Yamada, Takenobu Tokunaga	**Research Papers** Location: **Phramingkwan** Building 2, Second Floor **"Beyond Anxiety": Unveiling the Emotional Washback of the High-Stakes Hanyu Shuiping Kaoshi (HSK) on L2 Chinese Learners through Control-Value Theory (CVT)** Yang Yang	**Research Papers** Location: **Room 401** Building 3, Fourth Floor **Expanding Construct in EAP Speaking Assessment: Defining and Operationalizing a Critical Thinking Perspective** Shengkai Yin	**Research Papers** Location: **Room 405** Building 3, Fourth Floor **Enhancing Inter-Coder Reliability in Online Think-Aloud Protocols Through Visual Behavior Analysis** Ananda Senel, Nathaniel Owen, Oliver Bigland	**Research Papers** Location: **Poonsapaya** Building 2, Third Floor **Are Proctors of High-Stakes Language Assessments Fair?** Will Belzak, Alina von Davier	**Symposium** Location: **Duangduen** Building 3, Ground Floor **West Has Met East: A Transnational Language Assessment Literacy Project of Southeast Asian Languages** ***Chair(s):*** Ahmet Dursun ***Discussant(s):*** Catherine Baumann, Ahmet Dursun *Presentations of the Symposium*

Date: Saturday, 07/June/2025						
	Performance of Generative AI models on Academic Writing Tasks: A Systematic Review Livija Jakaite, Vitaly Schetinin, Chihiro Inoue, Stanislav Selitskiy	**Human-AI Collaboration Patterns of EFL Learners in AI-Assisted Academic Writing** Shasha Xu, Xiang Xu, Fanxi Shen	**Validating an Assessment of L2 Interactional Competence in Online Text Chat** Xingcheng Wang	**An Evidence- and Consensus-based Approach to Ethical AI for Language Assessment** Carla Pastorino-Campos, Evelina Galaczi	**Investigating Standard Setter Cognition** Doris Moser-Froetscher, Stefanie Hollenstein, Robert Hilbe	**Enhancing Indonesian Reading Proficiency: Lessons from a Multi-Year Reading Proficiency Test Pilot Program** Sakti Suryani, Erlin Barnard
	A Diagnostic Classification Model Analysis of Thai EFL Learners' Academic English Writing Apichat Khamboonruang	**Digitalizing Language Tests for Migrants: Investigating a Multicultural Test Population's Digital Literacy and Target Language Use** Benjamin Kremmel, Eva Konrad, Doris Moser-Froetscher, Keri Hartman		**Adapting and Evaluating Formative L2 Comprehensibility Assessment Scale** Aki Tsunemoto, Rie Koizumi, Makoto Fukazawa, Yo In'nami, Ryo Maie, Mariko Abe	**Intercultural Considerations When Developing Materials and Assessments for Minoritised Languages** Lynda Brigid Taylor, Jill Wigglesworth, Rosalie Grant	**Exploring Learners' Perspective of Proficiency-oriented Performance-based Assessments in Burmese** Chan Lwin, Maw Maw Tun, Ye Min Tun, Kenneth Wong **Expanding Horizons: Inter-institutional and Transnational Collaboration in Assessment and Lesson Design and Development** An Sakach
5:00pm - 6:00pm	**CHULA Campus Tour** **Meet in front of Sumon** (Building 1, Ground Floor)					

Date: Sunday, 08/June/2025					
8:30am - 10:30am	**Research Papers** Location: **Ampai** Building 1, Ground Floor **Exploring Profiles of Researcher-teacher Collaboration in Language Assessment Literacy Studies** Beverly Baker, Lynda Taylor, Louis-David Bibeau	**Research Papers** Location: **Phramingkwan** Building 2, Second Floor **Investigating the Challenges of Language Assessment Across Contexts Through an Assessment Literacy MOOC** Carolyn Westbrook, Richard Spiby, Jordan Weide	**Research Papers** Location: **Room 401** Building 3, Fourth Floor **Towards a Framework of Critical Thinking for Assessing EAP Speaking** Shengkai Yin	**Symposium** Location: **Poonsapaya** Building 2, Third Floor **Beyond "In the Loop": Human-AI Collaboration in Human-centered Language Assessment** ***Chair(s):*** Eunice Eunhee Jang ***Discussant(s):*** Xiaoming Xi *Presentations of the Symposium* **Human-centered AI for Language Assessment Development and Administration** Jill Burstein, Geoffrey T. LaFlair, Alina. von Davier **Human-AI Teaming in Language Assessments through A Human-centered Approach** Eunice Eunhee Jang, Liam Hannah **Explaining AI Scoring Models to Humans** Erik Voss **A Framework for AI in Language Testing: Bidirectional AI-Human Alignment** Alistair Van Alistair, Jing Wei	**Symposium** Location: **Duangduen** Building 3, Ground Floor **Digitally Empowered Assessment of Interactional Competence in Second Language Contexts** ***Chair(s):*** Yunwen Su ***Discussant(s):*** Carsten Roever *Presentations of the Symposium* **AI Familiarity as a New Potential Source of Bias in Interactional Competence Assessment with Conversational Agents** Shungo Suzuki, Hiroaki Takatsu, Kotaro Takizawa, Ryuki Matsuura, Mao Saeki, Yoichi Matsuyama **An Ecological Perspective on Classroom Assessment of Pre-Service Teachers' Interactional Competence Using Technology-mediated Speaking Tasks** Soo Jung Youn **Assessing TAlkEZly Mediated L2 Interactional Competence Development in Blended Teaching Context** Chen Shen, Yaru Meng, Xi Qian
	Unpacking Test-Wiseness Strategies: Effects on Second Language Reading Performance Ray J. T. Liao, Kwangmin Lee	**Development and Validation of a Language Assessment Literacy Rubric: A Case for University-level English Teachers in China** Ling Gan, Xun Yan	**An Exploratory Study for Specifying The Q-Matrix in Cognitive Diagnostic Assessment of Chinese EFL Speaking Proficiency: Combining Theory with Data** Shuting Zhang, Lianzhen He		
		Promoting Assessment Literacy and Professionalization in Language Testing: Reflecting on the Role and Impact of the Studies in Language Testing Series Lynda Brigid Taylor, Nick Saville			

Date: Sunday, 08/June/2025					
				AI-Driven Diagnostic Assessment Grows Together with Learners: Towards Individualised Actionable Diagnostic Feedback for L2 Speaking Hiroaki Takatsu, Shungo Suzuki, Ryuki Matsuura, Miina Koyama, Yoichi Matsuyama	**Measuring L2 Interactional Competence: A Comparison of Human and AI-Mediated Roleplay Assessments** Yunwen Su, Xi Chen **Assessing Interactional Competence in a Digital Era: Developing and Validating a Rating Scale for a Computer-Based Paired Discussion Task** Sa Xiao, Yan Jin
10:30am - 11:00am	**Coffee Break** Location: **Sumon**				
11:00am - 12:00pm	**Special Session** Location: **Ampai** (Building 1, Ground Floor) **Playing the Peer Review Game: Tips on Academic Publishing from Seasoned Journal Editors** Talia Isaacs, Xun Yan, Jin Yan, Elvis Wagner	**Special Session** Location: **Poonsapaya** (Building 2, Third Floor) **Navigating the PhD Journey in Applied Linguistics and Language Assessment** Yena Park, Haeun Kim, Ing Kongchareon **Moderator:** Sun-Young Shin	**Special Session** Location: **Duangduen** (Building 3, Ground Floor) **Perspectives and Current Priorities - What are the Benefits of Collaboration between the Associations?** Nick Saville, Quynh Nguyen, Salomé Villa Larenas **Moderator:** Rama Matthew		
12:00pm - 1:30pm	**Lunch** Location: **Sumon**				
12:30pm - 1:30pm	**LAQ Editorial Board Meeting** Location: **Room 405** (Building 3, Fourth Floor)				

Date: Sunday, 08/June/2025				
1:30pm - 2:30pm	Location: **Ampai** Building 1, Ground Floor **SIG** **Automated Language Assessment (ALASIG)** Jing Xu, Xiaoming Xi	Location: **Room 401** Building 3, Fourth Floor **SIG** **Test-taker Insights in Language Assessment (TILASIG)** Andy Jiahao Liu, Ray Jui-Teng Liao	Location: **Room 405** Building 3, Fourth Floor **SIG** **Language Assessment for Young Learners (Young Learners SIG)** Mark Chapman, Veronika Timpe-Laughlin, Jeanne Beck	Location: **Poonsapaya** Building 2, Third Floor **SIG** **Language Assessment Literacy (LALSIG)** Rebecca Yaeger, Xun Yan, Sharry Vahed, Elsa Fernanda Gonzalez, Gladys Quevedo-Camargo
2:30pm - 3:30pm	**Cambridge/ILTA Distinguished Achievement Award Lecture** Location: **Poonsapaya** (Building 2, Third Floor) **Navigating by the Stars': Reflections from a Lifelong Journey towards Language Assessment Literacy and Professional Competence** Professor Lynda Taylor, University of Bedfordshire			
3:30pm - 4:00pm	**Closing** Location: **Poonsapaya** (Building 2, Third Floor)			
4:00pm - 4:45pm	**Khob Khun Break** Location: **Sumon** (Building 1, Ground Floor)			
4:45pm	**Bus to banquet - Meet in front of Sumon** (Building 1, Ground Floor)			
5:30pm - 8:30pm	**Banquet** Location: **Royal Orchid Sheraton Riverside Hotel Bangkok** (Ballroom 1, Second Floor)			

TABLE OF CONTENTS

Conference Schedule Overview 1
Message from the ILTA President 19
Welcome from the LTRC Chair and Co-Chairs 21
Conference Organization 23
ILTA Executive Board and Committee Members 25
Awards and Grants Committees 26
Award Winners 27
Housekeeping and Important Information 30
Maps and Floor Plans 30
Plenary Sessions 39
Samuel J. Messick Memorial Lecture 39
Alan Davies Lecture 41
Cambridge/ILTA Distinguished Achievement Award 43
Pre-Conference Workshops 45
Special Sessions 56
Opening Symposium 59
Symposium 1 64
Symposium 2 72
Symposium 3 80
Symposium 4 84
Symposium 5 89
Paper and Demo Summaries – Friday, June 6, 2025 94
Works-in-Progress – Friday, June 6, 2026 122
Paper and Demo Summaries – Saturday, June 7, 2025 133
Poster Presentations – Saturday, June 7, 2025 152
Paper and Demo Summaries – Sunday, June 8, 2025 170
Duolingo 2024 Doctoral Dissertation Awards 175

Message from the ILTA President

On behalf of ILTA, I would like to welcome all conference delegates to our 46th Language Testing Research Colloquium in Thailand! ขอต้อนรับสู่การประชุมวิชาการ LTRC! Bangkok – a vibrant and dynamic city – will provide the perfect location for our first ever LTRC in South-East Asia, a region of diverse linguistic and cultural contexts. The conference theme highlights this diversity: *Language assessment in multicultural contexts: West meets East*. I look forward to a range of papers, symposia, works-in-progress, and posters that expand on this theme, providing research-based insights into a range of global and local issues in our field.

I am grateful to our host institutions, Chulalongkorn University and Indiana University, and particularly to the conference chair and co-chairs, Jirada Wudthayagorn, Sun-Young Shin, and Punchalee Wasanasomsithi, as well as the other members of the hard-working conference organizing committee: Wutthiphong Laoriandee, Raveewan Viengsang, Chanisara Tangkijmongkol, Chariya Prapobratanakul, and Mintra Puripunyavanich. I know that the committee have worked tirelessly, and with great purpose, to make sure that this will be an unforgettable conference. I'm sure it will be! We appreciate your efforts, and we look forward to seeing your vision for LTRC come to fruition.

The organizing committee have received assistance and support from Dorrian Regan and Valerie Smith at Nardone (ILTA's management company) and has drawn on the experience of the LTRC Advisory Committee (Beverly Baker, Claudia Harsch, Benjamin Kremmel, Lorena Llosa, Conor McKeown, Heike Neumann, and Elvis Wagner). I would also like to acknowledge the hard work of members of the ILTA Executive Board in liaising on various matters, chairing awards committees, and generally supporting ILTA's mission.

LTRC is always well-supported through the sponsorship of many organisations within our community, and this year that tradition has continued. We are very grateful for the contributions of our platinum sponsors (Cambridge University Press & Assessment, Chulalongkorn University, Duolingo English Test, Eiken Foundation of Japan, IELTS, and Pearson English Language Learning), our gold sponsors (British Council English & Exams, Goethe-Institut, LTTC [Language Training and Testing Center]), our silver sponsor (Michigan Language Assessment), and our bronze sponsors (Lancaster University and WIDA). These sponsorships enrich all LTRC delegates' experience, and we appreciate all sponsors' generosity.

All of this planning, hard work, and support has led to an exciting conference program full of cutting-edge research: a feast for anyone interested in language assessment.

We begin with four workshops covering topics as varied as *Language test design and alignment to the CEFR* (Viphavee Vongpumivitch and Napat Jitpaisarnwattana), *Publishing your research in the AI era* (Qin Xie), *An applied introduction to no-code regression and machine learning for language assessment* (Vahid Aryadoust), and *Impact evaluation: Global perspectives and best practices* (Micheline Chalhoub-Deville, Hanan Khalifa, Eunice Eunhee Jang). A special thank you to all the workshop organisers/presenters for running these sessions.

Before the opening symposium and welcome reception, we also provide our usual newcomers session, run this year by Lia Plakans and Benjamin Kremmel. Please come along if you're new to LTRC, or even if you've been to multiple: all are welcome!

During the main conference itself, I am particularly looking forward to our three keynote speakers: Lyle Bachman – who will deliver the Messick Lecture (supported by ETS), Carsten

Roever – who will deliver the Alan Davies Lecture (supported by the British Council), and Lynda Taylor – the recipient of this year's Cambridge/ILTA Distinguished Achievement Award (supported by Cambridge University Press & Assessment and ILTA). Alongside, there will be an impressive number of papers, symposia, WiPs and posters. I would also like to draw special attention to a range of special sessions and four SIG (Special Interest Group) meetings that will take place on the final day.

ILTA members are asked to attend the Annual Business Meeting, which will take place 12:30-14:00 on Saturday 7 June. You will have received some papers prior to this meeting, so please take some time to read through this information in advance if you haven't already. We will have time to discuss a range of topics, including discussion of LTRC proposals for 2028. If you're not an ILTA member yet, please consider joining our professional community!

As ever, with LTRC, there is also a very full social program to keep us all entertained, including a welcome reception, networking dinners, a Chula campus tour, the banquet (where we will also give out awards), and a number of post-conference trips.

This is going to be a conference to remember. Thank you once again to the organizing team, and I look forward to meeting you all in Bangkok.

Luke Harding

ILTA President 2025

Welcome from the LTRC Chair and Co-Chairs

Message from the Chair and Co-Chairs,

It is our distinct honor and heartfelt pleasure to welcome you all to the **46th Language Testing Research Colloquium (LTRC 2025),** held in **Bangkok, Thailand** from **June 4–8, 2025.**

We are delighted to host this year's LTRC in the Land of Smiles, where rich cultural heritage meets warm hospitality and dynamic academic exchange. Whether you are joining us from near or far, as a returning LTRC attendee or a first-time participant, we warmly invite you to be part of what promises to be a memorable and inspiring week of shared learning, innovation, and connection in the heart of Southeast Asia.

This year's theme, ***"Language Assessment in Multicultural Contexts: West meets East,"*** reflects a growing imperative in our field: understanding and enhancing the real-world effects of language assessments on education systems, learners, and societies. Across the globe, language tests continue to shape policy decisions, gatekeeping mechanisms, classroom practices, and life opportunities. As we strive toward more equitable and responsible assessment, it becomes ever more crucial to examine the ripple effects of our work. Through rigorous scholarship, collaborative dialogue, and critical reflection, we can ensure that language assessment serves not only institutional efficiency but also human dignity and social good.

LTRC 2025 gathers researchers, practitioners, educators, policymakers, and students from diverse backgrounds—a key strength that fosters rich, cross-cultural learning. The LTRC offers a dynamic program of keynote talks, paper and poster sessions, symposia, work in progress, workshops, and other activities like SIGs, special sessions, and networking dinners. This year we accepted 95 research papers, 30 posters, 23 works in progress, and 8 symposiums. We also have four pre-conference workshops, four SIGs, and three special sessions.

Throughout the week, you will hear from the esteemed voices in the field, as well as rising scholars who are pushing the boundaries of how we think about assessment in the multicultural contexts from west to east. Whether your interests lie in test development, validation research, assessment for learning, technology-enhanced testing, or test fairness, and so on, we trust you will find sessions that spark your curiosity and deepen your understanding.

And of course, LTRC is more than an academic conference—it is also a celebration of community. It is where professional relationships are forged, where long-time colleagues reunite, and where early-career scholars meet mentors and collaborators who will support them throughout their journeys. We encourage you to take advantage of the many social and networking events planned, including our welcome reception, cultural night, and conference dinner. These moments of connection outside the formal program often become the most treasured memories of the Colloquium.

Bangkok, our host city, offers a vibrant backdrop for the conference. As Thailand's capital, it is a city of contrasts—ancient temples and modern skyscrapers, serene riverside scenes and bustling street markets. We hope you will find time to explore the cultural riches that the city has to offer. Whether you are visiting the majestic Grand Palace, riding a long-tail boat along the Chao Phraya River, shopping at local artisan markets, or sampling Thailand's world-renowned cuisine, there is something in Bangkok to delight every traveler.

The LTRC 2025 venue, **Chulalongkorn University**, is Thailand's oldest and most prestigious university, centrally located with easy access to major city attractions, accommodations, and public transport. We are proud to host you on our historic campus and to introduce you to the warmth, generosity, and intellectual spirit of our academic community. Our team has worked tirelessly to ensure that every aspect of the conference—from logistics and programming to hospitality and safety—meets the highest standards. We are also offering special support to first-time presenters and attendees to help you feel welcome and confident in participating fully in the Colloquium.

In closing, we want to express our sincere gratitude to the many individuals and organizations who have made this Colloquium possible. Our thanks go to the International Language Testing Association (ILTA) for their ongoing support and leadership. Special thanks also go to our platinum sponsors (Cambridge University Press & Assessment, Chulalongkorn University, Duolingo English Test, Eiken Foundation of Japan, IELTS, and Pearson English Language Learning), our gold sponsors (British Council English & Exams, Goethe-Institut, LTTC [Language Training and Testing Center]), our silver sponsor (Michigan Language Assessment), and our bronze sponsors (Lancaster University and WIDA) who also make this Colloquium possible. We also thank the reviewers, local committee members, volunteers, and Chulalongkorn University staff, whose dedication has brought this event to life. Finally, a heartfelt thank you goes to our local team in Bangkok — Wutthiphong, Mintra, Chariya, Raveewan, Chanisara, and Kornrawan—your behind-the-scenes work, unwavering commitment, and wholehearted effort truly made this Colloquium a welcoming and memorable experience for all.

Above all, we thank you, our participants, for bringing your curiosity, your scholarship, and your passion to this gathering. It is your presence that makes LTRC not just a conference, but an academic living and loving community of practice and inquiry. We hope that your time at LTRC 2025 will leave you feeling intellectually enriched, professionally inspired, and personally renewed.

Warmest regards,

Jirada, Sun, Punchalee

Chair and Co-Chairs
LTRC 2025 Bangkok, Thailand

Conference Organization

Organizing Committee

Chanisara Tangkijmongkol
Chariya Prapobratanakul
Jirada Wudthayagorn
Mintra Puripunyavanich
Punchalee Wasanasomsithi
Raveewan Viengsang
Sun-Young Shin
Wutthiphong Laoriandee

Student Assistants

Aye Thwe Han
Chudakarn Pakarasang
Garngullaya Chimpleewanasom
Itsra Namtapee
Kamonchanok Chaipak
Kris Kisawadkorn
Lin Chen
Lu Geng Suen
Metira Phatinuwat
Mink Sawasdeepon
Nang Phong Noan
Nur Atiqah Binte Surya Akmaja
Poonyawee Navetra
Sumi Brown
Thet Myat Noe Aung

Logo

Narasak Sirikanjanawong

Cover and Print Design

Wachirawit Pattong

Secretary

Kornrawan Assavasuwan

Abstract Reviewers

Aaron Batty
Alan Urmston
Alistair Van Moere
Antony Kunnan
Apichat Khamboonruang
April Ginther
Ardeshir Geranpayeh
Ari Huhta
Atta Gebril
Barry O'sullivan
Bart Deygers
Benjamin Kremmel
Beverly Baker
Brigita Seguis
Carol Chapelle
Carolyn Turner
Cecillia Zhao
Chatraporn Piamsai
Ching-Ni Hsieh
Claudia Harsch
Constant Leung
Daniel Isbell
David Macgregor
Deborah Crusan
Dina Tsagari
Doris Moser-Froetscher
Elvis Wagner
Erik Voss
Fumiyo Nakatsuhara
Gad Lim
Gary Ockey
Geoff LaFlair
Guoxing Yu
Heike Neumann
Jason Fan
Jee Wha Dakin
Jeff Stewart
Jill Wigglesworth
Jing Xu
John Read
Jonathan Schmidgall
Kathrin Eberharter
Khaled Barkaoui
Khanh-Duc Kuttig
Lorena Liosa
Luke Harding
Lynda Taylor
Margaret Malone-Hall
Mark Chapman
Mikyung Wolf
Napat Jitpaisarnwattana
Nathan Carr
Noriko Iwashita
Norris John
Olena Rossi
Patharaorn Patharakorn
Rei Koizumi
Ruslan Suvorov
Ryan Lidster
Saerhim Oh
Sahbi Hirdi,
Sally O'Hagan
Sari Luoma
Senyung Lee
Shahrzad Saif
Slobodanka Dimova
Sungworn Ngudgratoke
Susy Macqueen
Sutthirak Sapsirin
Talia Isaacs
Tanyaporn Arya
Tineke Brunfaut
Tony Green
Ute Knoch
Vahid Aryadoust
Veronika Timpe-Laughlin
Vivian Berry
Xun Yan
Yan Jin
Yasuyo Sawaki
Ying Zheng
Yo In'nami
Yong-Won Lee
Yunwen Su

ILTA Executive Board and Committee Members

ILTA Executive Board 2025

President: Luke Harding, Lancaster University
Vice President: Lia Plakans, University of Iowa
Secretary: Elvis Wagner, Temple University
Treasurer: Beverly Baker, University of Ottawa
Past President: Claudia Harsch, University of Bremen

Members at Large

Kathrin Eberharter, University of Innsbruck
He Lianzhen, Zhejiang University
Gad Lim, Cambridge Boxhill Language Assessment
Bart Deygers, Ghent University
Erik Voss, Columbia University (Communications Committee Chair)

ILTA Staff

Valerie Smith
Dorrian Regan

LTRC 2025 Chair and Co-Chair

Jirada Wudthayagorn
Sun-Young Shin
Punchalee Wasanasomsithi

ILTA Nominating Committee

Chair: Dylan Burton, Georgia State University
Heidi Banerjee, PSI Services LLC
Haeun (Hannah) Kim, University of Melbourne
John Read, University of Auckland

Graduate Student Assembly

Chair: Monique Yoder, Michigan State University
Vice Chair: Chenyang Zhang, University of Melbourne
Communications Chair: Duong (Zoe) Nguyen, Iowa State University

SIG Chairs

Automated Language Assessment (ALASIG)
Xiaoming Xi, Hong Kong Examinations and Assessment Authority
Jing Xu, Cambridge University Press & Assessment

Integrated Assessment (IASIG)
GoMee Park, University of Texas Rio Grand Valley
Renka Ohta, Educational Testing Service
Ping-Lin Chuang, Duolingo

Language Assessment Literacy (LALSIG)
Elsa Fernanda Gonzalez, Universidad Autonoma de Tamaulipas
Gladys Quevedo-Camargo, Universidade de Brasilia

Language Assessment in Aviation (LAASIG)
Natalia Andrade, University of Campinas
Ana Lígia Silva, São Paulo State University-UNESP
Angela Garcia, Carleton University

Language Assessment for Young Learners (Young Learners SIG)
Mark Chapman, WIDA
Veronika Timpe-Laughlin, Educational Testing Service
Jeanne Beck, Iowa State University

Test-taker Insights in Language Assessment (TILASIG)
Andy Jiahao Liu, University of Arizona
Ray Jui-Teng Liao, National Taiwan Ocean University

Awards and Grants Committees

ILTA Best Article Award

Chair: Aaron Olaf Batty, Keio University
Vahid Aryadoust, National Institute of Education, Nanyang Technological University
Mikyung Kim Wolf, Educational Testing Service
Susy Macqueen, Australian National University
Fumiyo Nakatsuhara, University of Bedfordshire

Robert Lado Memorial Award (to be awarded at LTRC)

Chair: Dylan Burton, Georgia State University
Slobodanka Dimova, University of Copenhagen
Kellie Frost, University of Melbourne
Atta Gebril, American University of Cairo
Noriko Iwashita, University of Queensland
Ray Liao, National Taiwan Ocean University

Cambridge / ILTA Distinguished Achievement Award

Chair: Antony Kunnan, City University of Macau
Carol Chapelle, Iowa State University
Claudia Harsch, University of Bremen
Nick Saville, Cambridge University Press & Assessment

Caroline Clapham IELTS Masters Award

Committee members: Dr Emma Bruce, Dr Hye-won Lee, Dr Nick Glasson, and Dr Tony Clark

ILTA / Duolingo Collaboration, Outreach and Professional Development Grant

Chair: Luke Harding, Lancaster University
Kathrin Eberharter, University of Innsbruck
Geoff LaFlair, Duolingo
Salomé Villa Larenas, Universidad Alberto Hurtado

TOEFL / ILTA Student Travel Grants

Chair: Lia Plakans, University of Iowa
Kathrin Eberharter, University of Innsbruck
Bart Deygers, Ghent University

Award Winners

Cambridge / ILTA Distinguished Achievement Award

Lynda Taylor, University of Bedfordshire

TOEFL / ILTA Student Travel Grants

Zhuohan Chen, Oxford University
Shanshan He, University of Western Ontario
I-Chun Hsiao, University of Iowa
Leyla Karatay, Iowa State University
Yoonseo Kim, University of Hawaiʻi at Mānoa
Xiaoxiao Kong, University of Melbourne
Pong-ampai Kongcharoen, Northern Arizona University
Meng-Hsun Lee, University of Toronto
Yiwei Qin, University of Ottawa
Burak Senel, Iowa State University
Xingcheng Wang, University of Melbourne
Shengkai Yin, University of Melbourne/Shanghai Jiao Tong University
Daniel Yu-Sheng Chang, University of Bristol
Yang Zhang, Zhejiang University

ILTA Best Article Award 2023

A closer look at a marginalized test method: Self-assessment as a measure of speaking proficiency
Paula Winke, Michigan State University
Xiaowan Zhang, MetaMetrics
Steven J. Pierce, Michigan State University
Published in: *Studies in Second Language Acquisition, 45*(2), 416–441.
https://doi.org/10.1017/S0272263122000079

ILTA / Duolingo Collaboration, Outreach and Professional Development Grant

Advancing Learning-oriented Language Assessment Practices Through Collaboration: Engaging Teachers, Policymakers, and Curriculum Developers
Alla Baksh Mohamed Ayub Khan, Universiti Sains Malaysia
Faridah Juraime, Malaysian Examinations Syndicate, Ministry of Education
Liying Cheng, City University of Macau

The SEMAPLE/EPIC/LAALTA 2025 Conference
Isadora Teixeira Moraes, Federal University of Minas Gerais

Inaugural Teaching and Testing (T&T) Conference in ELT
Jayakaran Mukundan, Taylor's University Malaysia

Examining the Comparability of ChatGPT with Human Raters in Assessing Intercultural Competence in the East Asian Context (Conference: NéALA 2025 "The Natural and the Artificial in Applied Linguistics: A time of paradoxes", University of Lorraine, France).
Dr Weejeong Jeong, Indiana University Bloomington

Samuel J. Messick Memorial Lecture Award (Sponsored by ETS)

Lyle F. Bachman

Alan Davies Lecture Award (Sponsored by British Council)

Carsten Roever

Robert Lado Memorial Award

To be announced at the LTRC Banquet

2025 Jacqueline Ross TOEFL Dissertation Award 2025

Dr. Nicholas Glasson
Left to their Own Devices: Exploring Interactional Practices in an Online Group Speaking Task
Supervisors: Nahal Khabbazbashi & Fumiyo Nakatsuhara

Caroline Clapham IELTS Masters Award 2024

Rahmad Adi Wijaya

Duolingo 2024 Doctoral Dissertation Awards

Wiktoria Allan, Lancaster University
Carla Consolini, University of Oregon
Jieun Kim, University of Hawaiʻi at Mānoa
Valeriia Koval, University of Bremen
Sebnem Kurt, Iowa State University
Jennifer Kay Morris, Lancaster University
Xue Nan, Beijing Language and Culture University
Chenyang Zhang, University of Melbourne
(See more details on pp. 177-178)

TIRF 2024 Doctoral Dissertation Grant Awardees in Language Assessment

Shishi Zhang, University College London
Dissertation Title: *Assessing Second Language Pragmatic Competence for Intercultural Communication: The Case of Pre-sessional Students in UK Higher Education*
Funding Co-Sponsor: British Council and TIRF
Advisor: Professor Talia Isaacs

Xiaoxiao Kong, University of Melbourne
Dissertation Title: *Exploring the Language and Communication Demands of Early Childhood and School Teachers in Australia: Implications for Language Assessment for Teacher Registration*
Funding Co-Sponsor: Cambridge and TIRF
Advisors: Associate Professor Jason Fan and Professor Ute Knoch

TOEFL 2025 Grants for Doctoral Research in Language Assessment

Chia-Hsin Yin, The Ohio State University
Fatima Montero, University of Maryland, College Park
Jieun Kim, University of Hawaii at Manoa
Marina Melo Cialdini, Sao Paulo State University (UNESP)
Mateus Souza, University of Limerick
Meng-Hsun (Hunter) Lee, University of Toronto
Sebnem Kurt, Iowa State University
Shengkai Yin, Shanghai Jiao Tong University and The University of Melbourne
Yichen Jia, Nanyang Technological University

Language Testing 2024 Reviewer of the Year Award

Chao Han, National University of Singapore, Singapore
Trevor Holster, Fukuoka University, Japan

Housekeeping and Important Information

If any questions come up, please feel free to ask our staff at the **registration desk** at any time.

Important Road Safety Tips
Please be cautious when crossing roads in Bangkok. Traffic can be heavy and unpredictable, and drivers may not always stop for pedestrians, even at crosswalks. We recommend using pedestrian bridges or designated crossings whenever possible, and always checking carefully in both directions before crossing.

Prayer Room

For those who wish to pray or have a quiet moment, the prayer room is located in **Building 6**. Please feel free to use it at your convenience.

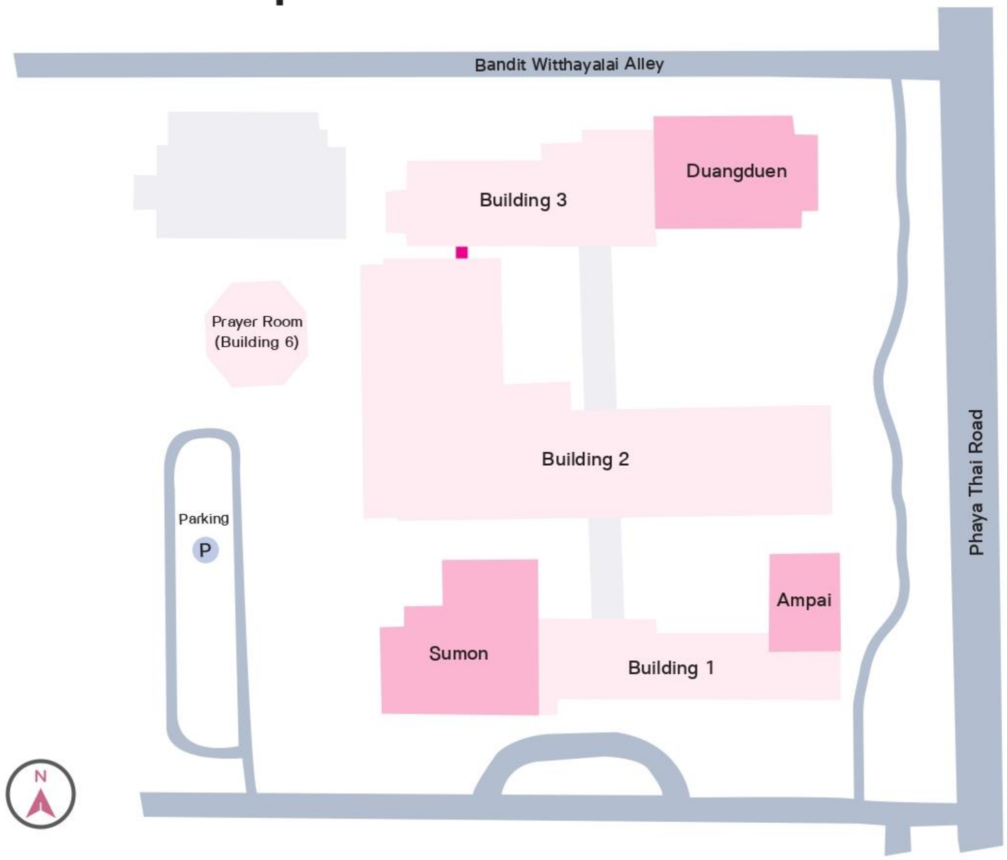
Wi-Fi Access Information
The provided Wi-Fi username and password can be used throughout all days of the conference during your time at Chulalongkorn University. We encourage you to take a photo of this information in case it is lost.

Further information can be found on the conference website:
https://www.culi.chula.ac.th/2025LTRC/index.html

If you want, you can use the #LTRC2025 for any social media postings.

Maps and Floor Plans

Venue map



Registration

Foyer in front of Poonsapaya, 3rd floor, Building 2

Opening

Poonsapaya, 3rd floor, Building 2

Workshops

Ampai, 1st floor, Building 1
Phramingkwan, 2nd floor, Building 2
Room 405 and 409, 4th floor, Building 3

Symposia and Research Papers

Ampai, 1st floor, Building 1
Phramingkwan, 2nd floor, Building 2
Poonsapaya, 3rd floor, Building 2
Duangduen, 1st floor, Building 3

Posters

Room 104, 1st floor, Building 1

Works-in-Progress

Sumon, 1st floor, Building 1

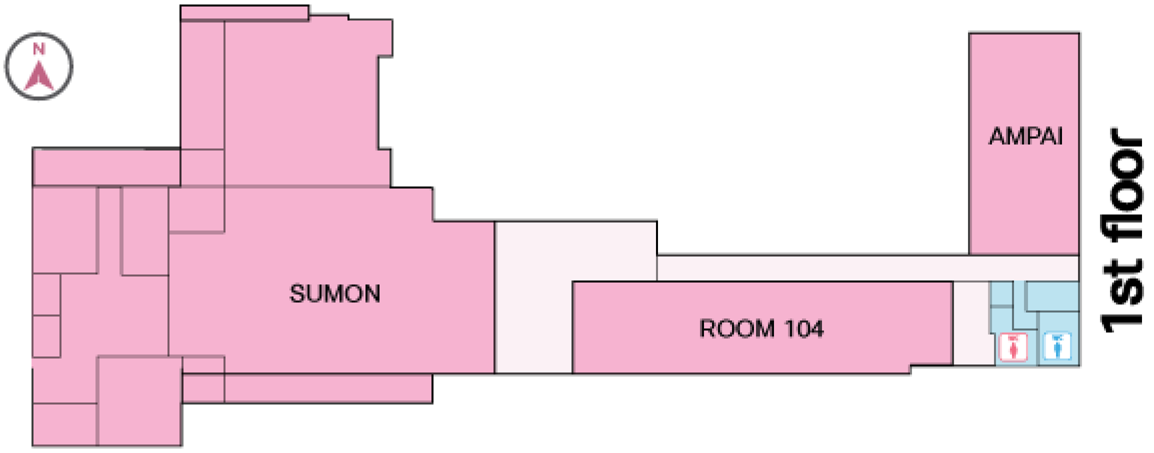
Welcome Reception, Coffee Break and Lunch

Sumon, 1st floor, Building 1

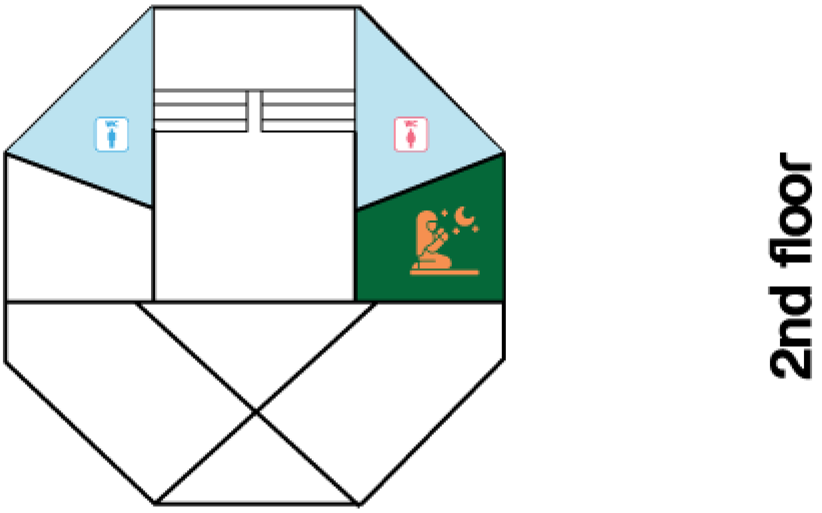
Building 1

Legend

🟦 Rest room 🟩 Prayer room



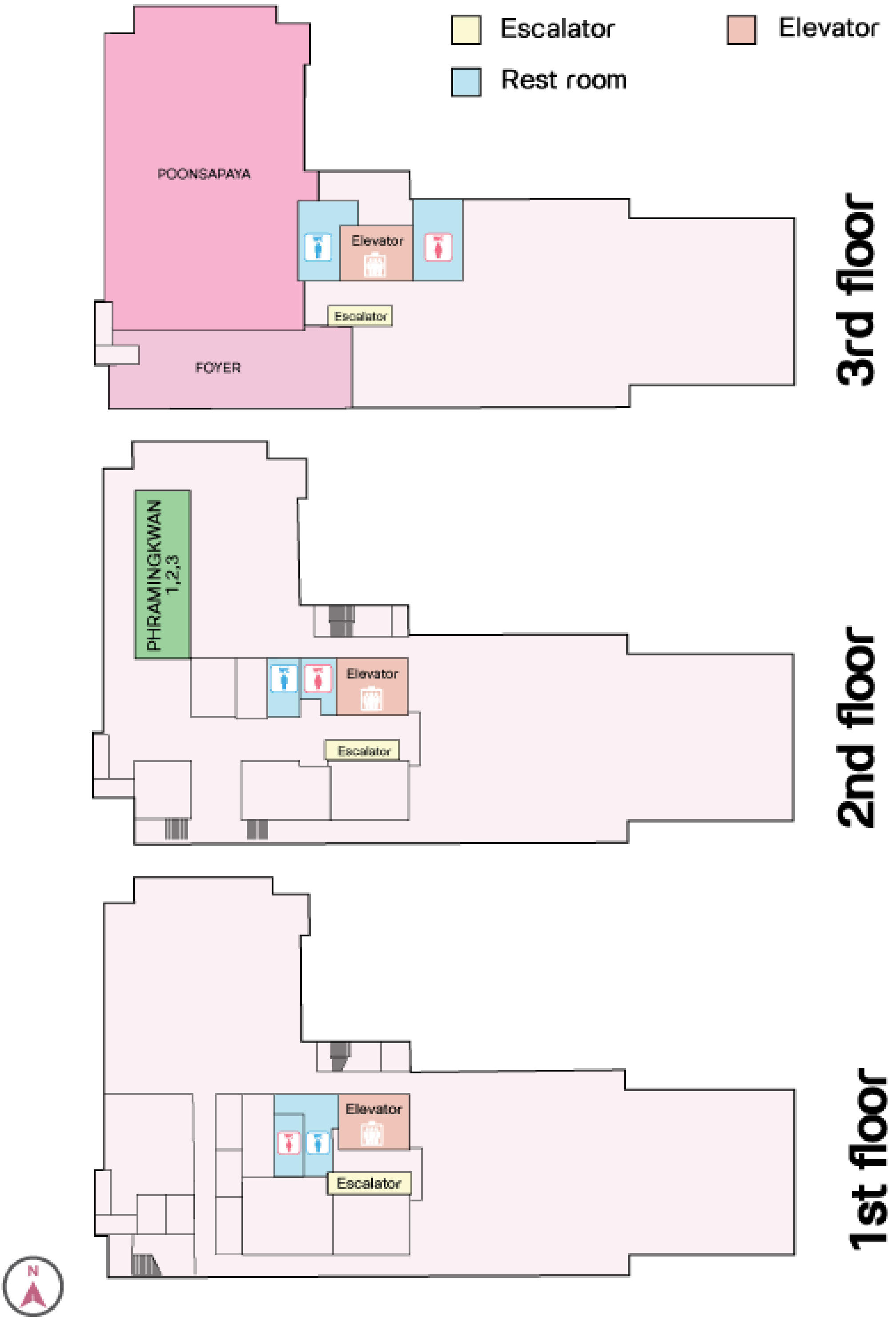
Building 6



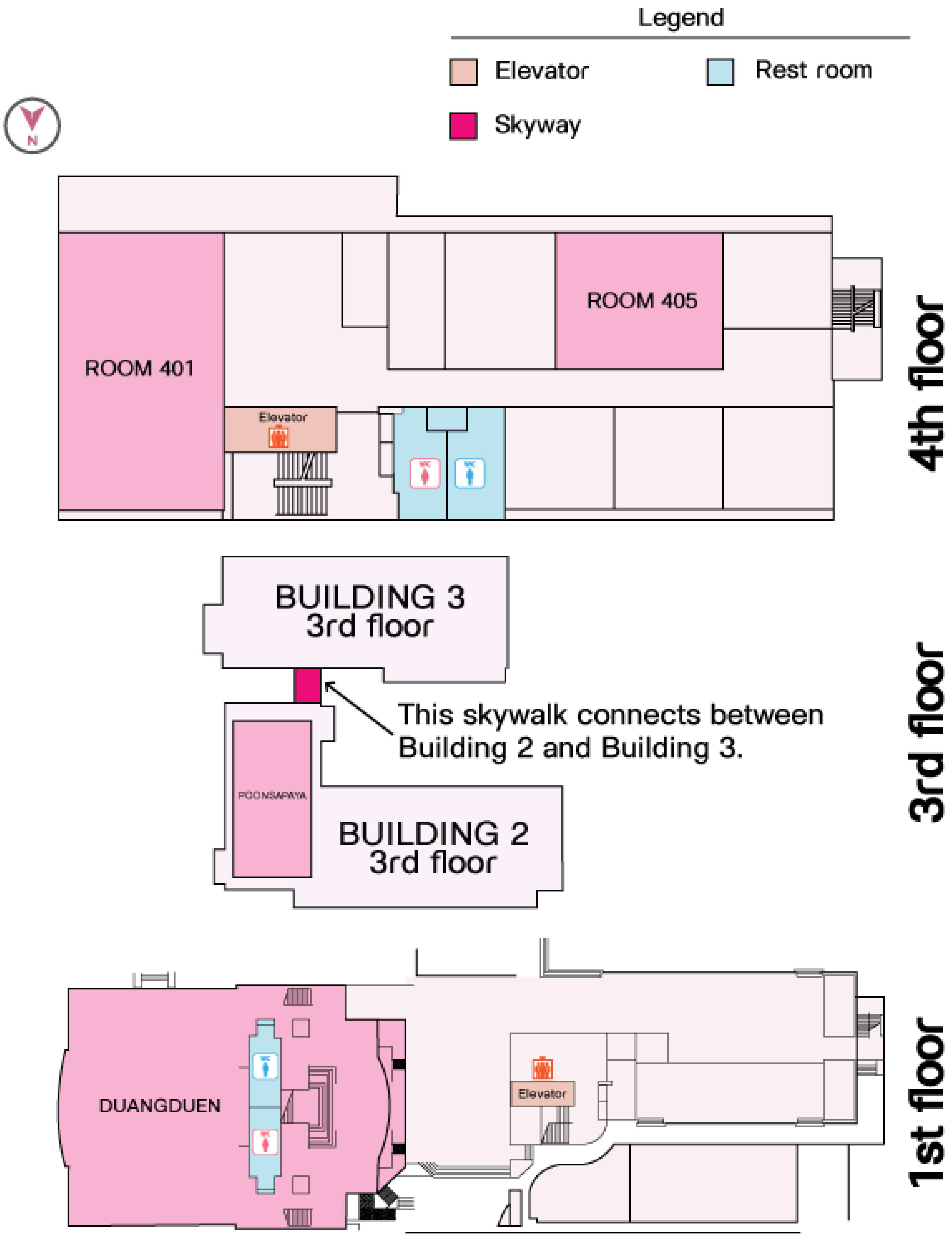
Building 2

Legend

🟨	Escalator	🟧	Elevator
🟦	Rest room		



Building 3


Legend
Elevator
Rest room
Skyway
N
ROOM 401
ROOM 405
Elevator
WC
WC
4th floor
BUILDING 3
3rd floor
This skywalk connects between Building 2 and Building 3.
POONSAPAYA
BUILDING 2
3rd floor
3rd floor
DUANGDUEN
WC
WC
Elevator
1st floor

Reflection

Session: ____

What did I learn?

____

____

____

____

____

Notes:

____

____

____

____

____

____

Session: ____

What did I learn?

____

____

____

____

____

Notes:

____

____

____

____

____

____

Reflection

Session: ____

What did I learn?

____

____

____

____

____

Notes:

____

____

____

____

____

____

Session: ____

What did I learn?

____

____

____

____

____

Notes:

____

____

____

____

____

____

Reflection

Session: ____

What did I learn?

____

____

____

____

____

Notes:

____

____

____

____

____

____

Session: ____

What did I learn?

____

____

____

____

____

Notes:

____

____

____

____

____

____

Reflection

Session: ______________________________

What did I learn?

Notes:

Session: ______________________________

What did I learn?

Notes:

Plenary Sessions

Samuel J. Messick Memorial Lecture

Sponsored by Educational Testing Service

A Personal Odyssey through Language Assessment and the Fields of Validity

Prof. Lyle F. Bachman, University of California, Los Angeles (UCLA)

Friday, June 6, 2025, 9:30am to 10:30am Location: Poonsapaya

In this presentation, I trace the course of my development as a language tester over the past half century, relating this to the evolution of validity theory and validation practice in the fields of language testing and educational measurement, during the same period. My initiation to language testing occurred in the 1970's, in Bangkok, when I unexpectedly became involved in the development of a placement and achievement test for students in an intensive English program. Over the next 50 years or so I participated in the development of and served as a consultant to a wide range of language assessments. These differed in terms of the variety of constructs they were intended to assess, the kinds of assessment tasks that were employed, and the uses for which scored-based interpretations were intended. These assessments also varied greatly in terms of assessment practice, including the amounts and kinds of research that were conducted in in support of test development and use.

Over this period, I also witnessed and to some extent contributed to the evolution of theoretical frameworks of validity and validation practices. I briefly describe these developments in validity theory and validation practice from around the mid-70's to the present. What has struck me most were the differences, over time, between conceptual frameworks of validity, on the one hand, and validation practices, on the other. I will argue that although theoretical "models" of validity have come and gone, the kinds of evidence collected to support the intended uses of assessments have remained virtually unchanged over this period. I will further argue that this disparity between validity theories and validation practice is not a problem for the field. This is because in the *real world*, where the uses of a given test affect real people, institutions, and society at large, *validation practice* is what counts. Furthermore, current validation practice in the field is quite rigorous, with the large-scale language test developers conducting appropriate research to support the intended score-based interpretations.

Nevertheless, it is my view that language testing institutions' validation practices do not include *routinely* collecting *systematic* evidence to support the decisions that are made on the basis of scored-based interpretations, and the consequences of these decisions for stakeholders. Rather, what one finds in this regard are numerous independent studies into different kinds of consequences, particularly the impact on instructional practices, and for different groups of stakeholders in different sociocultural settings. Furthermore, I have found that these studies are often not easily accessible. Finally, I suggest that ILTA can play a role in encouraging and assisting language assessment developers to create a rationale and a framework for systematically collecting and making accessible evidence to support the uses—decisions and consequences—for which their assessments are intended.

Lyle F. Bachman is Professor Emeritus of Applied Linguistics at the University of California, Los Angeles (UCLA). He is a Past President of the American Association for Applied Linguistics and of the International Language Testing Association. He has received numerous awards for his research and service to the profession, including the TESOL/Newbury House Award for Outstanding Research, the Modern Language Association of America Kenneth Mildenberger Award for outstanding research publication, the Sage/International Language Testing Association award for the best book published in language testing, the Lifetime Achievement Award from the International Language Testing Association, and the Distinguished Scholarship and Service Award from the American Association for Applied Linguistics. He has published numerous articles and books in language testing and other areas of Applied Linguistics, including *Fundamental Considerations in Language Testing*, and *Language Testing in Practice and Language Assessment in Practice: Developing Language Assessments and Justifying their Use in the Real World* (with Adrian Palmer), *Interfaces between Second Language Acquisition and Language Testing Research* (Co-edited with Andrew Cohen), *Statistical Analyses for Language Assessment*, and *Language Assessment for Classroom Teachers* (with Barbara Damböck). He has served on several committees of the National Research Council and has served as a member of the Board on Testing and Assessment, a standing board of the National Academies of Science, the Board of Trustees of the Center for Applied Linguistics, and the Technical Advisory Committee of the WIDA ACCESS test. He has also served as a consultant in language testing research projects and in developing language assessments for universities and government agencies around the world. His current research interests include validation theory and validation practice, classroom assessment, and epistemological issues in Applied Linguistics research.

Alan Davies Lecture

Sponsored by British Council

Assessment of Social Language Use: from Pragmatics to Interactional Competence

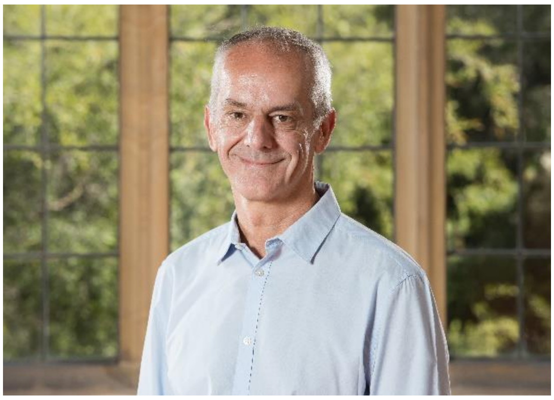
Prof. Carsten Roever, University of Melbourne

Saturday, June 7, 2025, 11:00am to 12:00pm Location: Poonsapaya

As early as 1987, Schegloff called interaction "the bedrock of social life – the primordial site of sociality" (p. 102). Participation in social life is the ultimate goal of learning a language, and learners' ability to use language in social settings has been conceptualized as "pragmatic competence" from a speech act pragmatics perspective (Leech, 1983) or "interactional competence" (IC) from a conversation analysis perspective (Schegloff, 2007). Developmental and assessment work on pragmatic-interactional abilities has grown exponentially and can be traced back as far as Farhady (1980), who investigated "functional language competence". Work informed by L2 pragmatics ranges from early studies by Shimazu (1989) and Hudson, Detmer and Brown (1995) to later ones by Timpe-Laughlin (2017) and Ellis et al. (2024).

This line of work has demonstrated that the ability to use language appropriately to social context can be assessed comprehensively but it remained focused on explicit knowledge and did not assess interactional abilities. By contrast, work under the IC paradigm has taken a holistic view of language use in social settings for social purposes. Definitions and frameworks like Pekarek Doehler's (2021), Young's (2019) and Taylor and Galaczi's (2018) variously feature micro-interactional aspects, the larger sociocultural context and the immediate social setting. Pioneering test development work by Youn (2013, 2015) demonstrated how social talk can be simulated in a testing setting and rating criteria developed bottom-up, while Galaczi's (2014) study showed that interactional features can be found and assessed in test taker talk elicited for proficiency assessment. Later work has delved deeper into ratable features (May et al., 2020) and also integrated the fulfillment of social roles from the perspective of membership categorization analysis (Dai & Davey, 2024).

Despite the lively research on the assessment of pragmatics and IC, no large-scale test incorporates assessment of social language use, be that from a pragmatics or IC perspective. This is likely due to practical difficulty in terms of test administration and scoring of measuring contextualized and interactive social language use, but also because it is not entirely clear whether measurement of IC offers additional, previously uncaptured variance relevant to score use. I will argue that measurement of IC is beneficial in terms of construct coverage, measures abilities not covered by proficiency measures, enables more informative scores, and can narrow the credibility gap between intended meanings of test scores and score users' interpretation of them. To address the practicality issue, AI delivered and scored interactions are an enticing option, situated monologs hold some promise, and existing measures can be upgraded to tap IC more directly. However, all these approaches come with caveats and require careful validation. I will lay out a research agenda for IC assessment development in the future.



Carsten Roever is a Professor in Applied Linguistics at the University of Melbourne. He holds a Ph.D. in Second Language Acquisition from the University of Hawai'i. Carsten's main research interests lie at the intersection of second language learning, interactional competence, and language testing, and he is also interested in conversation analysis and quantitative research methods. His most recent book is “Teaching and Testing Second Language Pragmatics and Interaction: A Practical Guide” published by Routledge in 2022.

Cambridge/ILTA Distinguished Achievement Award

Sponsored by Cambridge University Press & Assessment / ILTA

‘Navigating by the Stars’: Reflections from a Lifelong Journey Towards Language Assessment Literacy and Professional Competence

Prof. Lynda Taylor, University of Bedfordshire

Sunday, June 8, 2025, 2:30pm to 3:30pm Location: Poonsapaya

Language assessment literacy (LAL) is understood to have its roots in general educational contexts dating back to the early 1990s (e.g., Stiggins, 1991), but the concept was enthusiastically taken up in language assessment during the 2000s as part of the growing professionalisation of the field. This development built in part upon important theoretical and empirical work on test washback, impact and ethics conducted during the 1990s.

The concept of LAL evolved over time as language testing and assessment (LTA) matured as an academic field, embracing an expanding community of researchers and professional practitioners worldwide. Both researchers and practitioners explored a range of constituent elements that were hypothesised to make up the LAL ‘construct’, and they speculated on how the process of learning, competence-building and professional development in language assessment takes place (e.g., Malone, 2013; Baker & Riches, 2018; Kremmel & Harding, 2020).

The nature and variety of resources available to support LAL steadily increased, informed by the lived experience of LTA researchers and practitioners working in different contexts around the world, each with its own set of considerations and constraints – historical, educational, socio-political and lingua-cultural. Over the past 25-30 years, LAL has come to be seen as a core component of test development and validation activity, essential not only for test developers and teachers but for other test stakeholder groups, such as education policymakers, university admissions staff, prospective employers, test takers and their parents/carers. The need for relevant levels of theoretical knowledge and practical expertise is strongly felt in a variety of societal contexts, not only educational settings but also in recruitment, healthcare, migration and citizenship.

In this lecture I propose to look back at the trajectory of LAL development in our professional field since the 1990s. My aim will be: i) to reflect upon the journey travelled so far; ii) to consider where we find ourselves now and how the current landscape looks; and iii) to speculate where we might need to travel in the future. I will highlight some of the key developments relating to LAL over that period, including significant contributions made by both researchers and practitioners that helped advance our thinking and deepen our understanding of the LAL construct. Drawing on recent research findings, I shall consider some critical aspects of LAL, including: the diversity of test stakeholder groups and their needs, beliefs and attitudes; the highly context-based nature of LAL; the value of collaborative engagement within and between stakeholder groups; and the importance of audience- and context-appropriate language and discourse. In reviewing the development and growth of LAL in our field, I will also share some personal reflections from my own professional journey over a lifetime career in language learning, teaching, and assessment. Together we shall reflect on what it means to become a competent and confident professional committed to creating and using language assessments that benefit both individuals and society as a whole.

Lynda Taylor is Visiting Professor at the Centre for Research in English Language Learning and Assessment (CRELLA) at the University of Bedfordshire, UK. She has worked for over 40 years in the field of language teaching, learning and assessment, first as an ELT teacher and teacher educator, later as a materials developer and assessment researcher, particularly with IELTS and the full range of Cambridge English language qualifications. Her research interests include speaking and writing assessment, testing accommodations for language learners with special needs and the development of language assessment literacy. She was formerly Assistant Research Director with Cambridge Assessment English and has advised on test development and validation projects around the world, including Europe, North America and Asia. She has presented invited plenaries, research papers and practical workshops internationally, and has published extensively in academic journals, including *Language Testing*, *Language Assessment Quarterly*, *Assessing Writing*, *Journal of English for Academic Purposes* and *ELT Journal*. She has contributed numerous chapters to encyclopedia and handbooks in the field, including the *Annual Review of Applied Linguistics*. She authored or edited many of the volumes in CUP's *Studies in Language Testing* series and in 2023 she guest-edited (with Jay Banerjee) a special issue of *Language Testing* on accommodations in language assessment. In 2020 Lynda was elected President of the UK Association for Language Testing and Assessment (UKALTA) and she is currently serving a second 3-year term in that role. In 2022 she was awarded Fellowship of the UK Academy of Social Sciences (AcSS). Her most recent publications include a pair of volumes, co-edited with Beverly Baker and published in 2024, on the topic of *Language Assessment Literacy and Competence* (CUP). Lynda gave the Alan Davies Lecture at LTRC 2024 in Innsbruck and was recently awarded the Cambridge/ILTA Distinguished Achievement Award for 2024, to be presented during LTRC Bangkok in June 2025.

Pre-Conference Workshops

Workshop A: Language Test Design and Alignment to the CEFR

Viphavee Vongpumivitch and Napat Jitpaisarnwattana

Day 1: Wednesday, June 4, 2025, 9:00 am to 4:00 pm Location: Ampai
Day 2: Thursday, June 5, 2025, 9:00 am to 4:00 pm Location: Ampai

Workshop Description

Ever since its publication in 2001, the Common European Framework of Reference for Languages (CEFR) has expanded its influence far beyond Europe, and Thailand is no exception. In January 2014, the Thai Ministry of Education announced that the CEFR would be adopted as a framework for English language education reform, impacting curriculum design, classroom teaching, teacher training, learning objectives, and testing (Office of the Basic Education Commission, Thailand Ministry of Education, 2014). Over the past decade, stakeholders in language education have become increasingly familiar with the CEFR's six reference levels – A1, A2, B1, B2, C1, and C2. Numerous domestic English tests now report results linked to these levels.

Nevertheless, some language educators and assessment practitioners continue to have questions regarding the CEFR: Is the CEFR itself a test? How can new tests be developed based on the CEFR? How do we translate the CEFR's can-do statements into concrete test items? How should we interpret CEFR-based test scores? To what extent can we trust the claims made by test developers that their test results are aligned with the CEFR?

This workshop, presented in Thai at the very first LTRC to be held in Bangkok, seeks to address these questions by drawing on official training resources provided by the Council of Europe, the Association of Languages Testers in Europe (ALTE), the European Association for Language Testing and Assessment (EALTA), the British Council, and the UK Association for Language Testing and Assessment (UKALTA). The two-day workshop will consist of four parts:

Part 1 will introduce the core components of the CEFR based on the CEFR Manual (Council of Europe, 2001) and the Companion Volume (Council of Europe, 2020). Key aspects of the CEFR, including its action-oriented approach, illustrative descriptive scales, and concepts regarding language assessment, will be covered. Through group activities, participants will explore the CEFR's four modes of communication: reception (listening and reading), production (speaking and writing), interaction (spoken and written), and mediation (facilitating communication between individuals or groups who cannot communicate directly).

Part 2 will cover the fundamental considerations in language testing – validity, reliability and the test development process (Council of Europe, 2011). Participants will then learn about the required steps involved in CEFR test alignment, as detailed in the *CEFR Alignment Handbook* (Figueras, Little, and O'Sullivan, 2022). These include test specification, standardization, standard setting, and validation. Due to time constraints, this part of the workshop will *describe* each stage of the CEFR alignment rather than engaging the participants in actual test development or alignment.

Part 3 will showcase examples of English tests developed based on the CEFR, primarily the Cambridge Main Suites of English Examinations (https://www.cambridgeenglish.org/exams-and-tests/qualifications/). Additionally, we will also discuss a local English proficiency test at a Thai university to provide insights into the issues involving score interpretation and use within a Thai university context.

Part 4 will address performance tasks in CEFR-based tests that require human ratings. Topics will include rater training, rater behaviors, and the integration of automated scoring systems, such as e-raters with human raters in the rating process. The workshop will conclude by exploring the potential of AI-powered technology to support local raters in consistent scoring and reduce measurement errors caused by raters' behaviors.

By the end of this two-day workshop, participants are expected to gain a comprehensive understanding of the CEFR and the necessary steps for CEFR test alignment. Equipped with this knowledge, they will be able to critically evaluate claims made by test developers about the alignment of their tests with the CEFR, and take the right steps towards developing a CEFR-based test.

References:

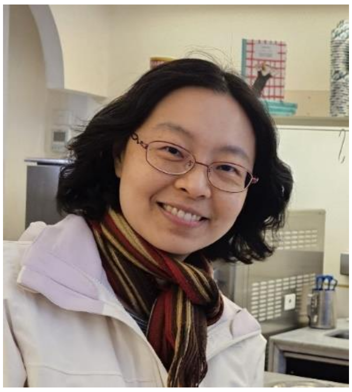
Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge University Press.

Council of Europe. (2011). *Manual for Language Test Development and Examining: For Use with the CEFR.* Association of Languages Testers in Europe (ALTE).

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment – Companion Volume.* Council of Europe Publishing.

Figueras, N., Little, D., & O'Sullivan, B. (Eds.) (2022). *Aligning Language Education with the CEFR: A Handbook.* British Council, UKALTA, EALTA & ALTE.

Office of the Basic Education Commission, Thailand Ministry of Education. (2014). *Guidelines Following the Ministry of Education's Announcement on the Policy for English Language Teaching Reform.* [Source in Thai] สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน กระทรวงศึกษาธิการ. แนวปฏิบัติ ตามประกาศกระทรวงศึกษาธิการ เรื่อง นโยบายการปฏิรูปการเรียนการสอนภาษาอังกฤษ. พฤษภาคม ๒๕๕๗.



Viphavee Vongpumivitch received her PhD in language assessment from University of California, Los Angeles. She has been teaching at the Department of Foreign Languages and Literature, National Tsing Hua University in Taiwan for the past twenty years. In 2015 she gave her first workshop on the CEFR to Thai language teaching and testing professionals while she was working as a visiting scholar at the Graduate Institute of Language and Communication, National Institute of Development Administration (NIDA). Since then, she had conducted several workshops on CEFR for English teachers around Thailand. She was one of the plenary speakers at the 5th International Conference on Foreign Language Learning and Teaching (FLLT 2018) organized by the Language Institute, Thammasat University.

Napat Jitpaisarnwattana is a lecturer of English and Computer-assisted Language Learning at Silpakorn University, Thailand. He recently finished his Master of Studies focusing machine learning and automated assessment at Homerton College, University of Cambridge. He received his PhD in Applied Linguistics from KMUTT and an MSc in Teaching English Language in University Settings from Oxford University. He is editor of Malaysian Journal of ELT Research and associate editor of rEFLections. His research interests include Computer-assisted Language Learning (CALL), Language MOOCs, the Internet of things, digital wellbeing, digital literacies, learning analytics, machine learning, AI in language education, technology-mediated language assessment and learning-oriented assessment (LOA).

Workshop B: Publishing Your Research in the AI Era: Sharing the Road Traveled with Budding English as a Foreign or Second Language Scholars

Qin Xie

Day 1: Wednesday, June 4, 2025, 9:00 am to 4:00 pm Location: Phramingkwan

Workshop Description:

Publish or perish. Publishing scholarly research is becoming unavoidable for academics working in universities worldwide. Publishing research in English poses considerably more challenges for EFL scholars working in developing countries. Such challenges could be overwhelming for budding scholars or EFL postgraduate students.

In this workshop, we will address some of the challenges by involving experienced scholars to share their insights and experiences of publishing research in the field of language testing and assessment. We will also walk participants through the publication procedures, discuss with the participants the Dos and Don'ts, introduce some emerging Generative AI-based tools, and explore their potential to assist publication in an ethical and fair manner. Participants will learn how to use Gen-AI tools to produce mini-videos based on their own professional or scholarly outputs for broader dissemination of their research output.

The workshop will have the following components:

1. We will walk the participants through the common procedures of academic journal publication. Experienced and stellar researchers will be invited to share their insights and experiences on key stages of these procedures.
2. We will introduce several recent AI-based tools and demonstrate how they can facilitate EFL novice researchers on their publication journey, especially in consolidating scholarly source materials and enhancement of the English language.
3. We will showcase how researchers can digitize their professional resources and research outputs in the form of searchable online databases and short videos to achieve wider dissimilation and greater social impact

The workshop will involve many interactions and step-by-step, hands-on practices of the tools introduced. Workshop participants will be asked to

1. Work in small groups to discuss their publication ideas and potential challenges
2. Work with case demonstrations and examples from existing materials to ground the workshop in real-world issues and applications.
3. Work on tasks to trial and explore the AI tools introduced
4. Be guided step by step on how to produce a one-minute video showcasing their research.

We believe the workshop program with experience sharing, case demonstration, and practical hands-on training will allow participants to traverse from the conceptual to the practical, achieve a clearer perspective toward academic publication and ethical uses of AI tools, and learn to use these powerful tools to assist their publication not only in English journals and but also in the emerging form of digital publication. The latter could make their research more accessible and achieve greater social, educational, and academic impact.

Qin Xie is Associate Professor of Language Assessment and Education at the University of Macau, where she trains English language teachers, focusing on their language assessment literacy skills. She also supervises doctoral and master students and has worked with them to publish their first research in academic journals. Her research interests include test washback, validation, and diagnostic assessment, from which she has published over 35 research articles in high-impact international journals such as *Language Testing, Language Assessment Quarterly, Assessing Writing, Language Teaching, Systems, and Educational Psychology*. Currently, Qin serves *Language Assessment Quarterly* and *Journal of Asia TEFL* as an associate editor; she is also on the editorial board of several other indexed journals and has been a regular reviewer for a dozen of journals in applied linguistics and language assessment. Qin will invite several guest speakers (TBC) to facilitate the workshop by sharing their experiences and insights about journal publications.

Workshop C: An Applied Introduction to No-code Regression and Machine Learning for Language Assessment

Vahid Aryadoust

Day 2: Thursday, June 5, 2025, 9:00 am to 4:00 pm Location: Room 405

Workshop description:

As Artificial Intelligence (AI) continues to advance at a remarkable pace, the field of language assessment must look beyond traditional methods to explore the potential of AI and machine learning (ML). Embracing these technologies is essential for adapting to technological shifts and understanding their impact on assessment practices. In response to this need, I am proposing a one-day workshop designed to teach no-code regression and ML, using JASP, a free software with a user-friendly graphical interface. Specifically, the workshop aims to provide language assessment professionals with practical tools and insights to running multiple linear regression in two ways: the statistical method and ML-based method. Next, the ML algorithm in JASP is used to compare the results of regression modeling and ML analysis in terms of fit, accuracy, precision etc.

Some of the concepts and methods covered in this workshop include data preprocessing and preparation for regression analysis, understanding the principles behind multiple statistical and ML-based linear regressions, and their assumptions, conducting and interpreting multiple linear regression using traditional statistical methods, and exploring the fundamentals of ML algorithms as they apply to language assessment. By the end of the session, participants will gain a foundational understanding of integrating regression and ML into language assessment analytics and develop the skills necessary to make informed methodological choices in their research and practice. At the end, limitations of no-code machine learning are discussed, and further considerations are explored for when and how to incorporate code-based machine learning approaches for greater flexibility and customization.

The primary participants for this workshop include language educators and assessment professionals, graduate students, and researchers interested in incorporating no-code machine learning into their practice. Participants must install JASP on their personal laptops before attending the workshop.

Vahid Aryadoust is Associate Professor of language assessment at Nanyang Technological University in Singapore, with additional roles as Honorary Associate Professor at UCL, London, and Visiting Professor at Xi'an Jiaotong University, China. His research interests include generative AI in language assessment, meta-analysis, and sensor technologies such as eye tracking, brain imaging, and GSR. Dr. Aryadoust has published extensively in reputable journals and authored several books and book chapters with leading publishers. He has led numerous assessment research projects funded by educational institutions in Singapore, the US, the UK, and Canada. He serves on the Advisory Board of various international journals and was awarded the Intercontinental Academia Fellowship (2018–2019). Dr. Aryadoust has received ILTA's Best Article Award (2024) for his research on sensor technologies in measuring cognitive load in listening assessment, NIE's Teaching Excellence Award, NTU's Teaching Award (School), among others. He is a proponent of knowledge-sharing and equity in education, exemplified by his YouTube channel "Statistics and Theory," which won the John Cheung Social Media Award in 2020 for its innovative use of social media.

Workshop D: Impact Evaluation: Global Perspectives and Best Practices

Micheline Chalhoub-Deville, Hanan Khalifa, and Eunice Jang

Day 2: Thursday, June 5, 2025, 9:00 am to 4:00 pm Location: Phramingkwan

Workshop description:

Impact evaluation is a vital process that helps organizations and policymakers measure the effectiveness of programs and interventions. This full-day workshop is dedicated to enhancing participants' understanding of impact evaluation emphasizing global and local perspectives and best practices that inform effective evaluation strategies.

The workshop commences with an introduction to impact evaluation, outlining various types of evaluations, models and frameworks and discussing the intersection between testing and evaluation. Participants will explore the goals of impact evaluation, which include assessing program effectiveness, informing policy decisions, and enhancing accountability. An initial hands-on activity will encourage participants to share their experiences with evaluations thus fostering a collaborative learning environment from the outset (Rossi et al., 2004).

Following this introduction, the focus will shift to defining evaluation outcomes and developing indicators. Participants will learn how to articulate clear objectives for their evaluations and differentiate between short and long-term outcomes. The workshop will emphasize the importance of developing indicators that are specific, measurable, appropriate, relevant and timebound. In a paired exercise, attendees will define outcomes and indicators for a hypothetical program, applying the concepts learned (Kusek & Rist, 2004).

The next session will cover evaluation design methods, and participants will learn how to select the most suitable design based on factors such as context, resources, and timelines. In small groups, they will then choose an appropriate evaluation design for a given scenario and justify their selection, thus reinforcing their understanding of design principles through practical application (Chen, 2015).

Participants will explore different data collection methods, including surveys, interviews, focus groups, and observations and discuss ways to enhance the reliability and validity of the instruments. They will create a short survey or interview guide based on the outcomes and indicators developed earlier and practice conducting interviews in pairs or small groups, allowing them to experience data collection firsthand (Fowler, 2014).
Next, participants will engage in a group case study exercise, where they will first review a successful evaluation and then develop an evaluation plan based on a given scenario, outlining objectives, indicators, design, and data collection methods. Each group will present their plan to the larger audience for constructive feedback.

The workshop will conclude with a discussion on best practices and ethical considerations in impact evaluation, emphasizing informed consent, confidentiality, and stakeholder engagement. A final wrap-up session will allow participants to share insights gained throughout the day and discuss how they can apply these practices in their respective contexts.

The workshop is designed to empower participants with the knowledge and skills needed to conduct impactful evaluations that inform decision-making and enhance program effectiveness. By the end of the workshop, participants will possess an understanding of impact evaluation in language testing, equipped with the tools and strategies necessary to implement effective evaluations in their own contexts. They will develop a nuanced understanding of how context shapes evaluation processes and leave with actionable next steps to ensure their programs are not only effective but also responsive to the needs of all stakeholders, ultimately contributing to improved educational outcomes.

Detailed Outline

Session 1	**Introduction to Impact Evaluation** - Purpose & goals (e.g., assessing effectiveness of initiatives) - Different types of evaluation (outcome, output, etc.) - Different evaluation frameworks (e.g., Log frame, TOC, Kirkpatrick's 5 levels)
	Hands-On Activity: Participants briefly share experiences with evaluations they have conducted or been involved in.
Session 2	**Defining Evaluation Outcomes & Indicators** - Identifying goals & objectives - Developing short- and long-term outcomes - Developing SMART indicators
	Hands-On Activity: Participants work in pairs to define outcomes and indicators for a hypothetical program.
Session 3	**Evaluation Design Methods** - Overview - Selecting the right design
	Hands-On Activity: In small groups, participants select an appropriate evaluation design for a given scenario and justify their choice.
Session 4	**Data Collection Methods** - Overview of data collection methods - Instrument reliability, validity and feasibility - Sampling (discussion on probability and non-probability)
	Hands-On Activity: Participants create a short survey or interview guide based on the outcomes and indicators they developed earlier. They then practice conducting interviews in pairs or small groups.
Session 5	**Group Case Study Exercise** - Review a successful case study
	Hands-On Activity: In groups, participants develop an evaluation plan based on the case study, outlining objectives, indicators, design, and data collection methods. Each group presents their plan to the larger group for feedback.
Session 6	**Best Practices and Ethics** - Ethical considerations: informed consent, confidentiality ..etc - Engaging stakeholders and ensuring cultural sensitivity
Session 7	**Wrap up** - Participants share key insights from the day and discuss how they can apply these practices in their work. - Recap of Key Takeaways - Open Floor for Questions - Feedback and Evaluation of Workshop

Dr Hanan Khalifa is an international language testing & evaluation expert who developed national and international examinations; aligned curricula and tests to standards; and evaluated donor funded programs. For two decades, Hanan led Education Reform & Impact Evaluation work at Cambridge University Press & Assessment English and advised ministries of education globally. She is currently leading a Pan Arab initiative on developing a conjoint measurement scale for Arabic language for use in multilingual and multicultural communities.

As an academic and a Council of Europe expert, she has contributed to and led on several impactful work, e.g., the socio-cognitive model for Reading (Khalifa & Weir 2009), the New Companion volume of the CEFR (2018, 2020), Qatar Foundation Arabic Framework (2022). Dr Khalifa has won several international awards and presented and published on various language education topics.

Eunice Eunhee Jang is a Professor at the Ontario Institute for Studies in Education, University of Toronto. Specializing in diagnostic language assessment, AI applications, and program evaluation, Dr. Jang has led large-scale assessment and validation initiatives in collaboration with various stakeholders, such as the Steps to English Proficiency (STEP) language assessment framework for Ontario public schools. She has been actively engaged in strengthening language proficiency requirements through standard-setting studies for professional regulators to determine the language proficiency of internationally educated healthcare professionals for immigration purposes and licensing. Dr. Jang is the author of "Focus on Assessment," which offers educators insights into assessing K-12 English language learners. She is a recipient of the University of Toronto’s David Hunt Graduate Teaching Award, Tatsuoka Measurement Award, Jacqueline Ross TOEFL Dissertation Award, and IELTS MA Dissertation Award. Currently, Dr. Jang is leading the BalanceAI and APLUS projects, which critically examine the impact of advanced technological innovations on language and literacy assessments in K-12 and postsecondary education contexts.

Micheline Chalhoub-Deville holds a Bachelor's degree from the Lebanese American University and Master's and Ph.D. degrees from The Ohio State University. She currently serves as a Professor of Educational Research Methodology at the University of North Carolina at Greensboro (UNCG) where she teaches courses on language testing, validity, and research methodology. Prior to UNCG, she worked at the University of Minnesota and the University of Iowa. Her professional roles have also included positions such as Distinguished Visiting Professor at the American University in Cairo, Visiting Professor at the Lebanese American University, and UNCG Interim Associate Provost for Undergraduate Education. Her contributions to the field include publications, presentations, and consultations on topics like computer adaptive tests, K-12 academic English language assessment, admissions language exams, and validation. She has over 70 publications, including books, articles, and reports, has delivered more than 150 talks and workshops. Additionally, she has played key roles in securing and leading research and development programs, with a total funding exceeding $4 million. Her scholarship has been recognized through awards such as the ILTA Best Article Award, the Educational Testing Service—TOEFL Outstanding Young Scholar Award, the UNCG School of Education Outstanding Senior Scholar Award, and the national Center for Applied Linguistics Charles A. Ferguson Award for Outstanding Scholarship. Professor Chalhoub-Deville has served as President of the International Language Testing Association (ILTA). She is Founder and first President of the Mid-West Association of Language Testers (MwALT) and is a founding member of the British Council Assessment Advisory Board--APTIS, the Duolingo English Test (DET) Technical Advisory Board, and English3 Assessment Board. She is a former Chair of the TOEFL Committee of Examiners as well as a member of the TOEFL Policy Board. She has participated in editorial and governing boards, such as Language Assessment Quarterly, Language Testing, and the Center for Applied Linguistics. She has co-founded and directed the Coalition for Diversity in Language and Culture, the SOE Access & Equity Committee, and a research group focused on the testing and evaluation in educational accountability systems. She has been invited to serve on university accreditation teams in various countries and to participate in a United Nations Educational, Scientific, and Cultural Organization (UNESCO) First Experts' meeting.

Special Sessions

Playing the Peer Review Game: Tips on Academic Publishing from Seasoned Journal Editors

Moderator: Talia Isaacs (co-editor of Language Testing)
Panelists: Yan Jin (co-editor of Language Testing in Asia), Elvis Wagner (co-editor of Language Assessment Quarterly), Xun Yan (co-editor of Language Testing)

Day 3: Sunday, June 8, 2025, 11:00 am to 12:00 pm Location: Ampai

This session will be a panel discussion among four editors for three prominent language assessment journals, with Talia Isaacs acting as the moderator. The journal editors will first introduce their respective journals, and describe the role, audience, and unique features of their journals. The editors will also describe the in-house screening process for their journals, and discuss reasons why articles might be internally rejected and/or rejected after the first round of peer review. The discussion will conclude with a piece of advice about navigating the peer review process that they wish they had known when starting their own academic careers, as well as general suggestions for authors going through the journal submission process. This discussion will last about 30 minutes, and the final 30 minutes of the session will be devoted to a question-and-answer session between the audience and the journal editors. If time allows, the editors will also debate how journal publishing has evolved over the past five years, describe how they are managing those changes as editors, and share anecdotes about difficulties, challenges, or dilemmas they have faced.

Navigating the PhD Journey in Applied Linguistics and Language Assessment

Presenters: Yena Park, Haeun Kim, Ing Kongchareon
Moderator: Sun-Young Shin

Day 3: Sunday, June 8, 2025, 11:00 am to 12:00 pm Location: Poonsapaya

This special session brings together current and former PhD students to share their experiences navigating doctoral studies in Applied Linguistics, with a particular emphasis on language assessment and related fields. The session is designed to offer practical insights and peer-to-peer guidance for prospective and current graduate students who are pursuing or planning to pursue a PhD. Panelists will reflect on their academic journeys, highlighting key challenges and lessons learned throughout their programs. Topics to be discussed include time management strategies for balancing coursework, research, and personal commitments; essential skills and knowledge required for careers in academia and industry; and practical tips for succeeding in graduate school. The session will also cover advice on identifying and applying for competitive grants, fellowships, and internship opportunities that can enhance professional development. In addition, panelists will provide guidance on navigating the dissertation writing process, including topic selection, research design, and working with advisors. Strategies for publishing and presenting research during the PhD program will also be shared, with an emphasis on building an academic profile and engaging with the broader research community. This session will be especially valuable for graduate students seeking support and direction, as well as for faculty members interested in understanding and supporting the evolving needs of PhD students in the field of language assessment and applied linguistics more broadly.

Perspectives and Current Priorities - What are the Benefits of Collaboration between the Associations?

Presenters: Nick Saville (ALTE) - Quynh Nguyen (AALA) – Salomé Villa Larenas (LAALTA)
Moderator: Rama Matthew

Day 3: Sunday, June 8, 2025, 11:00 am to 12:00 pm Location: Duangduen

This special session serves as the first ever collaborative project between the three regional associations of language assessment from three continents: the Association of Language Testers in Europe (ALTE), the Asian Association for Language Assessment (AALA) and the Latin American Association for Language Testing and Assessment (LAALTA). This initiative aims to promote collaboration and networking in the field of testing and assessment not only within but also across regions, with a hope to set a model for similar inter-association fora in the future.

The presenters are the current presidents of ALTE and AALA and the first president of LAALTA (2019-2021), while the moderator has been an active member of ILTA and collaborated with members of these associations.

The session features a series of moderated presentations and joint discussions by the presenters to cover the following topics:
1. Benefits of membership in regional professional associations;
2. Regional trends and priorities in research on multilingual testing and assessment;
3. Multilingualism and World Englishes; and
4. Inter-association collaboration - why, what and how?

The session will end with an interactive discussion on the topics. The moderator will invite the audience to share their views on the presenters' concerted call for regional and inter-regional collaboration on testing and assessment research and practices.

Presenters:
Association of Language Testers in Europe (ALTE)
Nick Saville is Director of Thought Leadership, at Cambridge University Press & Assessment (English) and Secretary-General of the Association of Language Testers in Europe (ALTE)
Asian Association for Language Assessment (AALA)
Quynh Nguyen is the founding director of the Center for Testing and Assessment at the University of Languages and International Studies, Vietnam National University, Hanoi (ULIS-VNU) and currently the Director of the Department of Research, Science and Innovation at ULIS-VNU. She is also the President of the Asian Association for Language Assessment (AALA).
Latin American Association for Language Testing and Assessment (LAALTA)
Salomé Villa Larenas, Assistant Professor at Universidad Alberto Hurtado, Chile, is co-founder and first president of the Latin American Association for Language Testing and Assessment (LAALTA).
Moderator
Rama Mathew, formerly Department of Education, Delhi University, Delhi is an active member of ILTA and researches and writes on language assessment and teacher development.

Opening Symposium

East Meets West: A Multifaceted Interaction of Constructs and Contexts in Language Testing and Assessment

Chair(s): **Liying Cheng** (City University of Macau, Macau S.A.R. (China))
Discussant(s): **Micheline Chalhoub-Deville** (University of North Carolina at Greensboro)

Thursday, June 5, 2025, 5:00pm to 6:30pm Location: Poonsapaya

Our field of research reflects a multifaceted interaction of constructs and contexts since its first annual Language Testing Research Colloquium (LTRC) more than four decades ago in 1979. This inaugural meeting brought together a small group of applied linguists from various countries, each bringing unique backgrounds and research interests, though their perspectives were somewhat limited by their specific contexts, as they addressed the challenges and issues in language testing. This was the first LTRC and set the scene for what has become the primary conference for a vibrant and important field of research. The founding of the *International Language Testing Association* in 1992 cemented the increasingly international ties. With this came a broadening of theoretical perspectives which fed into the growing literature fostered through journals such as *Language Testing*, *Assessing Writing*, *Language Assessment Quarterly* and *Language Testing in Asia*.

Over the years, we have also witnessed significant paradigm shifts including moving from testing to assessment, from examining issues with English as a second/foreign language to those of bilingual and multilingual perspectives, from assessing a single language to assessing multi-languages and multi-literacies, from a homogeneous test-taker group to test-takers with multilingual, multicultural, and diverse educational backgrounds. In recent years, we have seen the field move away from the dominance of large-scale international and global testing to more local (or localised) and *glocal* assessment. Similarly, we have seen a move away from a primarily measurement-driven approach to the operationalisation of validation to a position which recognises the importance of the context-of-use constructs to actively engage with the key stakeholders who represent that context. The professionalization of our field has also expanded beyond its traditional strongholds of the USA and the UK, with critical contributions from scholars across the world.

The papers in this symposium each provide a unique perspective on the evolution of language testing theory and practice to meet the increasingly diverse contexts of test development and use. *O'Sullivan* sets the scene by reflecting on the development of testing practice globally while highlighting the need for locally inspired scholarship around validation. *Jin & Zhang* discuss this thinking in their validation of a locally developed oral proficiency scale, offering insights for future research on the context validity of language proficiency scales. *Sawaki* speaks to the evolution of constructs in one local setting. She reports on the changing nature of how constructs are viewed within a given context by tracing the evolution of a target construct in the Japanese university entrance examination. Cheng discusses the implications and consequences of the paradigm shift from testing to assessment, focusing on large-scale testing in China and educational assessment in Canada. *Khalifa* continues this focus on consequences as she explores how impact frameworks have been employed in assessment systems in multilingual and multicultural societies across the Middle East and

Southeast Asia. Finally, *Kunnan* looks to the future, reporting on two experimental projects. The symposium finishes with a reflective commentary by the discussant, Chalhoub-Deville.

Presentations of the Symposium

The Testing Circle: East to West and Back Again

Barry O'Sullivan (British Council)

These days we tend to look to the so-called 'traditional' sources of test development, administration, and validation theory. In this way, we tend to look to the USA and Europe as the true originators of our testing world. The reality of how we have arrived at the testing cultures we now see around us is far more complex.

This presentation begins with an historical overview of the emergence of standardised competitive testing in Imperial China, where a merit-based selection system was introduced in 605 CE. The system, known as the Kējǔ, was to remain in place for the next 1,300 years. During this time, much of what we recognise as typifying modern testing systems was developed. Competitive examinations spread to other countries influenced by China, most notably Korea, Vietnam, and Japan. In later years the approach influenced the introduction of competitive examinations in Europe and the USA. While the practice of testing clearly developed in imperial China, it is equally clear that the scholarship of testing only seriously began in the early 20th century. Much of this scholarship was centred around individuals such as Edward Thorndike at Teachers College, New York. The approach mainly focused on how well the test performed statistically. This emphasis on the measurement approach continues to dominate much thinking in the USA. During this time the European approach has been to concentrate primarily on the construct to be tested, though here to the more recent trend has been to consider the construct-in-context.

This presentation explores the circle of developments in the testing world. While we learn from each other's practices, there are culturally and linguistically bound reasons why our approaches differ. The presentation ends by calling for a recognition of these differences and a broadening of the scholarship of language testing to reflect them.

Context Validity of Language Frameworks: A Comparative Study of the CEFR and the CSE Oral Proficiency Scales

Yan Jin, Lin Zhang (Shanghai Jiao Tong University)

Context has long been recognized as a crucial component of communicative language ability (CLA). Moving beyond psycholinguistic CLA models (e.g., Bachman, 1990; Weir, 1993), Chalhoub-Deville (2003) proposed a socio-culturally mediated interpretation of the L2 construct, defining it as "ability-in-user-in-context." This notion emphasizes the dynamic interaction between a language user's abilities and situational facets. In language assessment, Weir's (2005) socio-cognitive framework introduced the concept of context validity, encompassing both intrinsic task characteristics and the external socio-cultural contexts of task performance (Taylor, 2011).

Context validity refers to the degree to which the linguistic demands and situational features of test tasks represent those of real-world language tasks. When evaluating language proficiency scales, context validity pertains to the extent to which a scale meets the needs of its application context and is culturally appropriate (Vandergrift, 2006). To date, alignment studies have explored level correspondences between tests and scales, and validation in alignment has primarily centered on procedural, internal, and external validities. There remains a lack of empirical research specifically addressing the context validity of scales. Grounded in the local construct notion and socio-cognitive framework, this study examines the context validity of two language proficiency scales, the Common European Framework of Reference (CEFR) developed in Europe and China's Standards of English Language Ability (CSE), in their alignment with the College English Test–Spoken English Test (CET–SET). Findings suggest that both scales are generally suitable for alignment with the CET-SET developed in China. However, the CSE, as a localized scale designed for English language education in China, demonstrates advantages over the CEFR in level setting and descriptor features. This study provides validity evidence for a locally developed oral proficiency scale, offering insights for future research on the context validity of language proficiency scales and their application in diverse linguistic and cultural contexts.

A Historical Review of the Target Construct Reflected in the Design of The National University Entrance Examination in Japan

Yasuyo Sawaki (Waseda University)

Japan is known as having a unique testing culture, where high-stakes assessment drives various aspects of school-based English language instruction and assessment (Kuramoto & Koizumi, 2016; Sasaki, 2008). In particular, the modern national university entrance examination (henceforth, the national test) in place since the 1980s plays a significant role in school-based English language education. The national test, currently taken by approximately half a million candidates each year, is used for first-stage candidate screening by most national and public universities and as part of institutional admission testing of various types by many private universities across the country. Given the function of the national test as an achievement test of high school English mandated by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), the test design closely reflects how the target English language ability is conceptualized in the national course of study.

The present study explores how the conceptualization of the target construct and the design of the English section of the national test has evolved over the last four decades with changes to the name and content of the test from the First-Stage Exam (1979 to 1989), the National Center Test (1990 to 2020), and to the Common Test for University Admissions (CTUA; since 2021 to present). Toward this end, multiple data sources including previous test content analyses, past test papers, test development committee reports, and policy documents will be qualitatively analyzed along with relevant literature. The results will be discussed from the perspectives of the role of the context of language use in construct definition and task design (Bachman & Palmer, 2010; Chalhoub-Deville, 2003; Chapelle, 1998), as well as the introduction of the CEFR and some aspects of critical thinking skills emphasized in the current MEXT course of study (2018) to the CTUA test design.

Consequences: Constructs in Context

Liying Cheng (City University of Macau)

Constructs in context is a term traditionally used in personal psychology to discuss the interactions among people, event, and meaning (Anderson & Kirkland, 1990). Constructs in context is also used in disciplines such as business (Yadav, Seth, & Desai, 2017) and leadership (Pathak & Jha, 2024) to examine and predict patterns of research trends. In language testing and assessment, we adopt the term "ability-in-user-in-context" (Chalhoub-

Deville, 2003) and demonstrate how consequences, as the results of test use in context, need to be studied at the interaction of constructs and contexts (Cheng & Sultana, 2022; Chapelle, 2020).

Based on contemporary validity theories which aim to integrate multiple perspectives into a socio-culturally situated argument on the alignment of testing and assessment practices, values, and consequences (Brookhart, 2013; DeLuca, 2011; Moss, 2003), this presentation illustrates the research trend on consequences of test use (often called washback and impact of testing) in contexts. This presentation discusses the paradigm shift from testing to assessment, from traditional psychometric approaches to validity to a socio-culturally situated validity argument. A shift reflects varied research studies on consequences of test use of large-scale testing in China (single location, single test-taker group, and single language background) to test use of educational assessments within a multilingual, multicultural, and multilevel educational setting in Canada.

Impact frameworks Evolution & Usage in Multicultural and Multilingual Contexts

Hanan Khalifa (MetaMetrics Inc)

Early philosophical treatises such as Plato's The Republic or Ibn-Khaldun's Muqaddimah advocated education as a tool which drives society to improve the quality of people's lives. To date, education is perceived as a change agent and much attention is given to measuring the outcome of the educational process and the impact of educational interventions. Introducing language assessment as part of educational reform initiatives is therefore not surprising and is increasingly frequent with three predominant aims: increasing language learning outcomes, providing public accountability and promoting desired changes in learning and teaching practices. Given that "good examinations are not guaranteed to produce positive results and bad examinations do not necessarily produce bad ones" (Wall 2004: xiv), test impact is a phenomenon of great interest to investigate but the complexity of educational systems makes it challenging to do so. Therefore, how can education providers effectively evaluate whether their interventions achieve the intended results? Various impact models and frameworks have been born in response to this question.

This paper discusses how impact models and frameworks evolved over the years in the east and west. In the 1980s and early 1990s, the language testing community witnessed the inclusion of washback and backwash principles in test development and validation processes. Several theoretical impact models have been born (see Bailey 1996, Watanabe 2004, Green 2007, Milanovic and Saville 2009, Saville 2012 to name a few) focusing on participants, products and the interaction between them; dimensions of washback, variability and intensity of impact, and action-oriented approach to impact). More recently, Khalifa (2022) introduced Impact framework which considers social return on investment and sustainable practices. After a brief discussion of impact models and frameworks, this paper explores how their use in assessment contexts in multilingual and multicultural societies in Southeast Asia, the Middle East, Europe and Central and South America.

Integrating Teaching, Learning, and Assessing: The SBA Approach

Antony Kunnan (City University of Macau)

The most underreported movement in language assessment is the Learning-Oriented Language Assessment (LOLA) approach which has championed the integration of teaching, learning, and assessing. Based on substantial research (Gebril, 2021; Jones and Saville, 2016; Purpura, 2014. 2021), there have been varied operational approaches that have addressed this issue. Some of them are the Scenario-Based Approach (SBA; Purpura, 2024), the Dynamic Assessment Approach (Poehner, 2025), the Diagnostic Assessment Approach (Huhta and Harding, 2025). Each of these address different aspects of this movement.

The SBA approach, the focus in this presentation, follows Purpura's (2021) performance moderators (instructional, socio-cognitive, affective, social-interactional, and technological dimensions). The primary goal of this project is to not only integrate assessments with teaching and learning but to help build test takers' learning capability through LOA and randomized control trials (with SBA as the treatment) in Vietnam. The projects focused on the instructional dimension and the cognitive dimension. Both projects were in different contexts: the first in pre-service teacher training and the second for school teachers.

SBA-based assessments were created with reading passages and tasks and source-based writing tasks for each context. Scenarios and tasks were constructed to provided language learning and learning assistance for initial incorrect responses. An LOA approach was used in scoring: for example, 4 points for correct response; 3 points for correct response after initial clue; 2 points for initial clues and explanations; and 1 point for final clue; and 0 point for final incorrect response. The main finding from both projects is that the value-added quality of the assessments show test takers can develop more capability in language learning through an LOA approach. In an accompanying survey at both sites, participants overwhelmingly supported the concept of learning assistance in assessments.

Symposium 1

Evolving Glocalization in Language Testing in Asia: Reflection and Implications

Chair(s): **Jessica Row Whei Wu** (Language Training and Testing Center, Taiwan)
Discussant(s): **Lynda Brigid Taylor** (Centre for Research in English Language Learning and Assessment (CRELLA), University of Bedfordshire, UK)

Friday, June 6, 2025, 1:30pm to 3:30am Location: Poonsapaya

This symposium, proposed for the 46th Language Testing Research Colloquium (LTRC) in Bangkok, Thailand, serves as a sequel to the symposium held at the 41st LTRC in Atlanta, Georgia, in 2019, which was inspired by the publication of *English Language Proficiency Testing in Asia: A New Paradigm Bridging Global and Local Contexts*. The 2019 symposium introduced the concept of glocalization in language testing, exploring the intersection between local contexts in Asia and global assessment standards. As detailed in the book, the development and implementation of standardized English language tests across Asia have reflected a growing need to balance local relevance with international recognition.
Six years later, significant advancements have taken place in the field, prompting a renewed examination of how large-scale language proficiency tests in Asia have adapted and evolved. These advancements include changes in delivery modes in response to new demands arising from educational reforms and the global pandemic, which have led to a widespread adoption of online and hybrid testing formats. Additionally, the advent of Generative AI has prompted reconsideration of ethical guidelines, test constructs, item writing, scoring, and the provision of feedback to learners, reshaping both the technical and theoretical foundations of language testing.

This symposium will feature four presentations, which will focus on major language proficiency test developed and used in Asia (in chronological order by the launch year of the test), including the College English Test (CET) in China, the Test of English Proficiency (TEPS) in Korea, the General English Proficiency Test (GEPT)/ the BEST Test of English Proficiency (BESTEP) in Taiwan, and the Vietnam Standardized Test of English Proficiency (VSTEP). These presentations will explore the challenges and innovations encountered in aligning these tests with international standards while maintaining their cultural and educational relevance. Although the primary focus is on language testing in Asia, the discussions will have significant implications for testing practices beyond Asia, particularly in understanding the balance between local contexts and global standards.

A discussant will critically analyze these presentations, offering insights into the broader implications for both locally-produced and internationally recognized tests. The symposium aims to deepen the ongoing dialogue on glocalization, drawing on the foundations laid in the 2019 publication and the accompanying symposium, while honoring the memory of our esteemed colleagues, Prof. Cyril Weir and Dr. Jamie Dunlea, who were not only strong believers in the glocalization of language testing but also key figures in its realization. Their contributions have profoundly shaped the field, and this symposium seeks to honor their legacy by continuing the work they so passionately championed.

The symposium aligns closely with the LTRC 2025 theme, "Language Assessment in Multicultural Contexts: West Meets East," by examining the convergences and discrepancies in language testing practices between local and global paradigms. It will contribute to the broader understanding of how language testing practices are shaped by social and cultural influences, enriching the global discourse on language testing.

Presentations of the Symposium

Translation Assessment in Large-Scale Language Testing: A Case Study of the College English Test

Yan Jin (Shanghai Jiao Tong University)

Translation, recognized as a key mediational activity in language frameworks like the Common European Framework of Reference for Languages (CEFR), is rarely assessed in international language tests due to test takers' diverse first language backgrounds. However, in the Chinese context, Mandarin as a shared written language enables the incorporation of a translation task in English language assessment. This presentation examines the evolution of the translation component of the College English Test (CET), a large-scale English as a foreign language test for non-English-major college students in mainland China.

Introduced in 2013, the CET translation task was designed in response to the government's call for promoting Chinese culture, history, and contemporary developments. This discourse-level task requires test takers to translate a culturally-themed paragraph from Chinese to English. The task varies slightly between CET Band 4 and Band 6, with test takers translating 140-160 and 160-180 Chinese characters, respectively. Since its inception, translation has become an integral part of the CET, significantly impacting English language teaching and learning at the tertiary level.

This presentation will address three key questions:
1. Why is translation assessed and what are the design features of the CET translation task?
2. What are the challenges and recent developments in CET translation assessment?
3. How might the rapid advancement of AI technologies impact the future of translation assessment?

By exploring these questions, this presentation aims to contribute to the assessment of translation in large-scale language testing. It will delve into the ongoing discourse on cultural representation in localized language assessments and examine the balance between assessing language skills and cultural knowledge. The discussion will also consider implications for international language tests, offering insights into potential applications of translation tasks in broader assessment scenarios.

Glocal Tests and Test Glocalization: The Case of TEPS

Yong-Won Lee (Seoul National University)

Nowadays, the term glocal tests is used increasingly in the field of language assessment to describe a variety of large-scale English proficiency tests that are locally developed outside the inner-circle countries of English but are created, administered, and researched in accordance with global standards of practice and quality assurance. The glocal tests can be defined much more broadly as well because glocalization usually refers to the provision of a product or service with both the global use and local adaptation in mind (Dendrino, 2013; Weir, 2020).

In fact, such test glocalization can be achieved by creating a new test from the scratch with both global and local visions taken into consideration or by altering the existing tests for either global or local use. Relatedly, there could be two possible approaches to glocalizing the existing tests: (a) globalizing local tests; and (b) localizing global (or international) tests (Weir, 2020; Wu, 2020). Although such perspectives on glocal tests provide useful insights into creating locally- and contextually-appropriate assessment systems, however, a more active program of research seems urgently needed now to clarify on what we exactly mean by the glocal test characteristics and what are the guiding principles for creating glocal or glocalized tests appropriately in a particular locality/culture.

With these as a background, the main goals of the current presentation are to: (a) review previous research and existing theoretical frameworks on developing the local, global, and glocal tests; (b) analyze TEPS, a locally-developed, large-scale English proficiency test primarily for South Korean EFL learners, from the glocal test viewpoint, and (c) identify major issues and challenges that need to be carefully considered in creating glocal tests and adapting the existing tests for global or local uses. The major research findings are discussed along with their implications for assessment development and validation.

Evolution of English Proficiency Testing: Journey from GEPT to BESTEP

Rachel Yifen Wu, Anita Chunwen Lin (Language Training & Testing Center, Taiwan)

This presentation examines the evolving need for English language tests in Taiwan, shaped by the dynamic interplay of socio-cultural contexts, educational policies, and test use. Over the last decade, the role of English language testing in Taiwan has shifted from a traditional approach focused on assessing what students have already learned (assessment of learning) to one that uses assessments to support and enhance ongoing learning (assessment for learning). This shift underscores the need for cohesive alignment among teaching, learning, and testing practices.

The presentation will offer a comparative analysis of two locally developed English proficiency tests, introduced 24 years apart in response to evolving educational policies: the General English Proficiency Test (GEPT), launched in 2000, and the newly introduced BEST Test of English Proficiency (BESTEP) in 2024. While both tests share a common objective of assessing English proficiency, they reflect different educational priorities and advancements in test design over time. The presentation will explore the distinct purposes behind each test, examining how changes in educational policy have shaped their test constructs, task designs, and result reporting mechanisms. It will also highlight how each test incorporates learning-supportive features to enhance language proficiency, as well as the key differences that set these local tests apart from global English proficiency exams.

The presentation will also address the efforts made to tackle the challenges of making these local tests internationally transportable qualifications. Finally, the presentation will explore implications for developers of glocalized tests and propose future directions for incorporating the latest technical and theoretical advancements in test design and services.

The Evolution and Devolution of Language Tests in Vietnam - an Ecological Review

Quynh Thi Ngoc Nguyen (Vietnam National University, Hanoi)

This presentation takes an ecological perspective to discuss an overview of language tests in Vietnam during the past decade. Most of the facts and figures in the presentation are about VSTEP – the first national standardized English proficiency test in Vietnam, but the discussion will include other language tests that have been developed, evolved and devolved, during the past 10 years in the country. I will elaborate on multiple layers and aspects of glocalization of these tests by analyzing different implications of the ways they (and indeed other tests discussed in the proposed symposium) have been referred to in literature – as 'locally-produced', 'localized', or simply 'local' tests – in the context of Vietnam. As I asserted in my chapter in the 2019 publication of English Language Proficiency Testing in Asia: A New Paradigm Bridging Global and Local Contexts, not only do these terms refer to different aspects of difference between these tests and the existing renowned international tests, but each of them is also interpreted differently in different contexts. The presentation will provide more insights of these tests and their changes over time in response to contextual challenges and demands to explicate the operationalizations of these terms, which I believe will be useful for developers of glocalized tests in Asia and beyond. In addition, I will focus on how Vietnamese test developers deals with competition as well as cooperation in the testing ecosystem of their own tests, other locally-produced tests, and of course many international and imported tests. By nature, any ecosystem is a community of organisms in symbiosis. I will discuss how these symbiotic relationships result in the gradual evolution and devolution of their tests in different aspects. I will call for further research on this ecological approach on glocalized language tests in other parts of the world.

Symposium 1

The Promise and Perils of Investigating Writing Assessment in Other Languages Through the Lens of English

Chair(s): **Atta Gebril** (The American University in Cairo, Egypt), **Beverly Baker** (The University of Ottawa)
Discussant(s): **Slobodanka Dimova** (University of Copenhagen)

Friday, June 6, 2025, 1:30pm to 3:00am Location: Duangduen

In line with the LTRC 2025 theme that focuses on "Language Assessment in Multicultural Contexts," this symposium takes a critical stance towards writing assessment models and tools originally designed for English, and applied to other languages. This area has not received due attention in the literature, as indicated in a recent synthesis that found only 12 publications tapping into languages other than English (LOTE) in Assessing Writing (Yan et al., 2021). In an attempt to increase attention placed on other languages, Baker and Gebril (2024) recently guest-edited a special issue of Assessing Writing on LOTE writing assessment. Continuing in the same direction, this symposium sheds light on the challenges encountered when using English models and tools in conducting research targeting other languages. These challenges are usually associated with the unique characteristics of these languages, such as different orthographies, wide-ranging rhetorical traditions, diverse instructional settings, and varied socio-political expectations. One of the issues emerging in this line of research is the tendency for researchers to adopt English models without sufficient critical consideration of, or attempts to address, the unique characteristics of the language of interest. Such practices usually result in challenges related to data collection, analysis, and interpretation; this by default could lead to compromising overall research quality. The objective of this symposium is to raise awareness about these issues and share current research efforts by scholars who have brought this critical comparison with English to their work with different languages.

The symposium will follow a 90-minute program, including four presentations representing four different languages (Chinese, French, Arabic and German), and a discussant. The symposium will start with a 5-minute introduction, followed by a presentation from each speaker for 12-15 minutes. The discussant will be given the opportunity to share her perspective for around 15 minutes while the final 10-15 minutes of the symposium will be alloted for entertaining questions from the audience.

The first presentation looks into the unique subconstructs of Chinese writing ability, with a specific focus on linguistic accuracy. The second presentation includes a discussion of distinct understandings of lexical and syntactic complexity as applied to analyses of French academic writing. As for the third presentation, a critical evaluation of Hyland's metadiscourse model is offered to investigate its applicability to L1 Arabic argumentative writing. The fourth presentation discusses the construct of integrated writing in the German language. It is our hope that this symposium motivates other scholars to investigate the challenges and implications of adopting L 1 models within the context of LOTE writing assessment. Such investigations are crucial for better identification of L2 constructs and for building research expertise in local settings.

References

Baker, B., & Gebril, A. (2024). The assessment of writing in languages other than English (LOTE). Assessing Writing, 60, 100840.
Yan, X., Bowles, B., & Malone, M. (2021, June 14-17). A narrative synthesis on languages other than English [Conference presentation]. Language testing research colloquium virtual conference (LTRC), International Language Testing Association.

Presentations of the Symposium

Chinese Character Matters!: An Examination of Linguistic Accuracy in Writing Performances on the HSK Test

Xun Yan, Jiani Lin (University of Illinois at Urbana-Champaign)

Language testing is largely dominated by the assessment of English. In contrast, languages other than English are disproportionally underrepresented in assessment literature, although they feature unique assessment problems. Mandarin Chinese is a less commonly taught language in most parts of the world. Its orthographic and morphological system requires more time and developmental stages for learners to acquire. This source of difficulty can present unique challenges and opportunities for Chinese writing assessment. In addition, exposure to Chinese in both instructional and naturalistic settings can make a difference in language learning, given that the presence of Chinese language is noticeably different across countries and regions depending on historical, economic, and political factors. Thus, examining linguistic accuracy in writing performance across proficiency levels, linguistic, and geopolitical backgrounds might provide insights about the unique subconstructs of Chinese writing ability. In this study, we employed a corpus-based approach to examine the linguistic errors in 10,750 essays written by test-takers from 17 first language (L1) backgrounds on the HSK test. Based on both orthographic types and economic-geopolitical factors, we classified test-taker L1s into 3 groups. We first factor-analyzed a comprehensive array of error types to identify the underlying dimensions of Chinese writing accuracy. Then, dimension scores were included in regression models to predict HSK writing scores for different L1 groups. The results revealed five dimensions related to syntactic, morphological, and lexical errors. Among them, dimensions on character and word-level errors were stronger predictors of HSK scores, although the discrimination power was stronger for test-takers from L1s that are orthographically dissimilar and economic-geopolitically distant from Mandarin Chinese. These findings suggest that Chinese morphology (i.e., the acquisition of characters and how characters form words) constitutes a unique source of difficulty for L2 learners. We argue that morphological elements should be an important subconstruct in Chinese writing assessments.

Lexical and Syntactic Analyses Procedures Applied to the French Language: What Works and What's Lost in Translation?

Randy Appel[1], Angel Aria[2], Beverly Baker[3], Guillaume Loignon[4]
([1]Waseda University, [2]Carleton University, [3]University of Ottawa, [4]Université de Québec à Montréal)

In 2022, our team examined the lexical and syntactic features of French texts from the Test du Certificat de compétence en langue seconde (Second Language Certification Test) at the University of Ottawa. Our objective was to identify characteristics of the texts that differentiated high- and low-level writers, in order to inform both assessment revision and in-house French for Academic Purposes instruction.

We achieved this objective by conducting lexical bundle (LB) analysis, complemented with a tool developed expressly for French called the Integrated Lexical Syntactic Analyser (ILSA; Loignon, 2021). In this presentation, we discuss the following insights from this work from a French-language perspective:
· We discuss the challenges in finding/building relevant databases in French to support the use of ILSA;
· We explain the specific French language measures that can be extracted with ILSA based on automatic corpus annotation; and
· We review our findings in terms of the similarities and differences with the preponderance of research findings in English.

In general, similarities between our findings and those for English suggest overlapping constructs of lexical and syntactic complexity, though not completely. Use of the past participle, for example, may be a stronger indicator of advanced writing in French than in English.

It's important to note that any consideration of similar constructs between English and French is muddied by the fact that our tools still relied on theoretical understandings of writing initially developed for English. Most of the ILSA measures, for example, were inspired by Coh-Metrix. We suggest adaptations to procedures in future studies with French to better address its particular linguistic characteristics.

The Applicability of Hyland's Metadiscourse Model to L1 Arabic Writing: A Critical Cross-Linguistic Study

Abdelhamid M. Ahmed, Lameya M. Rezk (Qatar University)

This study critically evaluates Hyland's metadiscourse model to assess its applicability to L1 Arabic argumentative writing. Although validated for English contexts (Hyland, 2005), applying this model to L1 Arabic presents conceptual and practical challenges due to distinct linguistic features and rhetorical traditions (Al-Khatib, 2001; Al-Ali, 2006; Zaki, 2022). For instance, El-Seidi (2000) highlights the use of emphatic devices like "Inna and Anna," which lack direct English equivalents, and attributes differences in argumentative writing styles to L1 discourse transfer and other instructional factors.

The study employs quantitative analysis of argumentative texts, adapting Hyland's framework for the Arabic context. Using the Qatari Corpus of Argumentative Writing, which comprises 390 essays by the same Qatari L1 Arabic students (Ahmed et al., 2024), the findings reveal significant differences in metadiscourse use. L1 Arabic writers employ transition markers more frequently in their native language compared to L2 English (Ahmed et al., 2023). Additionally, the higher occurrence of reformulation and exemplification markers in Arabic (553.46 and 753.31) compared to English (114.34 and 1000, respectively) indicates L1 Arabic writers favour detailed explanations, while L2 English tends toward conciseness (Ahmed & Rezk, in press).

The study also addresses differences in annotation practices, noting how Arabic's inflectional structure and high-context communication style necessitate tailored annotation methods. This requires specialized training for annotators and adaptations in annotation tools. The proposed modifications aim to enhance the cross-cultural applicability of Hyland's model, thereby improving argumentative writing instruction in L1 Arabic. Overall, this research bridges a gap in comparative rhetorical studies and enriches our understanding of metadiscourse as a crucial aspect of academic communication across languages.

Beyond EFL – The Use of Theoretical Models for Integrated Writing Assessment in a German Context

Sonja Zimmermann (zimmermann@gast.de)

Integrated writing tasks have become a common tool for assessing language proficiency, particularly in academic settings. These tasks are widely used in English as a foreign language (EFL) assessments, where substantial theoretical models have been developed to understand the cognitive processes involved. However, the expansion of these models to other languages remains relatively unexplored, despite the increasing use of integrated writing tasks in various language proficiency tests worldwide.

This paper aims to bridge the gap by examining how theoretical frameworks for integrated writing, originally designed for EFL learners, can be adapted to other languages. Using German as a case study, the paper applies EFL-based models to an integrated writing task in a large-scale standardized proficiency test for university admissions. The task requires test takers to synthesize information from a reading text and a graphical input.

The aim of the presentation is twofold: On the one hand, results from a study will be presented that used a combination of eye-tracking and stimulated recalls to look into the cognitive processes involved in summarizing information from two different sources. The analysis revealed that the task elicits a specific interaction of basic reading and writing processes, and that the use of cognitive processes and the utilization of the two sources varied at different stages of the writing process. The processes described in this study mostly confirmed findings from existing research in the EFL context, implying that the current understanding of integrated writing can be expanded to other languages than English. On the other hand, the intention with this presentation is also to discuss some limitations of the existing integrated writing models, proposing a more comprehensive framework that takes into account different modalities of input material and diverse genres.

Symposium 2

Assembling, Adapting, Adopting: The Development and Implementation of Standards and Frameworks for Young Learners

Chair(s): **Barry O'Sullivan** (British Council, United Kingdom)
Discussant(s): **Barry O'Sullivan** (British Council, United Kingdom)

Saturday, June 7, 2025, 8:30am to 10:30am Location: Poonsapaya

A number of different language standards and frameworks are in widespread use around the world. These serve to form the basis of various elements of the learning system, facilitating their alignment with each other. For example, in O'Sullivan's Comprehensive Learning System (2020), such frameworks are the unifying element underpinning curriculum, delivery and assessment, thus enabling the system to function as a whole. The performance level descriptors (PLDs) of a framework may be employed to support the design of a syllabus, appropriate teaching and learning materials, and specification of constructs when devising assessment tools.
There are many examples of commonly used frameworks, differing in composition and implementation according to the purpose and context of use: in North America, ACTFL, the Canadian Language Benchmarks (CLB) and WIDA; in Europe, the Common European Framework of Reference (CEFR); in Asia, the Japanese CEFR, the CEFR-J, and the China's Standards of English (CSE).

One increasingly important purpose is for use with young learners, as seen in the linking of international tests to the CEFR and other frameworks. However, frameworks used with young learners differ from those used with adults as they also need to take into account the complexity of construct definition resulting from the characteristics of different age groups, for example, in terms of children's cognitive, metacognitive, social, L1 and literacy development (Butler, 2016). Therefore, frameworks need to be created or adapted to the particular educational context and be appropriate for the learners' age (Little, 2007), as well as supporting learning and providing positive washback.

This symposium discusses the development and implementation of frameworks in four diverse eastern and western contexts: China, where the custom-built CSE framework has been developed with reference to national curriculum goals; Japan, where the CEFR-J represents an adaptation for young learners in a specific learning system; the US, where WIDA standards were created to support multilingual ESL learners in schools; and in a global context from a UK perspective, where the CEFR has been used to develop assessment for specific learning materials. Researchers in these diverse contexts have responded with convergent and divergent approaches to the challenges of teaching and assessing young learners. They have all attempted to establish consistency among curriculum, delivery and assessment to meet the needs of relevant stakeholders (e.g., learners/test-takers, parents, administrators), with some even drawing on a common pool of PLDs (the CEFR). However, they have also faced various demands in terms of national policy requirements, EFL vs. ESL settings, differing age groups, monolingual versus multilingual contexts, localisation, and national vs. international outlooks in addition to the very specific educational cultures.

There is still a great deal for the field to learn about how standards, frameworks and PLDs are implemented and how they can positively impact on learning systems, particularly for young learners (McKay, 2006). By investigating such diverse experiences in one forum, the symposium aims to shed light on this gap in our understanding and on how different social, educational and cultural values impact on their use in curriculum and assessment.

Presentations of the Symposium

Principled Localisation of the CEFR: an Example of the CEFR-J

Masashi Negishi

The aim of our project group was to create a comprehensive framework for English language teaching that would link primary, secondary and tertiary schools in Japan. This led us to focus on the CEFR. However, as soon as we started researching the CEFR, we realised that Japanese learners of English were heavily skewed towards the lower levels of the CEFR. Therefore, we first conducted a branching of the lower levels so that learners and teachers could feel the progress of their learning. Then, using the CEFR Can Do descriptors and other Can Do descriptors that existed in Japan, we created Can Do descriptors for each level, validated them using self-assessment data, and revised them. The results were published as the CEFR-J, from which research was conducted on the CEFR-J-aligned RLDs (Reference Level Descriptions) and the tools for judging the level of input and output texts were developed. In the next stage of research, tests based on the CEFR-J Can Do descriptors were developed and validated, and a revised version was published. In the final stage, research was carried out with secondary school teachers on teaching based on the CEFR/CEFR-J.

This presentation will focus on a comprehensive overview of the use of the RLD tools and tests developed in the CEFR-J project to monitor the progress of secondary school students in Japan. The CEFR-J localises the CEFR to our needs, and the advantage of this is that we can always relate our findings to the ever-growing body of CEFR research. Our research shows that principled localisation can contribute back to the original CEFR, and that this bilateral relationship strengthens the framework by reflecting local needs and times.

Linking Materials to the CEFR to Develop a Standardised Approach to Assessment in a Global Context

Carolyn Westbrook, Johnathan Cruise (British Council)

In a Comprehensive Learning System (O'Sullivan, 2021), assessments should reflect what has been taught and standards, which may be local, national or international depending on the context, are at the heart of the system. Assessment, while time-consuming, is a fundamental part of a teacher's job as they need to provide learners and other stakeholders with accessible feedback about the outcomes and learners' progress.

This presentation will report on a project to align teaching materials for secondary-school learners to the Common European Framework of Reference for Languages (CEFR) and then to develop a standardised assessment approach for an international teaching context, balancing teacher workload with the need for fair and valid assessment, and standardised reporting. The teaching materials for each course comprise 10 specially-produced 'magazines' (i.e. modules) containing thematically-linked tasks, culminating in an assessed project. For each assessment, a can-do statement and corresponding performance indicators were selected to assess the task, and a standard reporting tool was developed. Teachers' perceptions of this new approach were investigated through post-training questionnaires, focus groups and interviews. Results suggest that teachers are able to use the can-dos and the performance indicators and that they feel the approach will lead to fairer marking through standardisation. They also feel that the approach will help them to better prepare learners for assessments and provide better feedback. However, there were differing opinions regarding the number of assessments that are feasible. In addition to briefly presenting the teachers' reactions to the approach, we will outline the challenges of creating a standardised approach for a global context. We will then highlight some issues related to aligning materials for younger learners to the CEFR and finally, the implications for developing frameworks for younger learners will be discussed.

Developing a Cognitive Diagnostic Computerized Adaptive Test (CD-CAT) based on China's Standards of English Language Ability (CSE)

Lianzhen He *(Zhejiang University)*

The China's Standards of English Language Ability (CSE), first released in 2018 and updated in 2024, serves as a unified framework for aligning English language teaching, learning, and assessment across different educational stages in China. The updated version incorporates refinements in typical communicative activities and salient features of language use, further solidifying the CSE's role as a reference for language assessment practices.
Against this backdrop, we developed a web-based cognitive diagnostic computerized adaptive test (CD-CAT) to assess students' English proficiency in a university in China. Grounded in the CSE, this test is administered to incoming undergraduate, graduate, and doctoral students in the university, providing individualized diagnostic score report across four domains (i.e., listening, reading, writing, and speaking). The score report includes detailed feedback on students' CSE levels, subskill mastery, and personalized recommendations for improvement. A large-scale operational test was conducted with 5,993 students. The results showed that acceptable classification accuracy could be achieved with just 25 items for each domain (i.e., listening, reading), representing a 25% reduction in length compared to the traditional cognitive diagnostic test. Despite the shorter test, diagnostic quality was even higher. These findings suggest that integrating cognitive diagnostic models with adaptive testing can effectively operationalize national language frameworks like the CSE, enabling more efficient, and learner-centered assessment.

This presentation will outline the development and implementation of the CD-CAT system. I will introduce key components, including attribute specification based on CSE descriptors, Q-matrix construction, selection of an appropriate cognitive diagnostic model (CDM), adaptive item selection strategies, and item parameter and ability estimation. Finally, I will discuss the implications for future test development based on language standards.

Proficiency Level Descriptors in U.S. Contexts: From Standards to Assessment to Instruction

Margo Gottlieb, Lynn Shafer Willner (WIDA at the Wisconsin Center for Education Research)

Proficiency Level Descriptors (PLDs) have been a stable component of WIDA Standards Frameworks designed for multilingual learners in Kindergarten through grade 12. Currently in its 4th edition, the PLDs have evolved to represent current linguistic theory operationalized in its large-scale annual language proficiency test, administered to almost 3 million students, and classroom-based assessment. Although language test scores should facilitate decisions about language proficiency (Bachman & Palmer, 2010), WIDA research (Park, 2024) indicates that the PLDs in the 2020 edition are challenging for educators due to their complexity and vagueness.

The PLDs are a series of matrices, representing 6 levels of language proficiency by 3 dimensions of language- discourse, sentence, and word/phrase- for 6 grade-level clusters (K, 1, 2-3, 4-5, 6-8, and 9-12). Our investigation centered on refining the current PLDs to increase their viability. Using mixed methods, this 2023-24 study examined linguistic complexity, use of technical vs. plain language, level of detail, and presentation formats to ascertain the effectiveness of PLD use by educators. Qualitative data revolved around written surveys, online focus groups, and interviews while quantitative data included frequency counts within written surveys and T-tests.

Findings indicate statistically significant differences in favor of plain language PLDs that utilize fewer words, bolding those keyed to language progressions. However, educators seemed to struggle shifting from a traditional, linear view of language development to a more functional one based on genre, purpose, audience, and topic. Analysis of digital options for manipulating the textual format of the PLDs yielded a preference for both for Microsoft and Google formats. Usefulness of test scores are contingent on their transparency and meaningfulness for designated interest groups (Papageorgiou & Manna, 2023). By increasing technological options and streamlining the PLD language, educators can more readily interpret assessment information, better understand language test results, and increase their classroom applications.

Symposium 2

Assessing Internationally Mobile Healthcare Professionals: Redrawing the Boundaries of Language Testing

Chair(s): **Gad Lim** (OET), **Peter Kim** (OET)
Discussant(s): **Lynda Taylor** (University of Bedfordshire)

Saturday, June 7, 2025, 8:30am to 10:30am Location: Duangduen

The conference theme of language assessment in multicultural contexts is nowhere more evident than in language testing for the purpose of international mobility. Migration brings together people from east and west, as well as from the global south and global north, necessarily creating socially and culturally constrained communication contexts where individuals' second languages need to be used. Writ large, it is not just migrants and their families who are impacted, but also the societies that they join and leave (Roever & McNamara, 2006). Thus, language testing for international mobility warrants additional research attention.

The language and communication requirements are further heightened and complicated in specific professional contexts such as in the case of internationally mobile healthcare professionals. First, the target language use domain is not one that language assessment practitioners are typically expert in, creating challenges in test definition and design. Second, the participants in these communication contexts can include migrant care providers and local care recipients, who are each put in positions of power and vulnerability, but in very different ways, potentially requiring greater skill for successful communication. Together, these challenges create a greater need to manage risk in an arguably high stakes assessment use context (Macqueen et al., 2020).

Because the healthcare context is particularly specific, it can reveal issues that language assessment researchers may not have considered in their theorizing about and investigations into assessment design and use. Thus, the context serves to interrogate and challenge the theories and models used in the field, helping to refine thinking about communicative competence, test design, policy, and impact.

Accordingly, this symposium brings together papers on the assessment of migrant healthcare workers' language skills, and its impact on their professional practice and on the systems in which they are embedded. The studies are set in Australia, Canada, and the UK, three top destination countries for doctors, nurses, and carers.

The first paper investigated communication challenges encountered by personal care assistants. The authors expanded an existing model of communicative competence (Celce-Murcia, 2008) to account for the workplace specific difficulties encountered by participants. The next paper deals with a context where authorities recognized an additional test (Chan & Taylor, 2020) and the positive social impact this has had. On the other hand, the third paper shows that there are implications if people preparing for different tests end up differently prepared for the workplace. Having multiple tests, therefore, may bring both positive (e.g., choice, efficiency) and negative impact (e.g., differing readiness), a complex matter which requires thinking about test impact in more nuanced ways. The fourth paper considers further directions in which policy and impact can be taken.

Together, the papers in this symposium not only provide insight into language testing in the healthcare setting, but also advance the field more broadly by challenging our theories and thinking. The discussant will help us map how our boundaries need to be redrawn, and the further questions we need to explore as we continue our collective journey.

Presentations of the Symposium

Modelling Communication Challenges of Aged Care Workers from Multilingual and Multicultural Backgrounds

Ute Knoch, Philipa Mackey, Sally O'Hagan, Ivy Chen (University of Melbourne)

Effective communication is central to the majority of activities in aged care settings (Bennett et al., 2016). In Australia, as in many English-speaking countries, personal care assistants (PCAs) working in aged care settings are increasingly from multilingual and multicultural backgrounds, with many growing up in countries where English is not the primary language. Many can start working without having to meet specific English language requirements, and anecdotal evidence suggests some struggle with the complexity of the required workplace communication. Such difficulties may result in safety issues, and impede carers creating meaningful relationships with older people in the care setting or successful working relationships with colleagues. To date, however, few studies have investigated what aspects of communication carers from culturally and linguistically diverse (CALD) backgrounds find difficult, nor have these difficulties been modelled theoretically. The aim of the current study was to capture communication difficulties of CALD PCAs and explore the relevance of Celce-Murcia’s (2008) model of communicative competence to communication in this context.

The interview-based study focused on three groups of participants: (1) thirty PCAs from CALD backgrounds, (2) twenty supervisors of PCAs, and (3) eighteen older people who were receiving care and/or nominated support people who participated on behalf of an older person. Interviews were conducted online, recorded and transcribed. The data were thematically coded to identify common themes of communication challenges. The findings show that the communicative challenges facing new PCAs from CALD backgrounds range from specific linguistic challenges to more workplace-specific communication problems. All six components of Celce-Murcia’s model were identified in the data. However, because the model was developed without a specific communication context in mind, additional areas were added to extend the model to workplace communication in aged care settings. The study has implications for test development and the training of PCAs from CALD backgrounds.

The Social Impact of Tests: A Multi-Stakeholder Evaluation of Introducing OET in the UK

Brigita Seguis (Cambridge University Press & Assessment)

The introduction of the Occupational English Test (OET) as an alternative to IELTS in the UK was perceived as a more equitable and relevant option for international nurses (Roberts, 2020). According to Nursing and Midwifery Council (NMC) data, just over 50% of nurses now submit OET scores as proof of their English language proficiency, overtaking IELTS in 2021. Given the central role that OET plays in the overseas recruitment landscape, this paper aims to evaluate the impact that the test has had on its stakeholders since its introduction in 2017.

The research adopted an exploratory, qualitative design and involved semi-structured interviews with healthcare recruiters, clinical educators, test preparation providers and former OET candidates.

The results of the study revealed a strong overall congruence between the OET content and actual communication in the workplace as perceived by key stakeholders. Former OET candidates in particular were readily able to link the skills acquired during test preparation to real life professional communication and day-to-day tasks they perform in the National Health Service (NHS). The study also showed that OET has had a positive impact on recruitment timelines, enabling candidates to meet their NMC registration requirements more quickly as well as leading to a higher number of international nurses recruited into the NHS.

In the final section of the study, a counterfactual analysis asked stakeholders to imagine what the situation would have been like had OET not been available as an option. Additionally, data from former IELTS candidates was included to provide a comparative perspective. This approach helped to gather alternative evidence for assessing the direct influence of OET on the changes observed within the broader healthcare landscape.

From Preparation to Practice: The Role of OET, IELTS, and PTE in Preparing Nurses for Workplace Communication

Jason Fan[1], Ute Knoch[1], Michael Davey[1], Ivy Chen[1], Sally O'Hagan[1], David Wei Dai[2]
([1]University of Melbourne, [2]University College London)

Despite a profusion of research on the impact of language assessments, few studies have focused on the impact of language assessments for professional purposes (LAPP) within a specific social context or policy space. This study explores the impact of the Occupational English Test (OET) in Australia where the OET, along with several other language tests such as the IELTS and PTE, is used as proof of English proficiency for healthcare professionals seeking professional registration. Specifically, this study explores nursing candidates' perceptions on the relevance of the OET tasks to the nursing domain and the extent to which their experience of preparing for and taking the OET prepared them for the language demands in the workplace, with comparisons to the IELTS and PTE.

Data were collected through semi-structured interviews with 30 nurses or midwives working in Australian healthcare settings, 15 of whom took the OET and 15 either IELTS or PTE for their professional registration. The interview data were coded and analysed thematically. The findings reveal that the nursing candidates perceived the OET tasks as highly relevant to their workplace communication, whereas the IELTS and PTE were seen as broader assessments with less specific relevance to healthcare contexts. Preparing for the OET had a positive impact on the participants’ readiness for workplace communication, as the healthcare-related tasks in the OET were perceived as beneficial for developing communication skills in scenarios such as doctor-patient consultations and writing referral letters. The IELTS and PTE were viewed as less tailored to healthcare-specific language demands. These findings highlight the importance of test authenticity and alignment with professional practice, with implications for the design and use of LAPPs in Australia and similar contexts where both LAPPs and general proficiency tests serve as a high-stakes gatekeeping function for professional registration.

Navigating Language Proficiency Testing Policies: Challenges for Migrant Healthcare Professionals in Canada's Health System

Eunice Jang[1], Maryam Wagner[2] ([1]University of Toronto, [2]McGill University)

In Canada, internationally educated healthcare professionals comprise between 12-35% of the workforce (CIHI, 2022). These professionals are indispensable to Canada's healthcare system. but must meet stringent language proficiency (LP) testing requirements. However, the specifics of required language levels can vary between professional designations, creating confusion and barriers for applicants. In this paper, we draw upon collaborations with the nursing and pharmacy regulatory bodies in Canada in which we: 1) reviewed and evaluated language testing priorities, and current practices; 2) identified their language communicative demands; and 3) evaluated various language tests for their use for these two professions. We use this research to discuss implications for language assessment policy.

A central policy issue is related to the confusion and barriers created for applicants because of the variability in the LP evidence required for the different professional designations. While general language tests assess broad skills, they often fail to measure the specialized language needed in healthcare. Specific-purpose tests could address these gaps, but they may be costly and not widely available, leaving professionals to navigate between multiple tests that may not fully capture their communication competencies for healthcare settings.

Our data illustrated shifts in both oral and written practices due to technology. Technological advancements, alongside the rise of assistive AI tools further complicate LP policy. There is a need to reconsider how communication skills in healthcare settings are defined and assessed. Lastly, LP testing policy presents ethical and equity concerns, particularly for professionals from non-English-speaking backgrounds or those facing financial and geographic barriers. There are calls to shift away from strict LP testing to supportive models for streamlining of credential recognition, provision of timely support for foreign-trained healthcare professionals, and alternative proofs of language proficiency. Policy reforms should balance rigorous standards with flexibility in recognizing diverse qualifications and experiences.

Symposium 3

West Has Met East: A Transnational Language Assessment Literacy Project of Southeast Asian Languages

Chair(s): **Ahmet Dursun** (University of Chicago, United States of America)
Discussant(s): **Catherine Baumann** (University of Chicago), **Ahmet Dursun** (University of Chicago)

Saturday, June 7, 2025, 3:30pm to 5:00pm Location: Duangduen

This symposium focuses on the deliverables and impact of a five-year (2019-2024) grant project titled “Professional and Materials Development to Strengthen Southeast Asian Language Instruction,” provided by the Luce Foundation to the Southeast Asian Language Council (SEALC). The project aimed to professionalize Southeast Asian language teaching in the U.S. and Southeast Asia. It provided critical support for language instructors' professional development in eight Southeast Asian languages (Burmese, Filipino, Hmong, Indonesian, Khmer, Lao, Thai, and Vietnamese) and the development of materials and resources for less commonly taught languages in the U.S.

A unique aspect of this grant project was its connection between U.S.-based instructors and colleagues in Southeast Asia, which offers mutual benefits (such as creating a global professional network, increasing collaboration, and accessing more authentic materials from the target culture) but also presents challenges (such as the contextual varieties, including students’ backgrounds, purposes, and motivations to study the language, and cultural differences in teaching methods).

The project involved 56 instructors (30 from the U.S. and 26 from Southeast Asia) in eight SEA languages, who participated in a multi-year series of workshops. These instructors first completed a 4-day ACTFL Oral Proficiency Interview (OPI) training. Familiarity with the ACTFL guidelines provided a common framework for understanding learning outcomes based on general principles of second language acquisition, not tied to language-specific materials or methodologies. In subsequent years, they embarked on an intensive journey of operationalizing the learning outcomes and designed and developed 104 sets of criterion-referenced, performance-based, proficiency-oriented reading and listening intermediate and advanced assessments, followed by re-aligning the instructional materials to the assessments.

The project showcased assessment literacy training resulting in tangible deliverables that played a pivotal role in shaping effective instructional practices and measuring students’ learning outcomes. Post-workshop evaluation surveys (n=56) indicated that the assessment literacy training empowered SEA instructors to make informed decisions when selecting and designing assessment instruments and planning instructional practices accordingly.

This symposium will open by describing the assessment literacy training model implemented in this comprehensive inter-institutional and transnational professional development project. It will also present key issues and challenges encountered as the project unfolded. It will then showcase three interconnected research studies stemming from the project across U.S. and SEA institutions. The symposium will conclude with a discussion of how Southeast Asian instructors transformed and innovated their language pedagogy to meet programmatic and student needs. Specifically, the symposium will make a case on how enhanced assessment literacy has reframed participating instructors' programmatic goals and expectations around curricula, increased confidence in presenting their expertise, and a well-defined sense of professional identity, giving them greater agency in their actual work.

Presentations of the Symposium

Enhancing Indonesian Reading Proficiency: Lessons from a Multi-Year Reading Proficiency Test Pilot Program

Sakti Suryani[1], Erlin Barnard[2] ([1]Harvard University, [2]University of Wisconsin-Madison)

This presentation will showcase findings from a multi-year study on newly-developed criterion-referenced, performance-based, proficiency-oriented reading assessments for language learners at the Southeast Asian Studies Summer Institute in 2022 and 2023. We aim to evaluate assessment methods for intermediate and advanced learners, highlighting the collaboration between US and Southeast Asian instructors to develop and refine assessment tools, demonstrating the power of international cooperation in educational research.

The presentation will, first, detail the rigorous process of developing and refining multiple test sets across proficiency levels. Our two-phase approach included:
Phase 1 (2022): Six intermediate tests for 12 first-year students, five intermediate tests for five second-year students, and five advanced tests for eight third-year students.
Phase 2 (2023): Three revised intermediate tests for 16 first-year students, three intermediate sets for three second-year students, and three advanced sets for three third-year students.

We employed a mixed-methods approach, combining quantitative analysis of test scores with qualitative analysis of student survey responses and teacher reflections gathered through focus group discussions. The presentation will focus on how initial pilot results from 2022 informed the refinement of our assessment tools in 2023. By addressing common issues—such as unclear rubrics or insufficient task familiarity—we improved the validity of the use and interpretation of our assessments. We will present these results from both years, demonstrating how student performance varied across proficiency levels. The results provided concrete examples of how assessment tools could gauge language proficiency and informed curricular decisions, such as introducing mock tests and specialized reading lessons. Finally, we will discuss the broader implications of piloting reading proficiency assessments, emphasizing the importance of collaboration among Southeast Asian colleagues, task-based learning, and integrating authentic materials. Through this approach, assessments become a critical tool for elevating language proficiency and meeting the diverse needs of students in Southeast Asian programs.

Exploring Learners' Perspective of Proficiency-oriented Performance-based Assessments in Burmese

Chan Lwin[1], Maw Maw Tun[2], Ye Min Tun[3], Kenneth Wong[4] ([1]Arizona State University, [2]Northern Illinois University, [3]ohns Hopkins University, [4]UC Berkeley)

The field of language assessment has evolved significantly over the past decade, with large-scale organizations introducing technology-integrated, standardized tests that are widely accessible. However, for less commonly taught languages such as Burmese, assessments still heavily rely on teacher-created materials. These assessments often focus on curriculum-specific content, which may result in learners displaying higher proficiency levels on familiar topics (Cox, Bown, & Burdis, 2015).

Proficiency-oriented assessments, which measure a learner's language ability independently of curriculum content, have been considered as a solution to provide a more accurate reflection of language skills. This study aims to discuss findings from a pilot project using newly-developed performance-based, criterion-referenced reading proficiency assessments with Burmese learners and explore their perspectives on the test's effectiveness and utility. More specifically, we investigated:
1. How do Burmese language learners perceive proficiency-oriented reading assessments?
2. What experiences and challenges do learners face when using proficiency-oriented assessments compared to traditional teacher-created achievement tests?
3. Do content and instruction impact learners' experiences with proficiency-oriented reading assessments?

The participants were selected from Burmese language courses at Cornell University and UC Berkeley (N=12). These courses attracted a diverse range of students aged 18-30, enrolling for personal interest, research, or to fulfill language requirements. Participants took the newly-developed reading tests at the end of the semester, followed by guided interviews conducted via Zoom. The data consisted of students' test scores and post-test interviews. Students' test scores were analyzed using descriptive statistics. Interview transcripts were analyzed for recurring themes for each research questions, following Creswell and Guetterman's (2021). Open coding was used to identify recurring themes, which were then be grouped into categories to draw conclusions for research questions.

The results showed students' performances were varied and directly impacted by instructor's curricular decisions, such as introducing proficiency-oriented learning activities and formative-assessment tasks.

Expanding Horizons: Inter-institutional and Transnational Collaboration in Assessment and Lesson Design and Development

An Sakach (Arizona State University)

This study investigates the impact of a collaborative professional development (PD) model for language teachers, incorporating inter-institutional and transnational exchange components facilitated in both in-person and virtual environments. Specifically, it examines Southeast Asian language instructors' reflections on their participation in collaborative PD projects initiated by the Southeast Asian Language Council (SEALC). These multi-year projects focused on designing and developing criterion-referenced, performance-based, proficiency-oriented reading and listening assessments followed up by instructional materials development, involving educators from the United States and seven Southeast Asian (SEA) countries. Each PD project spanned an academic year. Participants collaborated to design and develop reading and listening assessments. Then they worked together to create instructional materials using authentic materials to re-align their curricula to the newly-developed assessments.

The research utilizes observation notes, post-workshop evaluation surveys, and semi-structured interviews from the participating instructors (N = 56) to gain an understanding of how these PD experiences impact their teaching practice.

Data collected underscores the benefits of collective participation and collaboration among educators across institutions and borders. Teachers acknowledged the positive impact of assessment training on their curriculum and instructional practices, emphasizing the value of cross-institutional and international collaboration in building professional identity. Workshop participants reported that collaborations between US and SEA instructors help them bridge cultural gaps and gain a deeper appreciation for the diverse perspectives and experiences of students from different backgrounds and of needs. Despite the varying teaching contexts and diverse student backgrounds, including students' first language, heritage identities or learning motivation, the collaborative process helped educators enrich their teaching resources and foster a more inclusive and effective learning environment.

Symposium 4

Beyond "In the Loop": Human-AI Collaboration in Human-Centered Language Assessment

Chair(s): **Eunice Eunhee Jang** (OISE / University of Toronto, Canada)
Discussant(s): **Xiaoming Xi** (Hong Kong Examinations and Assessment Authority)

Sunday, June 8, 2025, 8:30am to 10:30am Location: Poonsapaya

Over the past few years, artificial intelligence (AI) has significantly reshaped the field of language assessment, influencing its design, operation, and utilization. AI-generated tests and automated scoring have become commonplace, making prompt engineering an essential skill for assessment experts. However, with increasingly diverse human interactions across sociolinguistic boundaries, it is critical to integrate AI technologies that prioritize human values and capabilities at the core of assessment design, implementation, and use. This symposium seeks to explore how a human-centered approach to AI in language testing and assessment can enhance validity, reliability, and fairness.

The primary objective of this symposium is to explore the current integration of AI technologies in language assessment, particularly in the design, scoring, deployment, and validation of language assessments. We aim to critically examine how AI technologies are used in the field of LT/A today, highlighting human-centered AI approaches, in terms of both their strengths—such as efficiency and consistency—and their limitations, including potential biases and challenges in capturing complex language use. Building on this foundation, five papers will delve into different approaches to Human-AI collaborations to enhance fairness, reduce bias, and maintain a nuanced understanding of human performance.

The first paper, "Human-centered AI for Language Assessment Development and Administration," focuses on how human-AI teaming can combine AI's scalability and consistency with human expertise to ensure responsible and adaptable assessment practices. Practical examples illustrate how human oversight refines AI outputs, balancing efficiency and validity. The second paper, "Human-AI Teaming in Language Assessments through a Human-Centered Approach," discusses dynamic partnerships that leverage AI's speed and human contextual understanding to address validation challenges. Applications in speaking and writing assessments show how AI supports raters by automating aspects like pronunciation analysis and revision pattern identification, enhancing scoring reliability and adaptability to diverse linguistic backgrounds. The third paper, "Explaining AI Scoring Models to Humans," highlights the need for transparency in AI-based scoring models. It examines various interpretability approaches to make AI decisions clearer for stakeholders, advocating for building trust, identifying biases, and better aligning AI evaluations with human values. The fourth paper, "A Framework for AI in Language Testing: Bidirectional AI-Human Alignment," proposes a conceptual model where AI not only supports but dynamically interacts with and influences human behavior. The framework aims to inform test construction and validity models by highlighting the dual role of AI in test tasks and assessment processes. The fifth paper, "AI-driven Diagnostic Assessment Grows Together with Learners: Towards Individualized Actionable Diagnostic Feedback for L2 Speaking," explores the use of Reinforcement Learning from Human Feedback to provide personalized feedback.

This symposium will address the 46th LTRC theme by exploring how human-centered AI approaches can bridge cultural differences in language assessment practices. Through a focus on transparency, contextual adaptability, and personalized feedback, the symposium aims to offer insights into how AI can support fair and valid assessments across diverse sociocultural contexts, reflecting both Western and Eastern perspectives.

Presentations of the Symposium

Human-centered AI for Language Assessment Development and Administration

Jill Burstein, Geoffrey T. LaFlair, Alina. von Davier (Duolingo)

This presentation situates language assessment development and administration (LADA) with AI as a collaborative endeavor between people and AI systems that incorporates the strengths of both in LADA. People bring deep expertise in language assessment and learning as well as the ability to think critically about and interpret the outputs of AI models. The strengths of AI are related to scale, pattern recognition, and consistency. AI models can generate content at scale, encode relationships between linguistic features and scoring or generation criteria, and make consistent evaluations of performances on tasks.

We argue that human-AI teaming is a manifestation of human-centered AI (Auernhammer, 2020; Wang, 2019). This extends human-in-the-loop test development to position people not only as gatekeepers of AI, but as partners with AI in the LADA process. We argue that human-AI teaming is a mechanism for achieving responsible AI (RAI; Burstein, 2024, NIST, 2023) in the context of assessment development and for maintaining the validity of the interpretations and uses of test scores.

We demonstrate how human-AI teaming works through the lens of a digital-first assessment and provide examples of how it can improve content generation, difficulty estimation, automated-scoring, and proctoring. For content generation, human experts can review and edit AI-generated test content. The feedback from people is incorporated in the training and refinement of AI models (von Davier et al., 2024). AI models can be trained on human responses and feature data to learn item difficulty. These models can then be used to create item difficulty parameters for new items (Yancey et al., 2024; Yaneva et al, 2024). Scoring models incorporate human decisions as part of their training prior to their implementation. AI models can monitor and provide feedback to proctors in real time about the consistency of their decisions for certifying test scores (Belzak et al., 2024).

Human-AI Teaming in Language Assessments through A Human-Centered Approach

Eunice Eunhee Jang, Liam Hannah (University of Toronto)

Ongoing discussions in the field focus on validating AI-integrated assessment systems, particularly those based on general-purpose AI models like LLMs. These systems have the potential to improve efficiency and reduce human bias, but their evaluation is often constrained by task-specific limitations. This raises questions about how best to validate AI-driven assessments across diverse contexts and systems (Wang et al., 2023).

The shift from a traditional “in the loop” model, where AI assists while humans maintain control, to a collaborative human-AI teaming approach offers a promising path forward (Caldwell et al., 2022; National Academies Press, 2022). By leveraging the complementary strengths of AI and human experts, this partnership not only enhances efficiency but also provides a framework for addressing validation challenges. Human-AI teaming fosters adaptability and deeper engagement with language complexities, which can help ensure that assessments remain fair, reliable, and contextually relevant across diverse settings. This hybrid model improves scoring reliability, reduces bias, and better accommodates varied linguistic backgrounds by uniting AI’s speed with the nuanced, contextual understanding of human experts.

This paper explores how a human-centered approach can be integrated into language assessment, with specific applications in speaking and writing assessments. For speaking assessments, AI tools can support human raters by automating the analysis of features like pronunciation, prosody, and fluency, while human experts ensure that scoring remains contextually accurate and culturally sensitive. In writing assessments, we examine how human-AI collaboration can be used to recognize and automate the identification of writing revision patterns, combining AI’s efficiency with human insight to handle complex, nuanced aspects of writing performance. By positioning human expertise alongside AI capabilities, this dynamic teaming model encourages transparency and fairness, ensuring that assessments remain accurate, nuanced, and relevant across diverse contexts.

Explaining AI Scoring Models to Humans

Erik Voss (Teachers College, Columbia University)

Automated scoring systems have progressed from operationalizing constructs with rule-based Natural Language Processing developed by human language testing experts to deep learning architectures. Artificial Intelligence (AI) scoring models based on deep learning architecture can achieve state-of-the-art performance in automated scoring tasks but come at the expense of models becoming less understandable, limiting the acceptance in high-stakes language assessment contexts (Xiaoming, 2023; Van Moere, 2024). However, current systems are referred to as “black-boxes” because it is unclear how the system arrives at its results with no explanation for which characteristics of an input are important. In an effort to understand these opaque systems, research has attempted to explain how AI systems reach a decision, recommendation, or prediction in a way that humans can understand, helping stakeholders understand and trust the AI models. Successful explainability could produce trust in the system and provide insights to identify unintended biases, risks, and areas for performance improvements.

This paper provides an overview of approaches that have been taken to understand and explain AI models. Complementary approaches including, post-hoc interpretability, intrinsic interpretability, and model transparency with diagnostic tools are methods that are applied at different stages in the development and training of AI models. These approaches have strengths and weaknesses. The combination of these approaches might maximize both the performance and explainability of AI models. A better understanding of how AI models evaluate natural language in alignment with human values is essential for understanding the role these scoring models will play in language assessment.

A Framework for AI in Language Testing: Bidirectional AI-Human Alignment

Alistair Van Alistair, Jing Wei (MetaMetrics)

AI's evolution has led to advances in assessment development and scoring. There is also a large untapped potential for incorporating AI into tests and the test taker experience, by, for example, representing real-world tasks wherein a test-taker can edit or review AI-generated text in response to a prompt. However, the pace of AI adoption within the language testing domain has generally been cautious, and consequently, we have yet to develop frameworks and validity models that sufficiently incorporate AI.

This paper lays out a framework for how AI should be incorporated in validity models, and how it can be positively implemented in assessment. One aspect of the framework is AI ethical principles which includes topics such as fairness and bias, explainability of the models, transparency concerning data privacy and training, and adversarial robustness. We believe these have been well documented as general principles (e.g. IBM, 2023), although not always well codified and understood in language assessment.

Another aspect of the framework is the bidirectional nature of AI-human alignment (Shen et al. 2024). Different from AI-human teaming which essentially aligns AI to humans, the bidirectional framework emphasizes the importance of aligning humans to AI, helping individuals and society adapt to AI advancements both cognitively and behaviorally. The limitation of a human-centered AI is that it is a static, unidirectional process (i.e., aiming to ensure that AI systems' objectives match humans) rather than an ongoing, mutual alignment (e.g. Leike et al, 2022). The bidirectional model views AI alignment as dynamic, allowing for ongoing mutual adjustment between AI and humans. It also considers human adaptation to AI, ensuring that the development process considers the end users' needs, capabilities and potential challenges in working with AI systems.

AI-driven Diagnostic Assessment Grows Together with Learners: Towards Individualised Actionable Diagnostic Feedback for L2 Speaking

Hiroaki Takatsu[1], Shungo Suzuki[2], Ryuki Matsuura[3], Miina Koyama[2], Yoichi Matsuyama[1] ([1]Equmenopolis, Inc., [2]Waseda University, [3]Carnegie Mellon University)

Language testers have admitted the potential of technologies such as AI in diagnostic assessment (DA; Alderson et al., 2015). One crucial evaluation criterion of DA is its consequential validity (Isbell, 2021), typically tested in terms of learning gains. Learning gains through DA can be realised by the identification of individual learners' profile (i.e., strengths, weaknesses) and individualised feedback (Chapelle et al., 2017; Lee, 2015). Meanwhile, for a reliable application of AI technologies, a large amount of data capturing the individuality of learners is required. However, it is unrealistic to collect sufficient data from target learners before they actually use the DA system. Therefore, to achieve AI-driven individualised DA, language testers might conceptualise DA practice as an ongoing process rather than a product, inviting learners as continuous contributors to assessment design (Alderson et al., 2015). We thus investigated the feasibility of AI-driven individualised diagnostic feedback, adopting one recent approach to aligning AI to human preferences—reinforced learning from human feedback (RLHF; Ouyang et al., 2024). Thirty-one Japanese learners of English repeated oral proficiency interview tasks six times as assessment and remedial learning tasks and received diagnostic feedback on their lexical use in speech. The feedback included the selected excerpts of students' unsatisfactory utterances and the possible paraphrases with vocabulary items slightly advanced to their proficiency level. To capture their individuality in feedback preference, they assessed each paraphrase in terms of actionability (cf. Lee, 2015)—how likely they can use the paraphrase by themselves—on a 5-point scale. We then built the prediction model of each learner's actionability scores of paraphrases. Results demonstrated the incrementally updating prediction models outperformed the batch-based models in prediction accuracy (75% vs. 61%), confirming the potential of RLHF for automatised individualised diagnostic feedback. We also highlight the challenge in the reliability of learners' actionability responses.

Symposium 5

Digitally Empowered Assessment of Interactional Competence in Second Language Contexts

Chair(s): **Yunwen Su** (University of Illinois, United States of America)
Discussant(s): **Carsten Roever** (University of Melbourne)

Sunday, June 8, 2025, 8:30am to 10:30am Location: Duangduen

This symposium explores how digital technologies can be leveraged to assess interactional competence (IC) in second language learners. Interactional competence (IC), the ability to co-construct meaning in specific socio-cultural and pragmatic contexts, is central to effective communication (Hall & Pekarek Doehler, 2011; Young, 2011). However, assessing IC remains a complex endeavor due to the co-constructed and context-specific nature of interactions (Galaczi & Taylor, 2018).

The assessment of IC, especially in L2 learners, requires tasks that reflect the social and dynamic nature of interactions. Paired or group tasks often achieve this goal but present logistical challenges due to the need for trained raters and human interlocutors. Additionally, variability due to interlocutor effects and differing interactional styles can complicate scoring. Recent advancements in digital technology, such as AI, virtual/augmented reality (VR/AR), and video conferencing platforms, offer novel solutions for these challenges by providing immersive and interactive environments that closely mirror real-world communication contexts. In fact, with the growing prevalence of digital communication in professional, academic, and social spheres, these technologies go beyond mere simulation, providing opportunities to assess IC in contexts that reflect actual digital interaction. The ongoing development of AI and other digital technologies also enhances the practicality of administering and scoring IC assessments. Potentially, such advances would enable immediate, AI-generated diagnostic information and feedback tailored to individual test takers. Over time, test takers, who are also learners, would be guided toward better interactional skills, which aligns with principles of learning-oriented assessment (May *et al*., 2020).

Despite these opportunities, there is a notable gap in the literature regarding the systematic integration of digital technologies into L2 IC assessment. While there has been considerable research on the use of digital tools in language assessment and learning, much of this work has focused on receptive or productive skills in isolation, rather than on the interactive process that is central to communication. This symposium aims to fill this gap and advance our understanding of how digital tools can redefine the construct of IC and its assessment. It brings together empirical studies examining the construct of IC (Suzuki *et al*.), rating scales (Xiao & Jin), and innovative approaches to IC assessment (Shen *et al*.; Su & Chen; Youn) in digital contexts as well as a critical review of current research about AI-powered IC assessment (Timpe-Laughlin *et al*.), followed by a synthesis by two discussants (Dai & Roever). Overall, the symposium contributes to a more comprehensive framework for evaluating IC that accounts for the evolving nature of communication in a digital age.

Presentations of the Symposium

AI Familiarity as a New Potential Source of Bias in Interactional Competence Assessment with Conversational Agents

Shungo Suzuki[1], Hiroaki Takatsu[2], Kotaro Takizawa[1], Ryuki Matsuura[3], Mao Saeki[2], Yoichi Matsuyama[2] ([1]Waseda University, [2]Equmenopolis Inc., [3]Carnegie Mellon University)

Given the spontaneous nature of real-world communication, L2 researchers and language testers have admitted the importance of assessing interactional competence (IC), which is defined as the ability to co-construct a purposeful interaction in a socially appropriate manner (Galaczi & Taylor, 2018). Meanwhile, for the valid assessment of IC, Galaczi and Taylor (2018) highlighted the challenges in balancing between interactional authenticity and variability. To solve this issue, the potential of conversational AI as a controllable but authentic interlocutor has been advocated and empirically tested (Gokturk & Chukharev-Hudilainen, 2023; Ockey & Chukharev-Hudilainen, 2021; Saeki et al., 2024). However, the use of conversational AI may introduce some construct irrelevant variance of test scores, for instance, by learners' familiarity with AI technologies (e.g., synthetic speech). To this end, the current study examined the contribution of AI-related attributes to IC scores assessed with AI interlocutors while controlling for linguistic and pragmatic competence.

Japanese-speaking learners of English (N = 100) completed three roleplay tasks with conversational AI (Saeki et al., 2024), as well as linguistic and pragmatic knowledge tests. Their roleplay task performance was automatically assessed on our own CEFR-based IC scale (Authors, XXXX). The linguistic and pragmatic tests include Productive Vocabulary Levels Test (Laufer & Nation, 1999), Maze task (Suzuki & Sunada, 2018), and spoken discourse completion tasks (Kang et al., 2021). To measure learners' familiarity with AI-based materials, we adapted Horkay et al.'s (2006) computer familiarity questionnaire.

Structural equation modelling showed that the latent variable of IC was primarily explained by that of linguistic and pragmatic competence. Meanwhile, the latent variable of AI familiarity only showed the negligible relationship with IC performance. These findings will be discussed with regard to the complex interplay between IC assessment with conversational AI agents, assessment task design and individual difference factors.

An Ecological Perspective on Classroom Assessment of Pre-Service Teachers' Interactional Competence Using Technology-Mediated Speaking Tasks

Soo Jung Youn (Daegu National University of Education)

This mixed-methods study examines the classroom assessment practices of Korean EFL pre-service teachers' classroom interactional competence in a technology-mediated, task-based speaking classroom. In order to understand the complex ecology of how pre-service teachers interact with the technology-mediated speaking tasks and classroom contexts, Chong and Isaacs' (2023) notion of Language Assessment Ecology was employed as the conceptual framework. The following research questions guided the study: (1) How does the classroom interactional competence of Korean EFL pre-service teachers develop in a technology-mediated task-based speaking classroom?; (2) What engagement, contextual, and learner factors mediate Korean EFL pre-service language teachers' engagement with the technology-mediated speaking assessment tasks? In alignment with real-life teaching situations, various speaking tasks were designed for both pedagogical and assessment purposes. These tasks included a combination of technology-mediated and face-to-face interactive speaking tasks. The technology-mediated speaking tasks involved the use of ChatGPT and Google Assistant on participants' mobile devices for self-assessment, technology-mediated interaction, and supplementary tools for completing complex speaking tasks. Following Walsh's (2006, 2013) model of classroom interactional competence informed by Conversation Analysis, the speaking tasks were designed to reflect various pedagogical goals of classroom interactions (e.g., managing instruction, providing corrective feedback, establishing a classroom context). These tasks elicited a range of interactional features (e.g., extended teacher turns, the use of transition markers, display questions, corrective repair). Multiple data sources were collected throughout the semester, including pre- and post-semester questionnaires, transcriptions of speaking task outcomes, reflection papers, and focus group interviews. The quantitative and qualitative findings are discussed in terms of the potential of technology-mediated speaking tasks for teaching and assessing classroom interactional competence. Furthermore, the ecological perspective on classroom assessment practices is discussed, emphasizing the importance of understanding unique contextual variables that mediate pre-service teachers' development of interactional competence.

Assessing TAlkEZly Mediated L2 Interactional Competence Development in blended Teaching Context

Chen Shen, Yaru Meng, Xi Qian (Xi'an Jiaotong University)

Interactional Competence (IC) is an important component in L2 speaking. Among the major IC features are four dimensions of verbal resources (i.e. turn management, topic management, interactive listening, and breakdown repair; e.g., Galaczi & Taylor, 2018). However, the longitudinal classroom-based mediation and assessment of these features are under-explored. Such issues could partly be addressed in Blended Teaching (BT) by leveraging the strengths of face-to-face and technology-mediated instruction and assessment (Graham, 2006). With chatbots beginning to undertake the role of L2 partners, the related research attempts are on the rise. Nevertheless, not enough attention is given to the integration of specially designed chatbots in BT in facilitating L2 learners' IC. To address this, the current study presents a 8-week pseudo-experiment of a TalkEZly (a special GenAI chatbot) mediated blended teaching and then assesses how learners develop in IC of the dimensions in question. A total of 108 first-year college L2 learners participated in the study, with the experimental group (n=58) receiving IC-based micro-courses before class, teacher guidance and peer practice in class, and GenAI partner-mediated interactive tasks after task. The control group (n=60) followed the same procedure except for the after-class TalkEZly-mediated interactions. Questionnaires to elicit the learners' demographic data and individual differences. In addition, focus group interviews were conducted for triangulation purpose.

Findings revealed that while the experimental group outperformed the control group, but not at all the dimensional level. This may be attributed to TalkEZly-related factors and their interactions with learner factors. Pedagogically, the study sheds light on the AI-empowered assessment and mediation in speaking classrooms. Technically, it inspires the further optimization of specialized AI chatbots.

Measuring L2 Interactional Competence: A Comparison of Human and AI-Mediated Roleplay Assessments

Yunwen Su[1], Xi Chen[2] ([1]University of Illinois, [2]University of Central Lancashire)

Recent research on the assessment of speaking highlights interactional competence (IC) as a key construct. Studies have shown that IC can be measured through various interactional features (e.g., Ockey et al., 2023; Roever & Kasper, 2018; Youn, 2020). Assessing IC typically involves tasks that prompt meaningful, interactive language use, such as roleplays and discussions. However, such tasks often require another speaker, which may reduce practicality (Ockey & Chukharev-Hudilainen, 2021) and increase rating complexity (Su & Shin, 2024), alongside interlocutor effects (Galaczi & Taylor, 2018). This mixed-methods study investigates the potential of using generative AI as an interlocutor for assessing IC in L2 English by comparing it with a native-speaking human peer interlocutor. The research addresses: (1) how interlocutor type (human vs. AI) impacts the severity and consistency of IC ratings; (2) differences in interactional features elicited by human versus AI interlocutors; and (3) raters' perceptions of AI interlocutors. Forty test takers completed a 6-item roleplay task targeting refusals of requests, invitations, and offers with a native-speaking human and an AI interlocutor (ChatGPT-4) two weeks apart. Four trained raters assessed the audio-recorded performances using a data-driven rubric covering two IC domains: Interactive Listening (supportive listening, comprehension efforts, smoothness, repair) and Sequential Organization (position, modification, justification, context awareness). Many-Facet Rasch Measurement (MFRM) compared IC ratings across interlocutors, raters, scenarios, and domains (RQ1). Interactional features were manually coded and analyzed quantitatively (correlations, multiple regressions, discriminant analysis) to identify reliable predictors of IC scores (RQ2). Finally, thematic analyses of rater interviews explored their perceptions of human versus AI interactions (RQ3).

Assessing Interactional Competence in a Digital Era: Developing and Validating a Rating Scale for a Computer-Based Paired Discussion Task

Sa Xiao, Yan Jin (Shanghai Jiaotong University)

Technological advancements in speaking assessment, such as video conferencing systems, provide valuable opportunities to capture the co-constructed nature of Interactional Competence (IC) in technology-mediated language assessments (Zhang & Jin, 2021). However, IC has yet to be systematically conceptualized, posing challenges in operationalizing the construct and evaluating performance on oral interactional tasks. This study addresses the need for assessing IC in computer-based paired speaking tasks, focusing on the College English Test — Spoken English Test Band 4 (CET-SET4) for tertiary-level English learners in China. The research aims to develop and validate an empirically-driven IC rating scale for the CET-SET4's paired discussion task.

A multi-phase, mixed-methods sequential design was employed to conceptualize, operationalize, and validate the IC rating scale. Initially, a preliminary IC framework was developed through thematic analysis of dyadic interactions and document review. Six macro components of IC were identified: Topic Management, Interactive Listening, Turn-taking Management, Breakdown Repair, Manner of Interaction, and Understanding the Situation and Context. These components, along with their subcategories, formed the foundation for the IC rating scale. The scale was refined and validated through expert workshops and a questionnaire survey to calibrate descriptor difficulty. A rating experiment involving 20 raters assessing 30 paired discussion performances was conducted. Results were analyzed using the Many-Faceted Rasch Model (MFRM) to examine rater consistency, task difficulty, and overall scale performance. Qualitative interviews with raters provided additional insights.

The findings contribute to the conceptualization of the IC construct, offer refinements to the underlying construct for computer-based paired discussion tasks, and highlight what raters can feasibly assess in real time. This research has significant implications for developing IC rating scales in various digital interaction contexts, addressing the growing need for assessing IC in technology-mediated language assessments.

Paper and Demo Summaries – Friday, June 6, 2025

Research Papers

Time: **Friday, 06/June/2025: 11:00am - 12:30pm** *Location:* **Ampai**

Promoting Multiculturalism in Arabic: A Pan-Arab Initiative for Language Preservation and Literacy Enhancement

Hanan Khalifa, Jing Wei, Alistair Van Moere
MetaMetrics Inc, United States of America

Arabic, one of the most widely spoken languages globally, occupies a crucial role in international communication and culture. Officially recognized by the United Nations and celebrated annually on World Arabic Language Day (Khan, 2021), Arabic faces numerous challenges. These include the impacts of globalization, the increasing dominance of English as a lingua franca, and shifting cultural dynamics, particularly within the Middle East and North Africa (MENA) region (Alfataftah & Jarrar, 2018). The preservation of Arabic's linguistic heritage is further complicated by its coexistence with a variety of regional languages and dialects.

In response to these challenges, the Miqyas Al Dhad initiative was introduced as a pan-Arab effort to enhance reading proficiency in Arabic. This initiative aims to assess and monitor literacy levels, measure text complexity, and build a database of age-appropriate books for learners. A consortium of linguists, educators, and assessment experts from across the Arabic-speaking world collaborated on this project. The initiative has developed over 120 reading assessments for learners aged 7-18, grounded in curricular standards from a range of Arabic-speaking countries. Using natural language processing (NLP) techniques, a corpus of over 2,000 textbooks has been analyzed to align student reading levels with text complexity. The initiative aspires to establish a universal framework for Arabic literacy assessment, applying psychometric and NLP methodologies.

This paper will examine the development of the Miqyas Al Dhad framework, addressing the challenges and solutions encountered, with a focus on stakeholder engagement and overcoming cultural and logistical barriers across the diverse Arabic-speaking world.

Translanguaging in Listening Assessment: Inclusion of L1 Responses in L2 Recall Tasks

Shelly Xueting Ye[1], Matthew Wallace[2]
[1]City University of Macau, Macau S.A.R. (China); [2]University of Macau, Macau S.A.R. (China)

Listening recall tasks, which require test takers to produce responses in an L2 based on audio input, may unintentionally assess L2 writing skills, thereby compromising the accurate assessment of listening abilities. This study addresses this issue by employing a mixed-methods approach, informed by the notion of translanguaging, to examine the effect of incorporating L1 responses in L2 recall tasks. The inclusion of L1 responses is explored as a potential means of reducing the influence of L2 writing proficiency on task outcomes and enhance task validity. In this investigation, a listening recall task requiring L1 responses was developed and administered to 102 learners with low and intermediate proficiency levels in English. Their performance was compared with that on a control task that required L2 responses. Additionally, follow-up interviews provided further insights. The results from both quantitative and qualitative analyses revealed that allowing L1 responses in recall tasks enhanced performance among intermediate-level learners, improved the accuracy of the listening construct representation, and was favored by participants as an effective method of assessing listening skills. These findings suggest that the use of L1 responses can enhance both the construct and face validity of recall tasks, thereby addressing initial concerns about validity. Consequently, this study not only contributes to the advancement of listening task development but also advocates for the integration of translanguaged practices in language assessment, promoting broader implementation of translanguaging in language education.

Interactive Videos in an L2 Listening Test: How do They Affect Test Scores?

Shanshan He
University of Western Ontario, Canada

L2 listening is widely recognised as a multimodal language skill that entails processing and understanding both auditory and visual information. However, the listening sections of most high-stakes standardised language tests still avoid video input, and one reason is the lack of understanding of how different visuals affect L2 listeners' test performance. It is also believed that processing visual information while answering questions and/or taking notes in a video-based listening test can be a cognitively demanding task for L2 test takers. To reduce the cognitive load on test takers, this study utilised an academic listening test with interactive videos (IV), which are videos divided into small "digestible" clips, each of which is followed by an embedded question.

This study explored whether IVs impacted listening test performance differently from traditional linear videos (TV) and examined how L2 learners' language proficiency and preferences of the video type affect their performance on listening tests with different types of video input. The reasons for participants' preferences towards certain video types complemented the quantitative results. The results demonstrated that participants performed better on the test items associated with TVs than on IVs. While participants' English proficiency and their preferences positively predicted their listening test performance, none of these results was statistically significant. This study contributes to research on video-based listening assessments by expanding this line of research to explore L2 learners' interaction with visual input and test items. This study also identified a significant gap between the design of assessments and the context of implementing them.

Research Papers

Time: **Friday, 06/June/2025: 11:00am - 12:30pm** *Location:* **Phramingkwan**

L1 Intuition vs. Three Popular LLMs: Can LLMs Mark L2–L1 Meaning Recall Tests?

Aaron Olaf Batty[2], Jeffrey Stewart[1], Laurence Anthony[3], Christopher Nicklin[4], Keita Nakamura[5], Kanako Tomaru[1], Stuart McLean[6]
[1]Tokyo University of Science, Japan; [2]Keio University, Japan; [3]Waseda University, Japan; [4]University of Tokyo, Japan; [5]Eiken Foundation of Japan; [6]Kindai University, Japan

Constructed response L2–L1 meaning recall tests have been demonstrated to be more predictive of reading ability than multiple-choice meaning recognition tests. However, because they require the marker(s) to be proficient in students' L1(s) in order to mark responses competently, the time and effort required can be substantial. Advances in large language models (LLMs), however, hold promise for automating the marking of such tests.

The present study explores the possibility of using LLMs to mark L2–L1 meaning recall tests using the responses of 611 Japanese learners of English to a 150-item test marked by two two L1-Japanese human markers and three popular LLMs: GPT (4o), Gemini (1.50 Flash), and Llama (3-8b). The data were analyzed with intraclass correlations, ANOVA, reliability analysis, and many-facets Rasch measurement (MFRM).

Correlations among all of the markers' scores and reliability indices were high. Post-hoc tests revealed significant differences between the human markers and both Llama and Gemini, which also had high Rasch Outfit indices. GPT was found to be roughly equally severe as the human markers, whereas Llama was stricter and Gemini more lenient. Thirty words with small-to-large human–LLM biases were identified. Close inspection of these words revealed that the humans used more information to judge than instructed, whereas the LLMs followed the prompt more closely.

The study demonstrates that using GPT or Gemini to mark L2–L1 meaning recall responses likely does not harm test reliability, but that current LLMs may not act in ways that an expert human marker

---

An Agile Approach to Utilizing AI Technology to Support Young EFL Students' Writing Skills

Mikyung Kim Wolf, Michael Suhan
ETS, United States of America

With the rapid advancement of generative AI technology using large language models, various AI tools have emerged for educational use. Among these tools, ChatGPT has gained traction in L2 writing due to its tremendous potential as a useful writing assistant for ESL/EFL learners. The present study explored how young EFL students interact with and utilize feedback generated by GPT-4 for their writing. This study was conducted within the context of developing an automated writing feedback tool to support potential TOEFL Junior Writing test users. Initially, we developed a prototype tool and employed an agile approach to incorporate users' feedback into its development. During this prototype phase, we implemented a series of prompt engineering exercises to provide feedback aligned with the construct and scoring

rubric of the TOEFL Junior Writing test. In our prototyping study, we addressed the following research questions:
1)What types of feedback do students seek in the tool?
2)How do students use feedback in their revisions?
3)What are students' perceptions of the GPT-based feedback tool?

A total of 14 students (average age = 12.8) participated in the study. Students' chat questions and writings were analyzed to identify patterns in the use of GPT-based feedback. The results indicated that, on average, students' writings slightly improved compared to the first draft. The most frequent type of question from students concerned organization, followed by language use and content. In this presentation, we will demonstrate the prototype tool and discuss the implications and lessons learned from our prototyping study.

Research Papers

Time: **Friday, 06/June/2025: 11:00am - 12:30pm** *Location:* **Room 401**

Cognitive Processes in Intertextual Summary Tasks: A Study of Multilingual Asian Test-Takers

Nathaniel Ingram Owen[1], Haiyan Xu[2], Oliver Bigland[1]
[1]Oxford University Press, United Kingdom; [2]University of Leicester

This study investigates the cognitive validity of an intertextual reading-into-writing summary task for academic admissions testing, examining how well current English proficiency tests align with actual academic writing requirements in higher education settings.

The research explored cognitive strategies employed by fifteen university students from India, China, Kazakhstan, and Oman through online think-aloud interviews. Participants completed a task requiring synthesis of two 150-word texts (a textbook extract and lecture transcript) into an 80-100 word summary. The study combined video-informed coding of visual behavior and verbalizations, analyzed through hierarchical clustering and mixed-effects modeling. Task performance was evaluated using an analytic scale incorporating CEFR C1-level mediation descriptors, with coding schemas informed by established cognitive models of reading and writing.

Analysis revealed substantial rater agreement of the coding scheme (Cohen's Kappa = .64) and demonstrated that the task successfully elicited appropriate organizational and synthesizing strategies across participants. The findings showed varying patterns of recursion to source texts, with participants generally favoring the first text. Notably, participants demonstrated substantial integration and transformation of source material through coherent recycling and paraphrasing from both texts, supporting the task's authenticity in assessing academic writing skills.

These findings enhance our understanding of how multilingual students engage with academic writing tasks and offer insights for teaching and assessment in multicultural contexts. The research highlights the importance of key academic skills like intertextual reading, paraphrasing, and information synthesis, contributing to the development of language assessment practices that reflect the diverse linguistic landscapes of contemporary academic environments.

Evaluating the Integration of Listening Comprehension in Paired Oral Communication Tests

Leyla Karatay[1,2]
[1]Iowa State University, USA; [2]Duzce University, Turkiye

This study explores the integration of listening as a distinct dimension in paired oral communication tests, addressing a gap in current rating scales that primarily focus on speaking and interactional competence. Although listening is a fundamental component of communication, its absence in these assessments leads to an underrepresentation of this critical skill (Buck, 2001; Vandergrift, 2007). Listening's complexity, involving emotional, cognitive, and behavioral dimensions, further complicates its assessment, particularly when it is subsumed under broader speaking categories (Aryadoust & Luo, 2023). To address this issue, a new rating scale was developed using the Evidence-centered Design framework (Mislevy et al., 2003), which includes an explicit listening comprehension component alongside interactional competence, pronunciation, fluency, and grammar/vocabulary.

The study involved 64 L2 students from diverse cultural backgrounds at a large U.S. university, with 32 video-recorded paired oral performances rated by six trained raters. A three-facet partial credit model using many-facet Rasch measurement (MFRM) analyzed 384 composite scores across examinees, raters, and scale dimensions. Rater consistency in evaluating listening was assessed using fit statistics. To gain insight into raters' decision-making processes, surveys, semi-structured interviews, and think-aloud sessions were conducted.

The findings from this study provides critical insights into how an explicit listening dimension can improve the construct representation of oral communication tasks in paired oral tests, ensuring fairness and validity in this context. Suggestions for future scale development, rater training, and further empirical research on assessing listening comprehension in paired oral tasks will be discussed.

Language Assessment Literacy Inventory Development: Understanding Learner Perspectives in Multilingual Contexts

Jiyoon Lee[1], **Yuko Butler**[2]
[1]University of Maryland Baltimore County; [2]University of Pennsylvania

This study addresses a significant gap in Language Assessment Literacy (LAL) research by developing and validating the first comprehensive inventory measuring English language learners' understanding of assessment practices. While LAL has traditionally focused on teachers and administrators, this research recognizes learners as key participants in the assessment practices. The study employed a mixed-methods approach, beginning with theoretical framework development based on Davies' (2008) LAL conceptualization and incorporating sociocultural perspectives. The inventory was developed through expert review and pilot testing, culminating in a validation study with 144 participants from 37 countries and 28 language backgrounds.

The final inventory measured four constructs: Assessment Knowledge, Assessment Principles, Assessment Skills, and Affective Factors, demonstrating strong overall reliability ($\alpha=0.942$) and robust construct-specific reliability (α ranging from 0.745 to 0.918). Results revealed that learners showed strongest understanding in Assessment Knowledge (M=4.32, SD=0.89) and valued feedback highly (M=4.54, SD=0.76), while showing less confidence in peer assessment (M=3.05, SD=1.23). Notable variations in assessment perceptions across cultural backgrounds emphasized the need for culturally sensitive assessment practices.

This research provides a validated tool for understanding learners' LAL needs and developing more inclusive assessment practices. The study's focus on learners' perspectives and cultural considerations makes it particularly relevant for advancing culturally responsive assessment practices in language education.

Research Papers

Time: **Friday, 06/June/2025: 11:00am - 12:30pm** *Location:* **Room 405**

Gender Representation in IELTS and NMET Reading Texts: A Comparative Analysis Using SFL and Social Actor Network Frameworks

Xiaoqin Huang, Xiangdong Gu
Chongqing University, Chonqing, People's Republic of China

This study compares gender representation in reading texts from ten IELTS tests (2013-2023, no official publication in 2014) and ten China’s National Matriculation English Tests (NMET, 2014-2023) using Systemic Functional Linguistics (SFL) and Social Actor Network frameworks. Drawing on the ETS Guidelines for Developing Fair Tests and Communications (GDFTC) and the 2024 UNESCO report on gender disparities, the research analyzes 25 IELTS texts and 27 NMET texts, focusing on gendered characters, pronoun usage, and professional roles. The analysis reveals a male-dominated representation in both tests, particularly in IELTS, where males account for 83% of appearances and 86% of pronoun usage, compared to NMET’s 58% male appearances and 56% pronoun usage. Female representation is notably limited, with binary gender expressions prevalent, and no acknowledgment of other identities. The study raises concerns about the impact of such biases in educational materials and emphasizes the importance of a holistic approach in test development to promote gender fairness.

Assessing Second Language Pragmatic Competence for Intercultural Communication: Test Localisation Targeting UK Pre-Sessional Students

Shishi Zhang
University College London, United Kingdom

Second Language (L2) pragmatic competence is deemed an essential component of language proficiency and vital to successful communication (Taguchi & Ishihara, 2018). Despite advances in assessing L2 pragmatic competence over the past four decades, challenges remain, including assessing pragmatic-relevant skills mirroring real-life language use, and operationalising a locally relevant test addressing diverse stakeholders’ needs and perceptions. The goal of this multi-phase project is to develop a needs-driven oral assessment tool of L2 pragmatic competence for intercultural communication in the academic domain targeting UK pre-sessional students. Adopting a pragmatist stance and drawing on the Socio-Cognitive Framework (O’Sullivan & Weir, 2011), the study employs a multi-stage exploratory sequential mixed methods design (Creswell & Plano Clark, 2018) and systematically engages target end-users (pre-sessional teachers and students) in the development and initial validation of the assessment tool. The researcher first triangulated a priori evidence from thematic analysis of 115 pre-sessional documents from seven UK universities coupled with semi-structured interviews with 16 pre-sessional teachers and eight pre-sessional students from 10 additional universities. Triangulation of the results culminated in two task types: (1) a group oral with textual input assessing pragmatic knowledge and use; and (2) video analytical tasks with filmed roleplays followed by questions assessing metapragmatic awareness. The tasks were then iteratively evaluated by seven pragmatics testing experts and experienced pre-sessional teachers through focus groups and written feedback, revised accordingly, and trialed with 11 pre-sessional students. Implications for enhancing washback in academic preparatory courses through a focus on groupwork and metapragmatic awareness will be discussed.

Assessing Language for Professional Registration: Does the IELTS Academic Capture Australian Teachers' Communicative Demands?

Xiaoxiao Kong
University of Melbourne, Australia

Since 2011, the IELTS Academic has served as an English language proficiency test for teacher registration in Australia at early childhood, primary, and secondary levels. The validity and appropriateness of this practice, however, have rarely been investigated.

Guided by an argument-based validation framework (e.g., Knoch & Chapelle, 2018) and needs analysis for language assessments for professional purposes (LAPPs; Knoch & Macqueen, 2020; Long, 2005), this study investigates the linguistic and communicative demands of early childhood and school teachers in Australia, as well as the appropriateness and adequacy of the IELTS Academic for assessing English language proficiency for teacher registration. Document analysis (n = 106), focus groups (n = 37), survey (n = 123) and interviews (n = 15) were conducted in three sequential stages to explore language use characteristics of important and frequently occurring workplace communication tasks, as well as teachers' views on the domain relevance of the IELTS Academic test tasks. Findings suggested differences in teachers' language demands across education levels, which translated to differences in the degree of domain representativeness of the IELTS Academic within the three education contexts. Additionally, a mismatch was found between the IELTS Academic and teachers' workplace communication in terms of task format and characteristics of expected response, raising concerns over the validity and appropriateness of the IELTS Academic for teacher registration purposes. Such investigations provide implications for policy formulation as well as the design and implementation of language assessment for teacher registration, which could in turn contribute to student outcomes

---

Research Papers

Time: **Friday, 06/June/2025: 11:00am - 12:30pm** *Location:* **Poonsapaya**

Survive and Thrive in a third Space: When Chinese Students Come to NZ

Qiuxian Chen[1], Gavin Brown[2], Yue Wang[2]
[1]shanxi University, China, People's Republic of; [2]University of Auckland, New Zealand

This study investigates the learning and assessment challenges faced by Chinese students in New Zealand. Six Chinese doctoral students enrolled in an Academic English Writing program participated in an in-depth focus group interview. The research adopts a critical incident approach to capture how these students navigated the contrasting assessment regimes between China and New Zealand.

The analysis reveals significant tensions between formative (learning-focused) and summative (evaluation-focused) assessment practices and values, with the students facing considerable difficulties in adapting to the new academic culture. The students' experiences highlight how their prior academic knowledge, learning strategies, and approaches to assessment were not always easily transferable to the new context. This forced them to negotiate a 'third space'—a liminal zone where the practices and expectations of their home and host countries intersect. In this third space, students had to reconcile their previous experiences with new modes of learning and assessment, often felt disoriented and struggled to balance the demands of the new and previous academic systems, and ended up with a unique set of adaptive strategies.

The findings underscore the complexities inherent in transitioning between these two disparate educational environments and highlights the broader implications of transnational education and the challenges that arise when East meets West in higher education. While adding useful suggestions to the preparatory package of prospective international students, this study hopes to provide insights for the assessment systems which have been seeking to establish formative assessment frameworks within their summative-dominated assessment cultures in the past two decades, such as China and other Asian contexts. It also underscores the complexities of international students' survival and thrival in the academic track and offers insights into how institutions can better support international students in navigating the diverse and sometimes conflicting expectations of different assessment cultures.

One Step Further: Understanding The Role of a Tailor-Made Genai-Powered Chatbot in Affecting L2 Learner Interaction and Engagement

Shangchao Min[1], Yuhong Gao[1], Jie Zhang[2]
[1]Zhejiang University, China; [2]Shanghai University of Finance and Economics, China

Generative artificial intelligence (GenAI), utilizing deep-learning models (Son et al., 2023; Niloy et al., 2024), engages students in natural, targeted, and context-aware interactions, enhancing the effectiveness of diagnostic assessments that identify learning needs. However, the extent to which GenAI-assisted adaptive learning based on diagnostic assessment results outweighs GenAI-assisted learning without adaptive mechanism awaits investigation. This study explores students' learner-chatbot interaction patterns and learner engagement when interacting with E-Talk in two task modes: free dialogue (FD) and thematic dialogue (TD).

Ninety students participated in this study, which lasted for one month. Data from multiple sources were collected to investigate learner-chatbot interaction and learner engagement in the two task modes, including records of learner-chatbot interactions, learner engagement questionnaire results, and transcripts of semi-structured interviews. The results showed distinct differences in the wordlists produced in the two task modes. The interaction pattern in the FD mode demonstrated that GenAI tended to introduce synonyms to expand the students' vocabulary for lower-level proficiency students, while extending from using synonyms to elaborating on new concepts for higher level proficiency students. However, students exhibited higher levels of emotional engagement in FD than TD, with no significant differences found for the other three types of engagement (i.e., behavioral, cognitive and social engagement) across the two task modes.

This study provided empirical evidence for the importance of diagnosis-enhanced designs in GenAI-powered L2 education and contributed to a wider discussion on the importance of diagnostic assessment in facilitating GenAI's role in fostering individualized learning.

Exploring Cultural and Pragmatic Challenges in AI-mediated Speaking Assessments

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

This study explores how a prototype AI interlocutor spoken dialog system (SDS) can elicit interactive oral performances across different cultural groups in high-stakes speaking assessments. Recent advances in Conversational Artificial Intelligence (AI) have allowed systems to mimic sociopragmatic behaviors such as politeness formulas, but AI may struggle to adapt its language based on context and interlocutor characteristics. This limitation in AI poses challenges in assessing interactional competence, a key component of oral proficiency, which involves dynamic, real-time communication.

Drawing on Kagitcibasi's (1997) framework, the study investigates the perceptions and oral performances of 100 non-native English speakers, divided into two groups: 50 from individualistic societies (e.g., Western Europe) and 50 from collectivistic societies (e.g., China, India, Japan). Participants' performances were rated by four trained raters using a modified Linguaskill General Speaking mark scheme. All participants completed a post-test survey, and semi-structured interviews were conducted with 10 participants and the four raters. A subset of candidate speech was transcribed and occurrences of interactive language functions were annotated.

Results will be reported in three main areas: (1) discourse analysis of the transcriptions will focus on turn-taking, politeness strategies, and clarification requests to examine cultural differences in interactional competence; (2) Multi-faceted Rasch Measurement (MFRM) will assess score reliability and potential cultural bias; and (3) descriptive statistics and thematic analysis will provide insights into participants' and raters' perceptions of the AI interlocutor's interactional behaviour and its cultural sensitivity.

Research Papers

Time: **Friday, 06/June/2025: 11:00am - 12:30pm** *Location:* **Duangduen**

Construct Comparability of TOEFL iBT and Duolingo English Test

Sara Cushing
Georgia State University, United States of America

This paper presents the results of a comparative construct analysis of the TOEFL iBT and the DET, relying on publicly available information about the two tests. I highlight similarities and differences between the two tests in terms of construct (what the test is intended to measure), cognitive and linguistic demands of test tasks, and how tasks are scored and items weighted. Implications for test score users are discussed.

Distinct Listening Biotypes and their Application in Test Validation: A Neuroimaging Study

Vahid Aryadoust
National Institute of Education, Nanyang Technological University, Singapore

This study introduces the concept of listening biotypes, which are distinct subtypes of listening processes characterized by specific brain activation patterns during listening tasks under both assessment and non-assessment conditions. The research aims to address limitations in traditional test validation frameworks, particularly the reliance on self-reported cognitive processes in listening assessments, which do not accurately capture unconscious cognitive mechanisms.

Using neuroimaging evidence from 109 participants in a functional near-infrared spectroscopy (fNIRS) study, we scanned regions involved in listening: the left dorsomedial prefrontal cortex, inferior frontal gyrus, and posterior middle temporal gyrus. Oxygenated hemoglobin levels, indicating neural activity, were measured and analyzed using a two-step clustering method. After comparing several models, a three-cluster solution emerged as the best fit, revealing three distinct listening biotypes, each associated with unique neuronal activation patterns.

In an ongoing follow-up study, we are examining the relationship between these biotypes and listening test scores, test methods, and test types. Two hypotheses are being tested: whether the biotypes are dependent on specific test methods or types, or whether they are independent of these “artefacts.” Ultimately, this study aims to offer a novel framework for understanding and categorizing listeners based on their neurocognitive responses, contributing to the refinement of listening test validation practices.

Metacognitive Awareness and Its Relationship with L2 Listening in a Model of Language Proficiency Using a Meta-Analytic Structural Equation Modeling

Yo In'nami[1], Mike W.-L. Cheung[2], Rie Koizumi[3], Matthew P. Wallace[4]
[1]Chuo University; [2]National University of Singapore; [3]University of Tsukuba; [4]University of Macau

Understanding language proficiency is crucial for research on validity and validation, as researchers must clearly define what aspects of proficiency are being measured to ensure the validity of the inferences drawn from test scores. In this respect, models of language proficiency are helpful. According to Hulstijn’s (2015) model, language proficiency consists of core components and peripheral components. While the model suggests relationships between language proficiency and core and peripheral components, the studies that informed the model were largely based on a series of investigations conducted in the Netherlands. To address this gap, we focused on one peripheral variable—metacognitive awareness—and examined its relationship with second-language (L2) listening. We collected and analyzed 29 Pearson correlation matrices that measured metacognitive awareness using the Metacognitive Awareness Listening Questionnaire (MALQ; Vandergrift et al., 2006) and its relationship to L2 listening. These studies were conducted in 15 countries. Results from meta-analytic structural equation modeling showed that a single-factor model fit the data well, consistent with previous studies (e.g., Goh, 2018). However, the varied factor loadings indicated that problem-solving, planning-and-evaluation, and directed attention strategies contributed more strongly to metacognitive awareness, while person knowledge and mental translation strategies had minor roles. Metacognitive awareness moderately predicted L2 listening (b* = .306), explaining 9.364% of the variance in listening, further supporting its peripheral role in Hulstijn's (2015) model of language proficiency. This relationship was moderated by listening test types (selected-response only vs selected-response and open-ended) and participants groups (secondary-school vs. university students).

Research Papers

Time: **Friday, 06/June/2025: 1:30pm - 3:00pm** *Location:* **Ampai**

Leveraging Generative AI for Interactive Assessment in Multicultural Contexts

Inyoung Na
Iowa State University, United States of America

This study explores the potential of Artificial Intelligence (AI) to enhance the assessment of interactional competence (IC) in large-scale, high-stakes oral assessments. Traditionally, IC, the ability to communicate effectively with others, has been assessed using paired or group tests. However, these tests introduce interlocutor variability due to differences in personality traits, proficiency, and speech varieties, making it difficult to assess individual performance. Attempts to address this through Spoken Dialogue Systems (SDS) have been critiqued for producing unnatural discourse. Recent advancements in generative AI technologies, specifically Large Language Models (LLMs), offer a promising alternative for creating more authentic and dynamic conversational partners, potentially reducing interlocutor variability and improving fairness in assessments, particularly in multicultural contexts.

This study compared the interaction patterns, test scores, and perceptions of test takers and raters when using human versus AI interlocutors in English Placement Tests (EPT). Ten international students participated in a four-minute paired discussion with either a human or AI interlocutor. Three experienced raters evaluated each speaker on fluency, pronunciation, IC, and grammar/lexis. Generalizability theory was used to examine the effects of test takers, raters, and testing conditions (AI or human) on score variability. The findings highlight the relative contribution of test takers to score variability and reveal both positive and negative perceptions toward AI versus human partners. Implications for validity, reliability, and authenticity will be discussed, along with recommendations for future AI-based assessments.

Can LLMs Generate Human-Like Responses for Training Fairer AES Systems?

Burak Senel
Iowa State University, United States of America

Automated essay scoring (AES) systems necessitate large datasets of human-written essays (HWE) for model training and evaluation. Additionally, underrepresented essays from certain L1s in model training is a persistent challenge, leading to validity issues. This issue is particularly pronounced for many Asian L1s in localized tests, where administrators often lack sufficient essays representative of these L1s. Large language models (LLMs), such as OpenAI's gpt4-o, can potentially address this gap by generating synthetic essays to enhance fairness in AES training. Therefore, this study aimed (1) to evaluate gpt-4o's ability to generate essays mirroring the textual features of HWEs and (2) to explore the use of LLM-generated essays (LGE) in AES training. From the TOEFL11 corpus, a subcorpus of essays by Chinese, Japanese, Korean, and Telugu speakers was created. Using gpt-4o, thirty LGEs per score level per L1 were generated with a five-shot chain-of-thought prompting approach. LGEs were compared with HWEs across 115 textual features extracted with GAMET and Coh-Metrix. Various AES models were trained under three conditions: HWEs only, LGEs only, and a combination of both. All models were evaluated on the same set of HWEs. Results showed that while LGEs captured the directional relationship between textual features and score levels seen in HWEs, they had less variability. AES models trained on both HWEs and LGEs achieved the highest agreement with human scores, improving the agreement scores substantially in some instances. Findings have important implications for enhancing fairness in AES systems, particularly for underrepresented L1s.

Are We Still Measuring the Intended Construct Through Computerized Dynamic Assessment? Insights from Confirmatory Factor Analysis

Meng-Hsun Lee
University of Toronto

This study investigated the impact of binary and dynamic scoring approaches on the dimensionality and construct validity of computerized dynamic assessment (CDA) for reading and listening comprehension when hints are given in written form. The study analyzed data from 472 post-secondary students who completed 16 listening and 21 reading multiple-choice CDA items, with most participants being English language learners. Each item provided scaffolding through a general hint after the first unsuccessful attempt and a specific hint after the second. Two scoring methods were used: binary scoring (0 or 1 point based on the first attempt) and dynamic scoring (0–3 points depending on the assistance needed). Confirmatory factor analysis (CFA) was conducted to examine factor structures, with five CFA models tested under each scoring method.

The CFA model comparison indicated that the bifactor model was the best-fitting across both scoring approaches (RMSEA < 0.02, SRMR < 0.06, CFI and TLI > 0.98). While a general comprehension factor explained the shared variance, several listening and reading items had low factor loadings (< .30) on their specific factors, particularly with dynamic scoring. This suggests these items did not uniquely measure listening or reading skills after accounting for the general factor. In CDA, the dynamic scaffolding potentially introduced some construct irrelevant factors, which may explain the misfit for some reading and listening items. These results raised concerns about the validity of measuring reading and listening comprehension

via the multiple-choice format with CDA, highlighting the need to refine CDA tasks to better distinguish between language skills.

Research Papers

Time: **Friday, 06/June/2025: 1:30pm - 3:00pm** *Location:* **Phramingkwan**

Investigating Factors Influencing HKDSE English Writing Scores: A Five-Facet Rasch Analysis

Meng-Hsun Lee[1], Kuan-Yu Jin[2]
[1]University of Toronto; [2]Hong Kong Examinations and Assessment Authority

The study addressed research gaps by identifying five potential factors that could influence HKDSE writing scores: students' writing ability, rater severity, scoring criteria, prompt difficulty, and scoring environments. The research question guiding this study was: To what extent do the five factors influence HKDSE English writing scores? Utilizing the 2022 HKDSE writing dataset, comprising scores from 456 raters for 44,931 students, the study employed the partial-credit many-facet Rasch model (MFRM) to analyze the impact of the five factors. All raters used an eight-category analytic scoring rubric with three dimensions: content, language and style, and organization.

The MFRM analysis revealed a wide range in students' writing abilities (–11.56 to 10.63 logits), with rater severity ranging between –1.64 and 1.06 logits. Approximately 5% of the raters exhibited misfit, necessitating re-calibration. Among the scoring criteria, content (–0.46 logits) was easier for students, while language and style (0.22 logits) and organization (0.24 logits) were more challenging. Although prompt difficulty varied (–0.32 to 0.46 logits), the impact of different writing prompts on writing scores was minimal relative to the wider range of students' writing abilities (22.19 logits). Similarly, the difference between OSM and iOSM was negligible, with OSM being slightly easier (–0.01 logits) than iOSM (0.01 logits). These findings suggests that while rater and prompt variabilities exist, the HKDSE scoring system is generally robust. Additionally, integrating iOSM offers logistical benefits without compromising the integrity of the writing assessment, as costs associated with physical assessment centers could be reduced.

The Effects of Automated Writing Evaluation (AWE) on EFL Students' IELTS Writing

Napat Jitpaisarnwattana[1], Nick Saville[2]
[1]Silpakorn University, Thailand; [2]Cambridge University Press and Assessment, UK

This study aims to investigate the effects of Write and Improve as an AWE tool to provide writing feedback on students' IELTS writing. The study also seeks to understand students' attitudes towards Write and Improve (W&I) and whether participating in the program can facilitate autonomous learning. Learning behaviours that may contribute to writing improvement is also examined. The data were gathered from a group of university students (N=53) taking a course called English Preparation for Standardised Tests at a Thai university. Data on the effects of W&I on the students' writing performance were collected via a comparison of pre- and post-tests' score. The students' attitudes and autonomous learning practices were collected through questionnaire and semi-structured interviews. Data on

learning behaviours were collected through learning logs generated by the program. The findings reveal that W&I intervention had a positive impact on the students' performance in a post-test in both writing tasks, with the largest gains being evident in the grammatical range and accuracy criterion. The writing improvement was most prominent in the intermediate group (Band scores 5.5). Moreover, the students generally held positive attitudes towards the intervention and thought that taking part could foster their learner autonomy. Finally, students who completed more writing tasks were more likely to perform better in the post-tests. This study concludes that AWE have the potential to help students improve certain aspects of their writing; however, AWE implementation should be designed with clear pedagogical goals and teachers should still be involved in the feedback loop.

Research Papers

Time: **Friday, 06/June/2025: 1:30pm - 3:00pm** *Location:* **Room 401**

From Text to Speech: Exploring Content-Related Features in an Integrated Listening-Into-Speaking EAP Task

Nahal Khabbazbashi, Fumiyo Nakatsuhara, Chihiro Inoue, Johnathan Jones
CRELLA, University of Bedfordshire, United Kingdom

This research examines the effectiveness of integrated tasks in English for Academic Purposes (EAP), focusing on learners' ability to synthesise information from (listening) source texts in their spoken performances. Despite the promise of integrated tasks to enhance authenticity and align with the EAP construct(s), little research has addressed their success in fostering higher-order skills involved in synthesising and summarising information. The content-related characteristics of source text(s) input and the ways in which they might be integrated in candidate output also remains largely unexplored. For our study, we analysed transcripts from 60 spoken performances across varying proficiency levels and first-language backgrounds in response to a prototype listening-into-speaking task. Data analysis involved segmenting source texts and spoken responses into idea units (IUs) and evaluating aspects like reproduction type and accuracy by matching candidate IUs against source text IUs. The findings revealed that the quantity of content in source materials directly influenced candidate output, with higher proficiency correlating with a greater number of IUs produced. While all proficiency levels showed similar awareness of key ideas, higher proficiency candidates often utilised more concrete examples from the source texts. Notably, content accuracy did not consistently increase with proficiency, and macro-proposition reproduction—where candidates extend or summarize input—was the least common type across groups. We will discuss the implications of our study for the development and validation of integrated EAP assessments. We will also consider how learners' performance may be influenced by culturally-specific ways or organising and presenting information.

A Mixed-Methods Investigation of Test-Takers' Performances on Integrated Speaking Tasks: Considering Task Design, Modality of Input and Proficiency Levels

Sathena Chan[1], Lyn May[2]
[1]University of Bedfordshire, United Kingdom; [2]Queensland University of Technology, Australia

While integrated tasks are increasingly used in both high and low-stakes language assessments, the construct of integrated speaking tasks is relatively under researched. In response, the researchers employed a mixed-methods approach utilising panel discussion, text quality analyses and statistical analyses to explore how task design and input modality might impact test takers' performance at different proficiency levels. The dataset consisted of stimulus materials for Reading/Viewing-Speaking (R/V-S) and Listening-Speaking (L-S) tasks, and spoken responses to the two task types from 150 test takers at three proficiency levels. An expert panel was convened to identify the features of spoken summary elicited through the R/V-S and L-S tasks, followed by a comprehensive analysis of performance features from 150 test takers on each task. The analysis of panel discussion resulted in the identification of eight integrated speaking features and a detailed profile of integrated speaking performances at low, mid, and high levels. While the results from two sets of one-way ANOVA showed significant differences in most features across the three proficiency levels, certain features were less effective in discriminating proficiency level. Results also suggest that test takers responded to the two integrated tasks with noticeable differences, indicating the different impact of task design and modality of input on test takers' performance. The mixed-methods approach provided valuable insights into the intricate dynamics of multimodal tasks, revealing the interplay of various task features and the importance of specifying their impact on test takers across proficiency levels during the test development cycle.

An Actor-Network Approach to Diversity in the Interpretation and Use of Concordances for Tests of English for Academic Purposes

Anthony Green[1], Leda Lampropoulou[2]
[1]University of Bedfordshire; [2]LanguageCert / Peoplecert

This presentation explores the complex network of actors and factors involved in the creation, dissemination, and use of concordance studies for large-scale English for Academic Purposes (EAP) exams. Concordances are crucial in helping institutions understand the comparability of language proficiency tests, but their interpretation varies widely across contexts.

Using Actor Network Theory (ANT) as a framework, we conducted a global survey of over 100 institutions delivering programs in English on five continents. Use of ANT supported the examination of the interrelated roles played by human and non-human actors, including institutions, individuals, technologies, and concordance studies themselves, in shaping how concordances are understood and used. Respondents included key stakeholders involved in interpreting concordances and applying results to admissions and policy decisions.

We examined three critical dimensions: Agency, exploring who uses concordance studies and in what contexts; Mediation, focusing on how these studies are transmitted and adapted within institutions; and Awareness, assessing how well stakeholders understand the purpose and implications of concordance studies. Our mixed-methods approach combined quantitative analysis of survey responses with qualitative coding of open-ended questions.

Findings highlighted significant variation in the use and interpretation of concordance studies, shaped by regional and institutional differences. Based on these insights, we offer recommendations for improving concordance literacy and suggest strategies for more effective dissemination of concordance results to meet the needs of institutions in diverse global contexts. This research provides a roadmap for enhancing the practical utility of concordance studies across the educational landscape.

Research Papers

Time: **Friday, 06/June/2025: 1:30pm - 3:00pm** *Location:* **Room 405**

Language Assessment Literacy: Development of Pre-Service English Teachers' Perceptions and Practices Before and After Doing the Practicum

Punchalee Wasanasomsithi[1], Benjawan Plengkham[2]
[1]Chulalongkorn University Language Institute; [2]Nakhon Pathom Rajabhat University

Language assessment literacy refers to the competence, knowledge, and understanding of teachers to design, validate, and make use of different assessment methods to measure and determine language learners' achievements and proficiency. To be able to do so, sufficient background training in language assessment is required. Without language assessment literacy, the quality of teaching and testing can be adversely affected.

This mixed-method study examined the level of language assessment literacy of 21 pre-service English teachers at a public university in Thailand with an aim to explore their perceptions of language assessment literacy and their assessment practices before and after completing a one-semester teaching practicum to determine if any changes had taken place after these pre-service teachers gained actual experiences of teaching and testing during their practicum. A pretest and posttest on test specification design, a perceptions and practices of language assessment literacy questionnaire, and semi-structured interviews were employed to examine levels of language assessment literacy before and after the practicum and to observe changes in their perceptions after they had gained actual experiences teaching and testing students during the practicum. Findings showed improvement in pre-service teachers' assessment literacy post-practicum, particularly in knowledge of assessment and competence in constructing tests. However, interviews revealed concerns about the practicality when translating language assessment theories into practices in the real-world contexts. Recommendations for teacher education programs to integrate practical, context-driven assessment into pre-service teaching training as well as for further research to shed more light on this critical issue for pre-service teacher training are presented.

Understanding Chinese Lexical Inferencing: Assessment and Impacts of Word-internal and Word-external Abilities

Yuxin Peng[1], Stanley Haomin Zhang[2], Cecilia Guanfang Zhao[1]
[1]University of Macau, Macau S.A.R. (China); [2]City University of Macau, Macau S.A.R. (China)

The current study challenges the existing practices in Chinese reading ability assessment that often use Chinese lexical inferencing skills as a proxy (e.g., Wesche & Paribakht, 2010; Zhang, 2016; Zhang & Koda, 2018a). As Chinese has a morphosyllabic writing system, the connections between graphemes and morphemes are more pronounced than those between syllables and graphemes (Koda, 2007). This characteristic encourages learners to depend more on morphological analysis when making lexical inferences (Ke & Koda, 2017; 2019). Given such observations, the study first developed and validated an instrument that measures Chinese lexical inferencing ability (CLIA), and then tested whether it is more a measure of word-internal ability or that of word-external comprehension ability.

The measure was validated, in the first place, among 323 L2 Chinese language learners using IRT and correlational analyses that evaluated its dimensionality, reliability, convergent validity, and relationships with reading ability measurements, i.e., decoding and comprehension skills. In the second phase, 195 participants completed additional assessments of word-internal abilities such as morpheme segmentation, morpheme discrimination (Zhang & Koda, 2018a), and word-external abilities such as oral and print vocabulary knowledge (Zhang, 2016; Koda, 2002) and reading comprehension (Zhang & Koda, 2018b). Multi-group SEM analyses conducted on the latent trait levels of CLIA revealed a clear developmental pattern in terms of the impact of word-internal and word-external abilities. Implications for the conceptualization and measurement of Chinese lexical inferencing skills are then discussed.

Proctoring Language Assessments in Multicultural Contexts

Alina A von Davier, Will Belzak, Rose Hastings, Basim Baig
Duolingo, United States of America

In recent years, the landscape of language testing has undergone a significant transformation, shifting towards digital-first language assessments that are delivered online. This shift has made remote proctoring an indispensable tool in maintaining the validity, integrity, and fairness of examinations. This paper will discuss the role of proctoring for language test validity, highlight recent innovations for improving score integrity, address prevalent challenges of remote proctoring, and identify best practices to optimize the remote examination experience in multicultural contexts. Accurately and fairly certifying the right sessions for a high-stakes test directly impacts the validity of the test: we want to make sure that honest test takers receive the score they deserve and dishonest test takers do not. Deciding fast whether a behavior is acceptable or not is a hard thing to do, especially in multicultural contexts, and needs additional research. In this study we build on the fairness and justice principles from Kunnan (2018) and Isbell, et al (2023).

Research Papers

Time: **Friday, 06/June/2025: 3:30pm - 5:00pm** *Location:* **Ampai**

Comparing Different Standard Setting Methods for Aligning Local English Reading and Listening Comprehension Test Scores with the CEFR

Supong Tangkiengsirisin[1], Sun-Young Shin[2], Suchada Sanonguthai[1]
[1]Thammasat University; [2]Indiana University, United States of America

This study investigates the comparability of two widely used standard-setting methods—the modified Angoff and Bookmark methods—in aligning the Thammasat University General English Test (TUGET) reading and listening scores with the Common European Framework of Reference (CEFR) levels. While previous research has explored these methods in general education and non-CEFR contexts, limited information exists on how different standard-setting approaches affect standard errors of cut scores and panelist agreement across multiple rounds when determining CEFR-aligned cut scores for distinct language skills. A total of 12 panelists with extensive EFL teaching experience and familiarity with both the CEFR and TUGET participated in the study. After a two-day online workshop for familiarization and specification, panelists were split into two groups, each using one of the standard-setting methods. Three rounds of judgment were conducted following Tannenbaum & Cho's (2014) guidelines. Results showed that panelist agreement improved across rounds, with greater consensus observed for the listening section than for the reading section. Furthermore, the Bookmark method yielded significantly higher agreement across multiple rounds compared to the modified Angoff method, aligning with findings from previous studies (Hsieh, 2013; Peterson et al., 2011). This study also highlights the challenges panelists encountered during the standard-setting process and offers practical insights for selecting suitable standard-setting methods in local English proficiency tests, which often operate with fewer resources than international assessments.

Can Expert Reviewers Accurately Identify and Explain the Reasons For DIF in Language Assessments?

Jacqueline Church, Will Belzak, Yigal Attali, Yena Park
Duolingo, United States of America

In language testing, differential item functioning (DIF) analysis is widely used to identify items that may unfairly advantage or disadvantage certain groups based on characteristics like gender, age, or country of origin. While DIF statistics provide valuable quantitative insights, their interpretation can be challenging for practitioners, especially when making decisions about whether items should be retained or removed. Although quantitative DIF analysis is an essential tool for ensuring fairness in language tests, there has been limited research on whether expert reviewers can identify and explain DIF in a way that aligns with statistical findings (Zumbo, 1999; Martinkova et al, 2017). This study investigated the ability of expert reviewers to detect DIF in test items and provide meaningful explanations for their judgments.

We analyzed response data from a large-scale English proficiency test collected between September 2023 and September 2024, focusing on writing and dictation tasks. Expert

reviewers were tasked with evaluating test items based on fairness and potential bias related to demographic characteristics such as gender, age, and country of origin.

This study aims to answer whether expert reviewers can accurately identify DIF, whether their judgments align with the statistical results, and whether the reviewers' own demographic characteristics influence their ability to explain DIF. The findings will inform recommendations for improving sensitivity review processes and integrating qualitative insights with quantitative DIF analyses to enhance fairness in language assessment.

Understanding the Language Assessment Literacy Needs of Junior High EFL Teachers in China: A Validation Study

Zhengqing Luo, Dunlai Lin
Beijing Normal University, People's Republic of China

This is a sequential mixed-methods study to investigate the LAL needs of EFL teachers in Chinese junior high schools, grounded in the LAL profile model. Initially, a 71-item Chinese questionnaire was administered to a broad sample of teachers. This questionnaire was adapted from Kremmel and Harding’s (2020) LAL needs survey, which was developed based on the LAL profile model. A subset of respondents was then invited to participate in semi-structured interviews. This paper, however, focuses solely on the quantitative findings from the questionnaire as part of the ongoing validation study.

The dataset originally comprised 8,125 responses, from which 3,006 were removed for failing an attention check. After three further rounds of data cleaning, 4,077 valid responses were retained. A confirmatory factor analysis was conducted using maximum likelihood estimation in R. Overall, the model demonstrates a borderline acceptable fit, though further refinements may be necessary to improve the CFI and TLI values. Additionally, the Mann-Whitney U Test, Kruskal-Wallis H Test, and ANOVA were conducted to examine differences in LAL needs across various groups, including gender, career phase, school type, school location, and prior knowledge in language assessment. The results indicated no statistically significant differences in LAL needs for groups defined by certain sociocultural factors, such as career phase. However, other factors, including school location and prior knowledge in language assessment, were found to significantly influence LAL needs. In conclusion, this study substantiates the LAL dimensions proposed by Taylor (2013) and validates the adapted LAL needs survey instrument for the Chinese context.

Research Papers

Time: **Friday, 06/June/2025: 3:30pm - 5:00pm** *Location:* **Phramingkwan**

Uncovering K12 Students' Engagement Strategies with ChatGPT's Feedback in Reading Assessments: A Lag Sequential Analysis

Ziqi Chen[1], Wei Wei[1], Jiawei Zhang[2]
[1]Macao Polytechnic University, Macau S.A.R. (China); [2]City University of Macau, Macau S.A.R. (China)

Exam-oriented assessments in East Asia, rooted in Confucian heritage, unconsciously lead students to focus on exam techniques and grades rather than holistic development (Tan, 2018). The recent emergence of Generative AI (GenAI) chatbots offers new opportunities for providing real-time formative feedback, delivering instant, human-like responses that help students evaluate their assessments and improve learning outcomes (Escalante et al., 2023; Ghafouri et al., 2024; Tseng & Lin, 2024). However, the impact of GenAI feedback on student engagement in the collectivist, teacher-centered educational context of East Asia has not been thoroughly explored.

This study investigates the engagement strategies used by Chinese secondary school students when interacting with GenAI-generated feedback during reading assessments. Through a mixed-methods approach, combining screen-recorded data analysis with qualitative interviews, the study involved 39 students who engaged with five types of feedback generated by ChatGPT (GPT-4), including revised answers, marking rubrics, learning strategies, additional tasks, and exemplar answers of varying quality.

Findings from the lag sequential analysis reveal that most students adopted a future-oriented feedback trajectory, starting with self-reflection and progressing to strategic learning techniques, marking rubrics, and exemplars. Instead of passively receiving feedback, these learners actively engaged with GenAI feedback and applied it to further tasks. Qualitative interviews indicated that students viewed GenAI feedback as targeted and relevant, preferring it for productive purposes, such as extracting learning strategies to enhance their approach to similar tasks.

English Exit Exam in Thai Higher Education: Test Characteristics and Teachers' Reflections on Their Practice

Anchana Rukthong[1], Punjaporn Pojanapunya[2], Somruedee Khongput[1]
[1]Prince of Songkla University, Thailand; [2]King Mongkut's University of Technology Thonburi, Thailand

An English Exit Exam (EEE) policy, which enforces the CEFR as guidelines for English language education and requires students to take an English exam before their graduation, has been implemented by higher educational intuitions across Thailand. This is to conform to the Office of Higher Education Commission's policy, introduced in April 2016 (Office of Higher Education Commission, 2016). For a better understanding of how the policy has been implemented in practice, this study analyzed a total of 12 sets of test materials from 12 institutions to investigate the characteristics of the exit exams used, 44 questionnaire responses from university lecturers and administrators involved, and 17 online interviews with the teachers. The analysis of the EEEs showed that different skills, sub-skills and linguistic

knowledge were targeted by the exam and various components of language skills such as reading, listening, writing, speaking, vocabulary and grammar, were included. A majority of the questionnaire respondents (84.6 %) supported the policy, considering it as a tool for student development. Opponents, however, question the effectiveness of the EEE. They argue that the test score alone does not guarantee high English proficiency and advocate for practices that encourage continuous use of English. Almost half of interview participants doubt that the exit exam results are aligned with the CEFR, which describes language proficiency in the form of "can-do statement" while the exam items mainly focus on grammatical and lexical knowledge.

Implementing Learning Oriented Assessment at an Institutional Level: Key Considerations and Challenges for Its Validity

Angeliki Salamoura
Cambridge University Press & Assessment, United Kingdom

Over the past two decades, Learning Oriented Assessment (LOA) has emerged as an influential approach in the area of classroom-based assessment and beyond (e.g. Gebril, 2021; Leung et al., 2018; Purpura and Turner, in press). Although the advantages of LOA within the classroom have been extensively discussed, less attention has been devoted to its application in a wider context: at the school, university or national education system. Yet, recent models of learning programmes highlight the importance of viewing assessment as part of the broader learning programme they sit in (O'Sullivan, 2020). This perspective is essential for understanding the conditions that may lead to their successful or unsuccessful application.

This theoretical paper presents a critical review of research on LOA conducted at institutional and national education levels (e.g., Gebril, 2021; Ho, 2015; Khan & Hassan, 2021; Leung, 2020; Li, 1998; Qi, 2005), aiming to explore the factors that affect its validity. Four key considerations and challenges for ensuring LOA validity at the institutional level were identified: establishing educational coherence (cf. O'Sullivan, 2020; Wang et al., 2024), fostering an understanding of LOA principles among all key stakeholders (not just teachers), building collective teacher efficacy (Hattie, 2017), and closing skill gaps for teachers.

The LOA studies reviewed draw from diverse cultural contexts and geographies, ranging from Australia and Hong Kong to Vietnam, Malaysia and Egypt, and allow for comparisons of practices across eastern and western settings. I will conclude by identifying areas in need of further research to complement the existing LOA literature.

Research Papers

Time: **Friday, 06/June/2025: 3:30pm - 5:00pm** *Location:* **Room 401**

A Longitudinal Investigation of the Relationship of Language and Academic Success of International Students

Donald Bruce Russell
University of Toronto, Canada

Valid test score interpretation and use is crucial to ensure that non-English background students have enough language proficiency to handle language demands in undergraduate study. The relationship of test scores and academic performance is complex and this study responds to calls for studies in multiple contexts that include other pathways to admission (Pakhiti, 2008).

The study applied multilevel modeling (MLM) to longitudinal data consisting of 804 international visa students at a Canadian university who completed a 24-week language program. IELTS scores were used as predictor variables and a MLM fixed and random effects mixed model was used to examine annual change in GPA over time and the effect of IELTS scores on this change. Regression models were used to explore the relationship of test scores and language program outcomes on time to graduate and graduation completion.

MLM results indicate that IELTS reading and writing scores are significantly associated with year 1 GPA which aligns with previous studies (Barkoui, 2024). Language program outcomes explain more variance in year 1 GPA than IELTS scores which aligns with Johnson and Tweedie's (2021) study. IELTS reading scores are significantly associated with GPA longitudinally which confirms the importance of reading scores and GPA (Dooey & Oliver, 2002) which has implications for cut score decisions and support. Other results indicate that language program outcomes are significantly associated with sessions to graduate and graduation completion. These results support validity claims of language programs (Thorpe, 2017) and have implications for program design.

Simulated Real-Life Tasks as a Tool to Investigate the Extrapolation Inference of Language Assessments for Professional Purposes

Fatima Montero
University of Maryland, College Park, United States of America

Following the argument-based approach to test validity (Chapelle et al., 2008), this paper presents the applicability of simulated real-life tasks as external evaluation criterion to investigate the extrapolation inference of language assessments for professional purposes (LAPP). This type of tasks can be used as an alternative to actual samples from the target context of language use when these are difficult to obtain. Additionally, these tasks can be purposefully designed to address the different social and cultural values that influence test validation processes. As such, a simulated task representing a doctor-patient medical follow-up phone call was specifically designed to investigate the extrapolation inference of the speaking section of the Spanish module of the Canopy Credential exam—a LAPP designed for the U.S. medical workplace. The simulated task was designed following the theoretical frameworks of task-based language teaching (TBLT) (Long, 2015), task-based language assessment (TBLA) (Norris, 2009, 2016), and the language proficiency interview (LPI) (Ross, 2017). 60 participants completed the exam and the simulated task in a counterbalanced order, and their performance on each instrument was rated by two different teams of raters. The many-facet Rasch model and linear/multiple regression analyses were employed to examine the appropriateness of the simulated task as external evaluation criterion. Correlation analyses were used to examine the extrapolation inference of the Canopy Credential exam. The results from this study show the applicability of simulated tasks for language test validity research, and their usefulness to examine LAPP designed for multicultural professional domains such as the U.S. medical workplace.

Working Memory among Chinese Learners in Compulsory Education and Its Relationship with English Achievement

Yinjie Tang
Beijing Normal University, People's Republic of China

This study explores the role of working memory (WM) components, phonological WM (PWM), executive WM (EWM), and visuospatial WM (VWM), in English achievement among Chinese EFL learners during compulsory education (Grades 1–9). While WM is vital in Second Language Acquisition, specific components have been underexplored, especially among young learners experiencing rapid WM development. The study involved 81 primary and 52 junior high school English learners. PWM was measured by the Nonword Repetition Task; EWM was measured by the Stop Signal Task, and VWM was measured by the Corsi Block-Tapping Task. English achievement was measured by a scenario-based online test evaluating listening, reading, and speaking skills in real-world contexts. Data analysis included descriptive statistics and correlation analyses. Results indicated that a) primary school learners had stronger EWM and VWM but weaker PWM compared with junior high learners; however, these differences were small and not statistically significant; b) in primary learners, both EWM and PWM showed significant positive correlations with overall English achievement. EWM correlated with reading and speaking skills, while PWM correlated with listening, reading, and speaking skills. VWM did not show significant correlations with English achievement; c) Among junior high learners, only PWM was significantly correlated with overall English achievement. The study highlights the crucial role of EWM and PWM in early English language learning, suggesting that emphasizing phonological awareness and abilities to concentrate in primary education may enhance English acquisition. These insights can inform EFL teaching strategies, indicating that targeting specific WM components can facilitate learners' L2 proficiency.

Research Papers

Time: **Friday, 06/June/2025: 3:30pm - 5:00pm** *Location:* **Room 405**

Examining Product and Process Features of Two TOEFL iBT Opinion Writing Tasks: A Validation Study

Huiying Cai[1], Ching-Ni Hsieh[2]
[1]University of Illinois Urbana-Champaign, United States of America; [2]Educational Testing Service, United States of America

In this study, we examined L2 writers' response characteristics and writing processes associated with the TOEFL iBT Independent Writing (IND) and Writing for an Academic Discussion (WAD) tasks, as well as the relationships between product and process features and essay scores. The participants included 1,095 Chinese-speaking college students. The students responded to the two tasks and filled out an exit survey during regular EFL classes. Students' responses were scored by official TOEFL iBT essay raters using the two tasks' official five-point holistic rating scales. Textual features of the written responses were extracted using natural language processing tools and techniques, focusing on text length, syntax, vocabulary, accuracy, and n-gram overlap with the prompt materials. Keystroke features related to writing strategy, pausing, speed fluency, and editing events were generated using a keystroke logger. The textual and keystroke features were analyzed using linear mixed-effects regression models.

Results showed that the response characteristics and keystroke features were largely comparable across tasks. Some noticeable differences between tasks included syntactic variety, academic vocabulary, unigram/bigram overlaps, initial pause time (before starting to write), and copy/paste events. These variations could be attributed to the differences in task demands. Textual features that exhibited score differences also appeared to distinguish students at different score levels in expected ways and were largely similar between the two tasks (except for text length). Findings of the study provide empirical evidence supporting score interpretations of the TOEFL iBT WAD task and offer practical implications for L2 writing and instruction.

Young Learner Effect in Assessment: An Interactional Multimodal Analysis

Gordon Blaine West, Jason Kemp, Shea Head
WIDA at the University of Wisconsin-Madison, United States of America

In this study, we draw on affect theory, an understanding of emotion as an embodied phenomenon (Ahmed, 2010), to explore multiple affective responses that young learners can have to a language test. As an embodied act, student affect is publicly displayed and can be co-constructed between students and test administrators (TA) (Benesch, 2020; Zembylas, 2007). We conducted a micro-analysis using an interactional multimodal analysis (Kress & Bezemer, 2023) of one young learner's responses to a test and how TAs responded to co-construct a positive affective experience that would allow the student to perform to the best of their ability. Drawing on data collected during cognitive labs as part of a large-scale test development project, we examined video recorded one-on-one administrations of a writing test between TAs and six-year-old students. We show how the students cycled through several affective responses to the assessment (e.g., boredom, annoyance, joy). Each affective response was embodied, and they deployed either resistance or engagement strategies. The TAs, in turn, employed strategies to engage the students. Findings show multiple ways in which young learners may display embodied affective responses to a language test and the strategies they may use to avoid or engage with testing as a result. They also show how TAs may effectively co-construct a positive affective experience with students. We highlight implications for reforms to test administration procedures and TA training for how to help young learners perform to the best of their abilities on language assessments.

Research Papers

Time: **Friday, 06/June/2025: 3:30pm - 5:00pm** *Location:* **Poonsapaya**

Unequal Trajectories? an Examination of the Role of Socio-Economic Status in Shaping Language Proficiency Development in a Mexican Higher Education Context

Nahal Khabbazbashi[1], Parvaneh Tavakoli[2], Edgar Emmanuell Garcia-Ponce[3], Gareth McCray[4]
[1]University of Bedfordshire, United Kingdom; [2]University of Reading, United Kingdom; [3]Universidad de Guanajuato, Mexico; [4]Keele University, United Kingdom

Our longitudinal study explored the factors influencing English language proficiency development in a Mexican higher education setting, with a specific focus on the role of socio-economic status (SES). A longitudinal repeated measures design was employed in which a group of university students took parallel versions of the TOEFL ITP and responded to a survey eliciting information on a range of personal and instructional factors at three time points over two academic years. Participants' English language instruction and learning experience, contact with and use of English outside university, motivation, and demographic information including SES were amongst variables elicited in the surveys. Linear mixed effects modelling was used to analyse the data. To account for the attrition rate and potential bias caused by missing values, we also conducted a sensitivity analysis.

Our results indicated some improvement in the participants' proficiency over time; score gains were statistically significant but small in magnitude. Of the variables analysed, SES, age of onset, and contact with English were the strongest predictors of proficiency – as measured by the TOEFL ITP. The magnitude of the effect was highest for SES, implying that a large amount of variance in proficiency is explained by learner access to social and economic resources. We will discuss the implications of these results focusing in particular on the role of SES, which is an often-neglected factor in the field of language assessment and one that we argue should gain increased visibility in relation to EDI and the future of language assessment.

Investigating Presentation-Mode Effects on L2 Graph-Based Writing: An Eye-Tracking Study

Daniel Yu-Sheng Chang
University of Bristol, United Kingdom

This eye-tracking study investigated how presentation modes influence raters' scoring and cognitive processes/load when assessing L2 graph-based writing. 53 raters scored 60 scripts (duplicate word-processed and handwritten) of graph-based writing, using an analytic rating scale, while their eye movements were recorded. Afterwards, raters completed the cognitive load questionnaire and the interview. Quantitative analysis included Many-Facet Rasch Measurement (MFRM), Wilcoxon tests and mixed-effects models, while qualitative analysis adopted the grounded theory approach.

The MFRM results showed that 42 raters (80%) demonstrated satisfactory rater fit, while rater severity ranged from 1.04 to -2.05 logits. Among the criteria, grammar (0.73 logits) was the most challenging, followed by content (0.17 logits) and vocabulary (0.16 logits), while organisation (-1.05 logits) was the easiest. Bias analyses revealed significant interactions between presentation modes and raters but yielded no significant interactions between presentation modes and criteria. Further, 14 out of 240 score sets (5.8%) differed significantly between handwritten and word-processed scripts, though legibility did not statistically explain the score variance in handwritten scripts.

For cognitive processes and load, while raters invested significantly higher cognitive load in assessing handwritten scripts than word-processed counterparts, their eye-movements seemed to associate with their rating quality across presentation modes. Raters tended to assign scores starting from the most salient script features (e.g., noticeable errors and initial impression) relevant to the criteria, rather than following left-to-right patterns on the rating scale.

This study contributes to understanding of presentation-mode effect on L2 writing assessments and offers insights for rater training and rating scale development.

Impact and Fairness of Retaking Single IELTS Test Components: Global Perspectives and Local Impacts in Asia

Hye-won Lee[1], Emma Bruce[2], Jan Langeslag[2], Tony Clark[1], Reza Tasviri[3]
[1]Cambridge University Press & Assessment, United Kingdom; [2]British Council, United Kingdom; [3]IDP Australia

Scores from English proficiency tests can significantly impact test-takers' lives, particularly in migration and higher education domains. Some candidates retake these tests multiple times to achieve desired objectives, often to meet specific sub-score requirements, and face frustration and anger due to repeated, unsystematic failures. Despite the importance of this issue, research on repeat test-takers is limited, with notable contributions only emerging recently. This gap underscores the need to examine the implications of repeat test-taking on the validity arguments associated with these tests.

To address potential risks associated with repeat test-taking, offering candidates the option to retake one component of their choice within a short timeframe has been proposed. This 'bias for best' approach aims to reduce construct-irrelevant variance from the initial attempt, such as detrimental effects of anxiety or unfamiliarity with the test, without compromising overall validity. The IELTS One Skill Retake (OSR) provides a valuable opportunity to investigate this further, with sufficient global data now available to understand scoring trends and test-taker perspectives.

This study employs a mixed-methods design, integrating scoring data from over 60,000 test-takers with insights from a large-scale questionnaire (n=578) and interviews (n=29). The analysis explores test-takers' characteristics, preparation methods, and overall experience, alongside scoring data, to identify factors influencing observed patterns. Additionally, data from test-takers in Asia is analysed separately to compare regional trends with global IELTS trends. The findings suggest the approach is viable for promoting test fairness, with slight regional variations noted. The discussion will explore broader implications for best practices in the field.

Research Papers

Time: **Friday, 06/June/2025: 3:30pm - 5:00pm** *Location:* **Duangduen**

Alternative Approaches to Reading Item Difficulty Calibration: Perspectives from Text Complexity

Ying Zheng[1], David Booth[2]
[1]University of Southampton, United Kingdom; [2]Pearson

This study investigates alternative methods for calibrating the difficulty of reading items in high-stakes English tests, with a particular focus on text complexity. Traditional calibration methods—such as field testing, live seeding and psychometric analysis—are resource-intensive and time-consuming. Our research explores the potential of technology-driven tools to streamline this process, with an emphasis on text complexity measures as predictors of reading item difficulty.

Three text complexity tools were evaluated: the Automatic Readability Tool for English (ARTE), Readable, and a proprietary toolkit. These tools assess a range of dimensions, including vocabulary, syntax, sentence structure, and discourse cohesion. A pilot study using 15 retired reading items informed the methodology for the main study, which analysed over 100 reading tasks.

Hierarchical linear regression was attempted to model the relationships between text complexity metrics and item difficulty, effectively capturing the nested structure of reading tasks (e.g., words within sentences, sentences within texts). Ridge Regression addressed multicollinearity among predictors, ensuring stable and accurate predictions. Results indicate that a combination of indices from Readable effectively captures key complexity dimensions and correlates strongly with item difficulty. Item type differences were identified. By integrating computational tools with traditional psychometric practices, this approach enhances the calibration process, providing a more nuanced understanding of reading difficulty.

This research provides a foundation for integrating technology-driven approaches into English language testing, particularly for low-stakes test calibration, with potential applications for high-stakes testing.

Integrating Learning with Assessment: Evaluating the SBA Approach with RCTs

Thi Ngoc Quynh Nguyen, Antony Kunnan, Do Thu Hoa
VNU University of Languages and International Studies

The Scenario-Based Approach (SBA; Purpura, 2024; Banerjee, 2025), has led the way in promoting the well-known concept of Learning-Oriented Assessment (LOA) Approach following Purpura's (2021) performance moderators (instructional, cognitive, affective, social-interactional, and technological dimensions). The primary goal of this presentation is to report on two SBA projects in Vietnam to help build test takers' learning capability through assessments.

A second goal of the study was to operationalize the SBA in specific EFL teacher education contexts. Although both projects are pre-service teacher education tertiary programs in Vietnam, the contexts are different: the first develops pre-service ESL teachers and the second is for future EMI teachers. In addition, a third goal of the project was to demonstrate a new experimental research design called randomized controlled trials (RCT) to examine the efficacy of the SBA.

In terms of the specific methods used, about 120 participants were recruited from intact classes that were randomly assigned to either the SBA intervention group or the non-SBA group in two Vietnamese contexts. A pre-experiment language assessment was conducted to establish the score equivalence of the two groups in terms of language proficiency. During the semester-long experiment, participants received the SBA treatment received once a week or twice a week. The results of the experiment showed that the SBA approach in both groups had much more effect on language learning of the participants in the study.

I Can Still Manipulate a Human Interlocutor: Test Taker Perceptions of Taking an Interactive Speaking Test with an Avatar-Enabled Spoken Dialogue System

Reza Neiriz
MetaMetrics Inc., United States of America

Spoken dialogue systems (SDSs) are gaining traction in oral communication (OC) testing due to their consistency and cost-effectiveness, providing an alternative to human interlocutors (Gokturk Tuney, 2020; Karatay, 2022; Ockey et al., 2023). However, their ability to replicate human-to-human interaction—especially for constructs like Interactional Competence (IC)—requires further exploration. This study investigates the effectiveness of Shahryar, a state-of-the-art SDS developed by the researcher, in assessing IC. Shahryar features an interactive avatar and an "online listening" capability, allowing it to listen continuously and manage interruptions, mirroring real-time human interaction. Ninety-six ESL/EFL learners engaged in 10 two-and-a-half-minute discussions with Shahryar, randomly assigned to either an avatar or still image condition. Participants completed surveys and interviews, which were analyzed using a phenomenological approach. Key themes include comparisons between human and SDS test experiences, the role of visual representations (avatar vs. still image), and areas for improvement in SDS design. This presentation will discuss the study's findings and their implications for enhancing the assessment of IC through SDSs, shedding light on their potential to deliver authentic and scalable language testing solutions.

Works-in-Progress – Friday, June 6, 2026

Time: **Friday, 06/June/2025: 5:00pm - 6:30pm** *Location:* **Sumon**

AI-Powered Listening Tests—How Reliable are They?

Junyan Guo
Wuxi University, China, People's Republic of China

Given the increasing interest in AI-driven item generation in language testing, this study explores the capability of Generative AI in developing the listening section of the College English Test—Band 4 (CET-4), a national test aiming at examining the English proficiency of undergraduates in China. The listening section of the CET-4 test has three parts: 3 news reports (7 questions), 2 conversations (8 questions) and 3 passages (10 questions). This study is composed of three stages. In the first stage, the research will analyze the corpus of the listening section of 35 CET-4 tests (2016-2024), including the genre, topic area, lexical richness, and targeting listening subskills in the questions. In the second stage, according to the results of stage one, prompt engineering and fine-tuning of prompts will be employed in ChatGPT to create authentic-test-paralleled listening scripts and test items of one news report, one conversation and one passage. Additionally, MURF.AI will be adopted to create audio input based on the generated script. University students will be invited to complete the authentic CET-4 listening test and the AI-created listening test. In the third stage, both expert judgment and psychometric models will be adopted to examine the quality of the output. Based on the findings, the opportunities and challenges provided by AI-powered tests will be discussed and suggestions will be made for future language test research and development.

Workshopping LAL: Engaging with University Admissions and Recruitment Staff in Language Assessment Literacy Development

Daniel M. K. Lam, Angela Gayton
University of Glasgow, United Kingdom

Language test scores are widely used in university admissions. Yet, decisions on setting/revising minimum scores and choosing which tests to accept are often not based on a sound knowledge of what the test scores (don't) mean but more on student recruitment priorities (O'Loughlin, 2011). This project aims to strengthen the testing community's impact on good practices in using language assessments (Kremmel & Harding, 2020) through actively engaging with test score users' language assessment literacy (LAL) development, and address two under-researched areas in LAL: what constitutes LAL among stakeholders in university admissions contexts, and what might be effective teaching/learning methods to develop it (Fulcher, 2020; Gan & Lam 2022).

To achieve these aims, the current project develops two 2-hour LAL workshops and deliver them to approximately 60 recruitment and admissions staff across UK universities. The workshops explore themes such as (in)appropriate inferences from language test scores, how comparable different tests used for university admissions are, as well as conflicting priorities in admissions decision-making and any creative solutions to reconcile the differing priorities. Hands-on activities to actively engage participants with the workshop themes include comparing test tasks with academic tasks and across different tests, comparing

speaking/writing samples at different score bands, and discussion on conflicting priorities among stakeholders in different roles.

This presentation will focus on the design of the LAL workshops, and report preliminary findings from participants' evaluation of the workshops through pre- and post-workshop questionnaires, reflections and focus groups.

Developing In-House English Proficiency Tests for Thai Universities: A Practical Approach to Aligning With CEFR

Worasuda Wattanawong, Siriphan Suwannalai, Dusadee Rangseechatchawan
Chiang Mai Rajabhat University, Thailand

This presentation explores the development of in-house English proficiency tests at Chiang Mai Rajabhat University (CMRU) and other universities in northern Thailand, with a focus on aligning these tests with the Common European Framework of Reference for Languages (CEFR). In response to a nationwide policy requiring universities to assess students' English proficiency, CMRU opted for the creation of in-house tests due to budgetary constraints. This initiative was part of a broader effort to ensure economic sustainability while maintaining academic standards. The presentation highlights the challenges and strategies encountered in test development, particularly in adapting the tests to reflect local cultural and academic contexts, while adhering to international standards. It also emphasizes the crucial role of professional development workshops, funded by the ILTA/Duolingo Collaboration and Outreach Grants, in training Thai lecturers on test design, item writing, and ethical considerations in language assessment.

Through collaborative efforts, these workshops facilitated the creation of CEFR-aligned, standardized tests tailored to the needs of regional universities. This presentation will discuss the results of these initiatives, showcasing the practical applications of test constructs and the ongoing collaboration between institutions to enhance the quality and relevance of English language testing in Thailand. This presentation offers insights into the sustainability of language assessment and the potential for expanding these models to other universities in Thailand and beyond.

How Test Takers' Attention Distribution on Source Text Affects Their Performance in the Continuation Task: An Eye-Tracking Study

Yang Zhang, Lin Shi, Lianzhen He
Zhejiang University, People's Republic of China

While previous research has indicated that effective use of source text is a key predictor of test takers' performance in integrated writing tasks, little is known about the underlying cognitive processes involved, particularly how varying patterns of attention distribution influence overall performance. To address this gap, the present study uses eye-tracking technology to examine test takers' cognitive processes when they engage with source text during a continuation task.

Prior research on the continuation task has highlighted that successful task completion depends on alignment with the source text. Interactive Alignment Model suggests that alignment is facilitated by a thorough understanding of the situation model, which includes five key dimensions: space, time, causality, intentionality, and references to individuals. In this study, the source text is divided into five types of context clues, based on these dimensions. Eye-tracking technology is employed to investigate test takers' attention distribution across these clues.

Two primary research questions are addressed: (1) What cognitive processes are involved when test takers engage with the source text? and (2) How does attention distribution across different context clues affect their writing performance and alignment? This study plans to recruit 60 Chinese L2 learners as participants. They will complete a continuation task on computer, with their eye movements recorded by Tobii TX300 eye-tracker. Mixed-effects models will be constructed to analyze the impact of attention distribution on performance. Findings of this study are expected to provide insights into cognitive mechanisms in integrated writing tasks and offer practical guidance for task design.

Using a GenAI-Based Conversational Agent to Assess Second Language Learners' Interactional Competence

Zhuohan Hou, Shangchao Min
Zhejiang University, People's Republic of China

This study explores the potential of using a GenAI-based conversational agent (CA) to assess interactional competence (IC) in paired speaking tasks. The research investigates how L2 learners perform when interacting with a CA versus a trained human interlocutor, focusing on linguistic and interactional features and the quantitative relationship among these features to predict speaking proficiency.

This research employs an exploratory sequential design consisting of 2 phases. Phase 1 developed the task, rating scale, a coding scheme of IC features, and a CA with a pilot study involving 30 university students. Phase 2 expands the study to 120 university students, each participating in 4 discussion tasks related to challenges in campus life - two with a CA and two with a trained human interlocutor, delivered in a counterbalanced order. Participants will also complete a final online survey and semi-structured interviews to compare the two task formats. Four trained raters will assess participants' overall speaking performance. Following criterial features defined in previous research, IC features will be coded into 6 domains: topic management, turn-taking management, interactive listening, non-verbal or visual behaviors, manner of interaction, and repair. Multiple analyses will be conducted, including (1) Rasch measurement of overall speaking score to evaluate rater severity, task difficulty, and test-takers' performance, (2) frequency and thematic analysis of survey and interview to explore test-takers' perceptions, (3) paired sample t-tests of overall speaking scores, linguistic and IC features for group differences, (4) conversation analysis of oral performance transcripts for qualitative examination of IC features, and (5) factor and regression analysis to investigate quantitative relationships between linguistic and IC features in predicting overall speaking score.

The Development and Initial Validation of a Standardized Test of English for University Admission Aligned with a National Reformed Curriculum

Thao Thi Phuong Nguyen, Chi Thi Nguyen, Yen Thi Quynh Nguyen, Hoa Quynh Nguyen, Quynh Thi Ngoc Nguyen, Duong Thuy Le
University of Languages and International Studies, Vietnam National University, Vietnam

The national curriculum has made its substantial impacts on educational practices related to teaching, learning and assessment (Hoang, 2022; Cunning, 2009). Starting in 2025, Vietnam National University (VNU) in Hanoi will include an English test as part of the High School Assessment (HSA) test, based on the reformed 2018 General Education English Curriculum (GEEC). This high-stake test with about 800,000 test takers per year is used as one of key methods to select potential students for branch universities in VNU. This study details an ongoing procedure of developing and designing the test. It outlines the process of developing the test from specifying test format, creating test specifications to test piloting. The 2018 GEEC was thoroughly examined to ensure that the required competencies, knowledge, and skills are accurately reflected in the test. The data analysis of the piloting phrase with about 300 high school students showed that the results were satisfactory when most of the items fit the expected level and student's performance based on Rash model analysis. Item banks have been developed to prepare for examinations from March to June 2025. The test will be validated through score analysis of the first two examinations. Content validity will also be examined with the sample tests to show the link between the 2018 GEEC with the test. Findings of the study are expected to answer the question how the test is aligned with the reformed curriculum and how its validity is performed across examinations, providing a reliable basis for future recommendations.

Feeding The Machine: A Comparison Between Analytical and Holistic Scoring to Inform an Automated Essay Scoring System

Joni Kruijsbergen, Fauve De Backer, Orphée De Clercq, Goedele Vandommele
Ghent University, Belgium

Writing is a crucial literacy skill essential for academic success (Wolf et al., 2024). However, assessing this skill remains challenging, especially in large-scale standardised testing. While selected-response formats like multiple-choice questions are easy to score, they fail to capture students' ability to use language in real-world contexts. Performance-based assessments are better suited for evaluating writing skills but require much more resources (Green, 2022). As automated scoring systems (AES) continue to grow in prevalence (Li & Ng, 2024), understanding the relationship between different human scoring methods and AES becomes critical. This study explores writing assessment in large-scale centralised tests administered to second-year secondary students in Flanders by comparing two scoring methods — analytic scoring and pairwise comparative judgment. Two research questions are investigated: (1) To what extent do the two human scoring methods produce similar scores? And (2) To what extent are the characteristics extracted by AES similar for both human scoring methods?

Exploring Raters’ Scoring Processes in Assessment of English-Chinese Consecutive Interpreting: A Qualitative Study Based on Retrospective Verbalization

Mengting Jiang
Xiamen University, People's republic of China

In rater-mediated assessment of interpreting, raters play a pivotal role of assigning numeric scores, based on certain scoring rules. An important line of research in the previous literature has examined psychometric properties of rater-assigned scores in interpreting assessment such as reliability and validity. Equally important is the scoring process in which raters make evaluative judgments and scoring decisions, as such process is closely related to scoring validity. While extensive research has investigated how raters evaluate monolingual writing and speaking, scant scholarly attention has been devoted to raters’ scoring processes in assessment of interlingual interpreting.

Against this background, we conducted a qualitative study to explore raters’ scoring processes in the assessment of bidirectional English-Chinese consecutive interpreting. Retrospective verbal reports were collected from two rater groups, including 29 teacher raters and 30 student raters. Half of the raters in each group were randomly assigned to evaluate either English-to-Chinese or Chinese-to-English consecutive interpretations of varying performance qualities. Preliminary qualitative content analysis helped us build a descriptive framework of raters’ scoring processes. This framework comprises two overarching dimensions (i.e., meta-cognitive and cognitive processes) underpinned by six major categories and 43 sub-categories. Additionally, rater experience and interpreting direction appear to have modulated raters’ scoring processes.

This study represents one of the first attempts to provide an empirically-based and fine-grained framework for elucidating the intricacies of rater cognition in interpreting assessment. Potential implications are discussed in relation to scoring validity, rater training, and development of automated scoring systems modeled on rater cognition.

Effects Of Response Language on Test-Takers’ Performance and Cognitive Processes in L2 Listening Recall Tasks

Phuong Nguyen, Ahmet Dursun
The University of Chicago, United States of America

Immediate recall tasks, requiring test-takers to write down all comprehended idea units, are valued for their ability to provide insights into learner-text interaction and to prevent test-taking strategies. However, a major concern is that writing proficiency may interfere with test-takers' performance as a construct-irrelevant variable. While most studies have focused on reading assessment and highlighted the benefits of responding in the first language (L1) over second language (L2), limited research has examined how response language affects L2 test-takers' performance, particularly in listening recall tasks.

This study explores (1) the effect of response language (L1 vs. L2) on performance in L2 listening recall tasks and (2) listeners’ cognitive processes during these tasks. Participants are 56 English-speaking learners of Spanish at intermediate and advanced levels. They will complete two recall tasks—an announcement and a message—in either English or Spanish. A counterbalanced design will be used to minimize ordering effects. Additionally, a subset of

8 participants will provide verbal recalls of their cognitive processes, which will be audio-recorded.

Responses will be graded by two Spanish raters, with Krippendorff's α calculated for inter-rater reliability. Verbal reports will be transcribed and coded for cognitive processes. Multiple linear regression will be used to analyze the effects of response language and proficiency level on test scores. To address the question about the listeners' cognitive processes involved during these tasks, themes identified from the verbal reports will be reported, illustrated by participants' quotes representative of these themes. Findings could have implications for the use of recall tasks in L2 listening assessments.

Exploring Task Designs Suitable for High-Stakes Spoken Japanese Language Assessment

Fumiyo Nakatsuhara[1], Chihiro Inoue[1], Atsuko Osumi[2], Yumi Horikawa[2]
[1]CRELLA, University of Bedfordshire, United Kingdom; [2]Japanese Language Proficiency Test Research Department, Japan Foundation

According to the Immigration Service of Japan (2023), over 3.4 million foreign residents live in Japan, including skilled professionals, students, and family members, with a significant number of semi-skilled workers in sectors like elderly care to address workforce shortages. Some visa types require applicants to demonstrate Japanese language proficiency through tests like the Japanese Language Proficiency Test, taken by 1.3 million candidates annually. However, these assessments lack an oral assessment, raising concerns about communication skills essential in sectors such as elderly care. There is an urgent need to address the lack of a spoken Japanese test for immigration purposes.

This Work-in-Progress presentation reports on baseline research exploring task designs for high-stakes speaking assessments for immigration purposes. The research comprised two phases: document analysis (Phase 1) and focus group discussions (Phase 2). Phase 1 identified 13 Japanese tests with spoken components. Of these 13 tests with multiple tasks, 11 tasks were selected for detailed analysis using the socio-cognitive framework for speaking assessments (e.g. Taylor, 2011; O'Sullivan et al., 2020) to reverse-engineer the test specifications. Phase 2 involved focus group discussions with 10 expert panellists on three testing purposes: everyday communication (CEFR A2), higher education studies (CEFR B1-C1), and occupational purposes (CEFR B1-C1). These discussions provided insights into language functions, honorific expressions, and cultural knowledge needed for each context. This presentation will invite the audience to share their insights and experiences regarding language assessments for languages other than English, particularly those for which relatively little research has been publicly available.

Prototyping an Information-Based Academic Writing Assessment

Chengyuan Yu
Kent State University, United States of America

To reflect the changing practice in academic writing in this age of information explosion, this study presents the preliminary qualitative prototyping of a multimodal information-based academic writing (IBAW) assessment for postgraduate students in a specific discipline, that is, education. Two research questions guided this study: (1) To what extent does evidence support the interpretation and use of the IBAW assessment? (2) What is the test-takers' IBAW process like? To explore these questions, semi-structured interviews were conducted with graduate students, writing instructors, faculty, and librarians, capturing insights into their experiences with IBAW. Analysis of these interviews identified research proposal writing as a prototypical task, encompassing all information literacy processes and receptive skills (listening and reading). In the IBAW assessment, test-takers watch a video lecture on a new concept in educational research and respond to a related prompt. They explore additional information on the Internet and academic databases, with a pre-writing task assessing their information literacy through multiple-choice and short-answer questions. Subsequently, test-takers compose responses to the writing prompt. The assessment underwent initial review by six graduate students and was taken by four graduate students in education, whose experiences were collected using a retrospective think-aloud protocol. Preliminary analysis indicates that the IBAW assessment effectively elicits cognitive processes aligned with the IBAW model (Yu & Zhao, 2021). Test-takers iteratively implemented IBAW processes, integrating planning and drafting while often omitting revision, possibly due to the timed test format. Feedback is welcomed for the ongoing development of the IBAW assessment.

The Predictive Power of Lexical Richness Indices in Chinese EFL Learners' Performance in L2 Speaking Tasks

Lin Shi, Yang Zhang, Lianzhen He
Zhejiang University, People's Republic of China

Lexical richness, a multidimensional construct comprising lexical sophistication, lexical diversity, and lexical density, is recognized as a key indicator of learners' writing proficiency (Zhang et al., 2021). However, due to the inherent difficulties in obtaining authentic spoken test corpora and the labor-intensive process of analyzing such data, the predictive capacity of lexical richness indices in evaluating L2 learners' performance in different types of speaking tasks has yet to be thoroughly investigated. The present study fills this gap by utilizing the advantages of authentic spoken texts to investigate the predictive power of lexical richness indices in Chinese EFL learners' performance in different types of L2 speaking tasks, hoping to provide evidence of lexical richness for automatic speech evaluation. This study analyzes speaking test data from the Undergraduate English Proficiency Test at a key university in China, where groups of 3-4 students perform both monologue and group discussion tasks. To ensure topic familiarity, 50 non-English majors will rate 15 potential topics, from which the 8 most familiar will be selected for analysis. Sixteen groups (approximately 60 participants) will be chosen for evaluation. Three experienced raters will assess speaking performances based on Chinese College English Test criteria, with a third rater resolving discrepancies. Audio recordings will be transcribed and preprocessed using Python to enhance accuracy. Lexical richness will be evaluated across three dimensions (i.e., lexical sophistication, lexical diversity, and lexical density), using over 400 indices calculated by various tools, with stepwise regression models applied to assess predictive power in relation to task types.

Test Scores and Speech Samples: Extrapolating from A Computer-Delivered Test of Speaking for University Admission to Group Oral Discussions

Yujia Zhou, Masashi Negishi, Asako Yoshitomi
Tokyo University of Foreign studies, Japan

The extrapolation of test scores to a target domain, i.e., association between test performances and relevant real-world outcomes, is critical to valid score interpretation (Chapelle, 2020; Kane, 2013). This is of particular concern in large-scale speaking tests delivered by a computer, which do not require synchronous communication (Roever & Ikeda, 2022, 2024). Despite its importance, the extrapolation inference for speaking assessments is underexplored (Fan & Yan, 2020), and previous studies (Brooks & Swain, 2014; Ockey et al., 2015) mainly focused on international proficiency tests such as TOEFL iBT.

This study, therefore, seeks to investigate the extrapolation inference for a high-stakes computer-delivered test of speaking for Japanese university admission, the British Council Tokyo University of Foreign Studies-Speaking Test for Japanese Universities (BCT-S). The overall research question of the study is "to what extent is performance elicited by the BCT-S reflective of performance on group oral discussion tasks, one of the main Targeted Language Use domains reported for the BCT-S?" Specifically, the study aims to examine (1) the relationship between students' BCT-S scores and their scores on a group oral discussion test, (2) the relationship between BCT-S scores and interactional features observed in speech samples of group discussions, and (3) the comparability of the quality of speech samples produced in the two tests regarding both scores awarded and discourse features observed.

It is hoped that the study will contribute to a better understanding of the test constructs of computer-delivered monologic tasks and provide insights into score interpretation of computer-delivered speaking tests.

Investigating self-identified language assessment literacy needs of teachers in North America, Europe, and Asia

Benjamin Kremmel[1], Luke Harding[2]
[1]University of Innsbruck; [2]Lancaster University

In their global survey of language assessment literacy (LAL), Kremmel and Harding (2020) identified the self-identified targets of language teachers across nine distinct dimensions of LAL. Their online survey, however, only investigated these targets on an aggregate level, without much in-depth exploration for how targets might vary in different regional contexts. Furthermore, while Kremmel and Harding's data is valuable in terms of what teachers would see as their optimal level of LAL, little is known about potential gaps between these targets and teachers' current levels of LAL. Although numerous studies have investigated local LAL training needs of teachers (e.g. Firoozi et al, 2019; Giraldo, 2018; Shahzadi & Ducasse, 2022), there has not yet been a comparative investigation across global regions with the same survey instrument of teacher's LAL needs, their self-assessment, and the gap in between across different geographical contexts.

This paper will thus report on findings from an online survey that used Kremmel and Harding's statements and asked international teachers (N=645 to date) a) what level of LAL they think their peer group need in their context and b) what level of LAL they themselves think they currently have. 160 teachers completed both sets of statements so far. Results are reported along the nine LAL dimensions of Kremmel and Harding's model, as well as at item level for individual statements. The analysis specifically focuses on comparing teacher responses from North America, Europe, and Asia to identify localized LAL needs.

Investigating the Feasibility of ChatGPT for Generating Passages and Items in Different Types of EFL Reading Tasks

Lin Shi, Yuhong Gao, Lianzhen He
Zhejiang University, People's Republic of China

The demand for various types of items in reading assessment necessitates novel item development methods. Given that automatic item generation (AIG) has boundless potential for creating reliable and diverse assessments (Attali et al., 2022) and with the development of generative AI represented by ChatGPT, AIG's potential in language testing has garnered increasing attention. However, there is still limited research comparing ChatGPT's feasibility of generating passages and items for different reading task types. This study is an attempt to address this gap by investigating the feasibility of using ChatGPT-4o to generate reading passages and items for three types of reading tasks (i.e., Cloze, Information Matching, and Multiple Choice) for Chinese College English Test-Band 6 (CET-6). A total of 90 passages are generated by ChatGPT, with 30 passages per task type, referencing 90 retired CET-6 passages to ensure alignment with task characteristics. For passage evaluation, a multi-dimensional framework, based on Bachman and Palmer's (1996) language task characteristics, is designed to evaluate readability, lexical properties, and syntactic features of the generated passages. For item evaluation, a set of quality criteria tailored to each of the three reading task types are developed, aligned with the test-point validity model (Li, 1997) and CET-6 item characteristics. Additionally, five experienced item writers will evaluate the quality of the generated items according to the customized moderation criteria, followed by semi-structured interviews to gather their perceptions. This study applies both quantitative and qualitative analyses, hoping to offer insights into the application of large language models in reading assessment development.

Exploring Assessment Literacy among EMI University Teachers in STEM Fields: A Mixed-Methods Study in Taiwan

I-Chun Vera Hsiao
The University of Iowa, United States of America

This work-in-progress study aims to explore the assessment literacy (AL) level of English as a Medium of Instruction (EMI) university teachers in Taiwan by developing and administering a survey informed by teacher interviews.

EMI has become a prevalent mode for delivering academic content in higher education globally. In Taiwan, the Ministry of Education implemented the Bilingual 2030 policy in 2018, significantly increasing the number of EMI programs and courses. Particularly relevant for EMI instructors, AL refers to the skills and knowledge of educational assessment that are essential to their professional expertise in evaluating student outcomes (Stiggins, 1997). Research on teachers' AL in Taiwan has primarily focused on the humanities and social sciences, leaving a gap in understanding how EMI teachers in non-language disciplines approach assessment practices. Therefore, this study focuses on science, technology, engineering, and mathematics (STEM) teachers, as STEM education requires assessments that evaluate theoretical knowledge, practical applications, problem-solving, and technical skills (Felder & Brent, 2024), necessitating a high level of AL among teachers.

An exploratory sequential mixed methods design was adopted, starting with semi-structured interviews with 10 teachers to gather insights into their assessment practices and challenges. This qualitative data informed the development of a quantitative AL survey, which will be pilot-tested with a small group of 75 EMI STEM teachers to assess clarity and relevance before being administered to a broader sample of 300 teachers (Mills & Gay, 2018). The findings will inform institutions and policymakers on areas for potential improvement.

Understanding Multilingual Communicative Competence: Exploring Possibilities for Language Assessment in Higher Education

Slobodanka Dimova
University of Copenhagen, Denmark

English for Academic Purposes tests have been the standard for assessing language proficiency for university entry and for diagnostic purposes in various English-medium instruction contexts. However, these tests often fail to capture the heteroglossic communicative competences required for disciplinary learning in multilingual higher education environments, even though research, especially in non-English-dominant university contexts (e.g., Sweden, China), has documented their role in classroom communication and examination, mostly in the form of translanguaging. Existing research on translanguaging at tertiary level is mainly descriptive and lacks models that are applicable in test design. Therefore, this project goes beyond description of translanguaging practices and focuses on identifying the characteristics of multilingual communicative competence that are crucial for academic success. More specifically, it explores the communicative competences needed to both 'process' and 'display' disciplinary learning for the purpose of developing an extended competence framework. Such framework will be useful for design of language assessment tools (especially post-entry) that can identify students' needs and provide relevant support. We are currently collecting transnational data across three university settings where both the local language(s) and English dominate in disciplinary instruction with the intention is to create a corpus of relevant tasks.

In this works-in-progress presentation, we present an overview of key findings and preliminary plans for data analysis. Then, we will seek participants' input and perspectives on how to proceed with capturing the role of multilingual competence in the process of task completion and the implications of the project data for multilingual assessment task design.

Grammatical Complexity in Thai EFL English-major Students' Writing

Pong-ampai Kongcharoen[1], Xinyu Zhao[2]
[1]Northern Arizona University, United States of America; [2]Nanjing University, China

Grammatical complexity in writing has seen differently from spoken register. NP complexity, especially, has been in a focus recently for written genre. In this study, grammatical complexity of Thai EFL English-major students' writing will be examined. There are two learner corpora which are the first academic writing course (first learner corpus) and research writing course (second learner corpus), and a reference corpus collected from research reports in British Academic Written English (BAWE) in this study. Biber Tagger will be used to extract the grammatical features of both learner corpora and reference corpus. Key Feature Analysis will be utilized to see the grammatical key feature of the three corpora. The results from this study will yield the direction of grammatical development and grammatical teaching for Thai EFL English-major students.

Rating Quality across Different Presentation Modes in L2 Writing Assessments: A Comparison of Human Raters and AI-Generated Scoring

Daniel Yu-Sheng Chang, Pu Pu
University of Bristol, United Kingdom

Recent research interest has shifted to AI-generated scoring (e.g., ChatGPT). Showing potential but remains context-dependent, it still lacks empirical evidence to complement or fully replace human raters. This mixed-methods study explored rating quality across different presentation modes (handwritten vs. word-processed) in L2 writing from two perspectives: human raters and AI-generated scoring. 60 essays (30 handwritten and 30 word-processed) were scored using an analytic rating scale (content, organisation, lexicon and grammar). In Phase 1, after 53 raters scored all essays, they completed a questionnaire and interviews to discuss their scoring decisions and perceptions of rating across the different presentation modes. In phase 2, the same batch of essays were uploaded to ChatGPT-4o for scoring and providing scoring explanations. Quantitative analysis included Many-facet Rasch measurement and mixed-effects models, while qualitative analysis adopted the grounded theory approach.

Although this project is still a work in progress, preliminary findings indicate: (1) presentation modes may cause some discrepancies in scores from human raters and ChatGPT, but most differences are not statistically significant; (2) presentation modes appear to have significant interactions with raters and ChatGPT, but no significant interactions were observed between presentation modes and the four rating criteria; (3) ChatGPT seem to demonstrate moderate reliability compared to human raters, as it is limited in accurately assessing test-takers' content-related performances, according to its scoring explanations. As a work in progress, we sincerely welcome feedback from the language testing community to optimise our analytical approach and enhance the presentation of findings.

Paper and Demo Summaries – Saturday, June 7, 2025

Research Papers

Time: **Saturday, 07/June/2025: 8:30am - 10:30am** *Location:* **Ampai**

Disentangling Multimodality in Speaking Assessment: The Interplay of Nonverbal Behavior, Affect, and Language in Estimates of Second Language Ability

John Dylan Burton
Georgia State University, United States of America

This study explores how raters use nonverbal behavior when scoring second language (L2) speaking tests. While past research shows that raters notice nonverbal cues like facial expressions, gestures, and paralinguistic features (Ducasse & Brown, 2009; Jenkins & Parra, 2003; May, 2011, Sato & McNamara, 2019), their role in language ability judgements remains underexplored (Jenkins & Parra, 2003; Plough et al., 2018). This study builds on past research to document the salience of nonverbal behavior and how it relates to test outcomes.

In this mixed-methods study, 100 raters evaluated 30 pre-recorded high-stakes speaking test samples conducted via Zoom. They provided scores on fluency, vocabulary, grammar, comprehensibility, and affective scales. Twenty of the raters then participated in stimulated verbal recall sessions to detail their thought processes while scoring. Results indicated that nonverbal behavior constituted 11% of the raters' comments, with particular focus on eye gaze patterns, mouth movements, and paralinguistic features. Raters integrated nonverbal behavior with verbal cues to assess listening comprehension and comprehensibility, primarily through perceived affect, such as confidence and engagement. However, raters rarely made direct connections between specific behaviors and language ability. Instead, behavior served as a heuristic, influencing their overall judgment rather than directly determining scores. The findings highlight the importance of nonverbal communication in L2 assessment, with implications for scale construction, rater training, and models of L2 communication.

The Impact of an AI-interviewer's Relationship-Building Dialogue Strategies on Language Performance

Fuma Kurata[1], Masaki Eguchi[1], Máo Saeki[1,2], Shungo Suzuki[1], Yoichi Matsuyama[1,2]
[1]Waseda University, Japan; [2]Equmenopolis, Inc.

This study investigates whether high-level engagement can be achieved in AI-delivered Oral Proficiency Interviews (OPIs) by altering the AI interviewer's behavior. Human-led OPIs struggle with standardizing interviewer behavior due to individual differences, affecting fairness and reliability. AI agents have been proposed to minimize variability and enhance standardization. As AI-based assessments become more accessible and formative, fostering high engagement and systematically building rapport with test-takers is essential while maintaining measurement quality.

The research compares learners' affective responses and performances using two AI agents: one with relationship-building behaviors and one without. Sixty EFL learners at a Japanese university participated. The relationship-building agent used friendly behaviors from the literature, including small talk, empathetic responses, positive feedback, and attentive nodding. The other agent provided neutral, simple responses without relational cues. Two parallel interview scenarios were counterbalanced to minimize practice effects, varying the order of AI agent presentations.

Participants completed an engagement and rapport questionnaire after each interview, and their language performance was assessed using a highly reliable automated scoring system (kappa above 0.9). Mixed-linear regression analysis showed significantly greater cognitive, emotional, and social engagement, as well as increased rapport with the relationship-building agent. However, there were no significant differences in behavioral engagement or final automated scoring results.

These findings suggest that incorporating relationship-building behaviors into AI interviewers enhances engagement and rapport without significantly impacting immediate language performance, thus achieving high standardization in assessments.

Advances in the Assessment of Interactional Competence: A Systematic Literature Review

Anh Nguyen[1,2], Noriko Iwashita[1]
[1]University of Queensland, Australia; [2]Hanoi University, Vietnam

Since Kramsch’s (1986) seminal work, the concept of interactional competence (IC) has been continuously elaborated in SL pedagogy and assessment over the past 30+ years. IC involves general linguistic knowledge and the ability to apply context-specific communicative strategies, including the roles of participants and how they navigate communication (e.g., Hall & Pekarek Doehler, 2011; Ross, 2018; Young, 2008, 2011). As IC gains prominence in speaking assessments, research has identified its key features, incorporated them into assessment tasks, and embedded them in rating scales across diverse contexts (Galaczi & Taylor, 2018). Despite the growing body of research, there remains a lack of synthesis in this area, and assessing IC in multicultural contexts presents challenges due to linguistic and cultural factors. This paper addresses these gaps by providing a systematic review of studies focused on IC assessment. Using PRISMA (Page et al., 2021), the review targeted 2019–2024 publications, identifying 60 eligible studies. Thematic analysis revealed four themes: key IC features, factors influencing IC, interaction modes, and research methodologies. Turn-taking was the most studied IC feature, focusing more on non-verbal interaction. Most studies used qualitative methods like conversation analysis, though mixed-methods research is rising. IC is also examined in broader contexts, including face-to-face, virtual, and testing scenarios. This study contributes to a comprehensive conceptualisation of IC in speaking assessment, identifying factors and modes influencing IC. It offers insights for developing effective tools and training materials and provides practical suggestions for assessing IC in multicultural environments, outlining future research directions.

Predicting Washback of a Speaking Assessment in the Japanese University Entrance Exam Context

David Allen
Ochanomizu University, Japan

In line with the ethos of 'working for washback' (Swain, 1985), washback research conducted prior to test use may support development of a theory of action (e.g., Chalhoub-Deville & O'Sullivan, 2020; Saville & Khalifa, 2016) that can facilitate the generation of positive washback and mitigation of unintended effects that may be expected to emerge. The present study focused on the BCT-S, a tablet-based speaking assessment for university admissions developed by the British Council and Tokyo University of Foreign Studies. Nine senior high school teachers (Grades 10-12) were familiarized with the test and participated in a 90-minute interview, in which they were presented with the hypothetical situation in which the BCT-S was introduced for university admissions. Interview data revealed a range of expected washback effects on teaching practices that were mediated by numerous factors and varied by school type. The perceived goals of English education and the expectations of students and parents were crucial in determining the current approach to English education at the school. The perceived alignment of this approach with the skills assessed on the test was revealed to be a crucial mediating factor of the expected washback. Teacher factors, learner factors, and test proximity were also identified as likely to affect the intensity of washback. Based on the data, a Theory-of-Action Plan was created that proposes specific actions including teacher training in both classroom practice and regular assessment, support for developing teachers' confidence in speaking, and the provision of additional information about the test.

Research Papers

Time: **Saturday, 07/June/2025: 8:30am - 10:30am** *Location:* **Phramingkwan**

Culturally Tailored Assessments: Investigating the Role of Personalized Images in Writing Tasks

Andrew Runge, Geoffrey T. LaFlair, Jacqueline Church
Duolingo, United States of America

There is strong interest in incorporating test taker perspectives into language assessment to address concerns about fairness and justice in test design (Hamid et. al, 2019; O'Sullivan, 2012). Recent studies have argued for the consideration of test takers' linguistic, cultural and substantive patterns to better serve the needs of test takers from diverse backgrounds (Mislevy, 2018). One way to do this is to include culturally relevant stimuli that are personalized to test takers' backgrounds, but doing so requires careful research to ensure that opportunities for test takers to engage with the tasks and "test their best" are consistent across all groups of test takers. This study contributes to that goal by investigating the effects of locally personalized stimuli on test taker performance and their perceptions of a written picture description task. We source photos from 157 different countries, reviewing them for task suitability and appropriateness for a global audience. We conduct a pilot study as part of the online practice test for a large-scale test of English proficiency. Participants' home country is used to select images for the picture description task. Participants receive zero, one or two images from their home country, and the rest from other countries. We collect and analyze participants' task responses along with survey responses to evaluate their familiarity with the content in the images, their self-assessed performance, and their preferences for localized

images. The findings from this study contribute to a better understanding of how personalized assessment tasks can influence test taker performance and perceptions.

Testing Extended Time Accommodations: Differential Effects on Language Test Performance

Ramsey Lee Cardwell, William Belzak, Jill Burstein, Ruisong Li
Duolingo, United States of America

Accommodations in high-stakes standardized tests are essential for fairness, as they reduce disability-related construct-irrelevant variance, supporting cognitive validity (Weir, 2005). Accommodations research in language testing is limited, and studies in other fields (e.g., mathematics) present mixed findings on whether accommodations benefit everyone (differential boost hypothesis) or only those with disabilities (maximum potential thesis). This study examines the impact of extended time (ET) on a computer-adaptive English test, focusing on its effect on test-takers with and without disabilities.

The experimental design included random assignment to a control group (standard timed conditions) or an experimental group (50% more time, except for speaking tasks) on an online practice test that simulates the official test. Of 8,988 participants, ~18% reported at least one condition that could qualify for test accommodations (e.g., ADHD, autism, or physical conditions).

Results showed that the autism, learning, psychological, and speaking condition groups did not benefit significantly from ET, while those reporting no disabilities showed a minor increase in scores, less than the test's standard error of measurement (SEM). Groups with ADHD, hearing, and physical conditions, as well as those who declined to answer, exhibited significant score gains beyond the SEM, often on tasks aligned with their specific challenges. Writing tasks consistently showed among the largest gains across groups.

These findings partially support the maximum potential thesis, indicating that ET mitigates disability-related variance but is not equally effective for all groups or tasks. A more nuanced approach to accommodations is needed to maintain cognitive and construct validity.

Developing and Validating a Self-Assessment Tool for Measuring Vietnamese Teacher Competence in Multiple-Choice Test Item Writing for Large-Scale English Reading and Listening Skill Tests

Phuong Viet Ha Ngo[1], Shelley Gillis[2], Cuc Nguyen[2]
[1]University Canada West, Canada; [2]The University of Melbourne, Australia

The study constructed and empirically calibrated a developmental competency framework to define the skills and knowledge required by teachers to write multiple choice test items in English reading and listening skills for high stakes assessment in Vietnam's higher education settings. Through a combination of desktop review and consultations with subject matter experts, the major tasks performed by Vietnamese EFL teacher item writers were initially identified and expressed as observable, indicative behaviours with varying levels of quality criteria. In total, the draft framework comprised 14 indicative behaviours and 45 quality criteria that were converted into a self-assessment tool for teachers to complete. One hundred and fifty EFL teachers from various tertiary institutions in Vietnam completed the online self-assessment tool. The results of classical test and item response theory analyses

demonstrated that the tool had satisfactory measurement properties. The findings had direct implications for theoretical models of multiple-choice test item writing for English reading and listening skill tests, policy-related issues for developing teachers' capability in multiple-choice test item writing, training and practice of EFL teacher item writers in Vietnamese higher education settings.

Examining the Willingness to Communicate (WTC) Scale in Advanced Learners of Languages other than English (LOTE)

Troy L Cox
Brigham Young University, United States of America

This study examines the Willingness to Communicate (WTC) scale's reliability and predictive power for oral proficiency among advanced foreign language (FL) learners. Building on McCroskey and Baer's (1985) WTC concept and MacIntyre et al.'s (1998) situational model, we explored WTC among 3,009 advanced learners of Spanish, Portuguese, French, German, Italian, and Russian at a U.S. university, most of whom acquired language skills through immersive missionary experiences. Participants completed the WTC section of the LASER instrument and the Oral Proficiency Interview – Computer (OPIc).

Findings reveal that WTC varied significantly across domains, with lower willingness reported for online communication compared to educational settings, contrasting with studies showing higher digital WTC. Rasch analysis indicated reliable measurement with a refined 4-point scale (Cronbach's α = 0.88). A small but significant correlation was observed between WTC and proficiency at intermediate levels, though WTC showed limited differentiation at higher proficiency.

Results suggest that WTC may be context-sensitive, particularly influenced by immersion experiences and specific interaction settings. These insights could inform customized assessment and training approaches for FL learners, emphasizing contextually driven, communicative engagement in advanced language acquisition

Research Papers

Time: **Saturday, 07/June/2025: 8:30am - 10:30am** *Location:* **Room 401**

Evaluating the Logic of a Policy-Driven National Language Test in Indonesia: The Test of Indonesian Proficiency (UKBI)

Rahmad Adi Wijaya
University of Melbourne, Australia

This study examines the Test of Indonesian Proficiency (UKBI) as a national policy instrument used by Indonesia's language planning agency, Badan Bahasa. Language tests, like UKBI, have become integral to policy agendas, with growing research exploring their discursive and ideological underpinnings. Badan Bahasa promotes UKBI as a multi-purpose, adaptive test across education, employment and immigration domains. Despite its far-reaching potential impacts, no coherent interpretive argument nor validation agenda exists to guide the evaluation of measurement claims, let alone values and ideologies promoted by UKBI. The

study addresses this gap by investigating how UKBI's score meanings, uses, and impacts are represented and legitimized by Badan Bahasa. An argument-based validation framework is combined with critical discourse analysis (CDA) methods to interrogate test related claims from 50 documents - two government policy texts, 8 UKBI documents (technical manuals, proficiency scales, and test taker guides), and 40 promotional texts.

Findings reveal that UKBI is promoted for gatekeeping and accreditation purposes, with the potential to improve national literacy, and, in response to the growing influence of English, to reinforce Indonesian national identity. Claims are supported discursively, rather than empirically, through strategies of authorization and rationalization (Van Leeuwen, 2007). Authorization involves aligning UKBI with international best practice, while the logic of intended policy effects is rationalized using historical discourses of Indonesian nationalism, and global discourses linking language testing to the realization of both individual aspirations and various national policy goals. The study highlights the importance of a validation agenda that integrates social-political dimensions of testing.

Participants as Co-Researchers: A Co-Analysis Approach to Language Assessment for Immigration

Coral Yiwei Qin
University of Ottawa, Canada

This presentation introduces a co-analysis protocol, a novel qualitative methodology that addresses the limitations of participant involvement in language assessment research. Traditionally, participants are treated as passive data sources, and their lived experiences and cultural insights are often overlooked during the interpretation phase. The co-analysis protocol shifts the role of participants to co-researchers, involving them in the review and development of themes from data. This method was applied in a study on language assessment for immigration purposes in Canada, where 10 participants from 10 focus groups contributed to refining key themes during follow-up interviews. For example, the theme "Participants' Testing Experience" was revised to "Participants' Testing History" to reflect cultural distinctions between the Chinese terms 经历 and 经验, showcasing the linguistic precision introduced by participants. Similarly, the broader theme of Integration was split into "Self-identity within a Multicultural Context" and "Chinese-ness and Shared Cultural Traits", emphasizing the depth of analysis made possible through participant engagement.

The co-analysis protocol challenges traditional, researcher-driven approaches by demonstrating how shared authority can enrich qualitative research. This presentation highlights how the protocol generates deeper, more reliable data by incorporating participants' cultural knowledge and subjective experiences, expanding the scope of findings beyond what conventional methods can achieve. Additionally, the presentation highlights the policy implications from the method, offering insights for restructuring language assessments to better support immigrant communities. The co-analysis protocol also provides a model for collaborative knowledge production, amplifying the voices of marginalized groups and fostering cultural sensitivity in research.

A Study in Contrasts: Investigating Human and Automated Scorer Evaluation of Textual Changes

Sarah R. Hughes[1,2]
[1]University of Cambridge; [2]Pearson, United Kingdom

While the vast majority of automated scoring research to date has hinged on whether humans and machines agree on scores, it is arguably more important to develop methods for determining why humans and machines agree. Do humans and machine scorers value the same things in the same way when determining a score? Evidence that demonstrates humans and machines are influenced by the same aspects of a response when determining a score would provide persuasive support for the construct validity of an automated scoring system. This presentation reports on the second phase in a project to expand the construct validity evidence for assessments that use automated scoring by integrating eXplainable AI (XAI) methods and expert human judgement to improve explainability of automated scoring decisions.

This phase focuses on contrastive explanation. Contrastive explanations provide an answer to the question, “Why this score rather than another?” The aim of the contrastive explanation is to identify the key differences between the two scenarios that resulted in different outcomes. This study employs two methods of identifying influential factors and revising essays responses: human judgement and XAI.

The analysis considers quantitative agreement statistics and qualitatively maps the construct identified by humans and machine. The findings of this study show the utility of this method of validation. As automated scoring systems become more commonplace in high stakes assessment, the findings of this study will be of interest to those seeking new methods of evaluating construct validity and ensuring accountability and transparency of scoring procedures.

Investigating the Construct of the Continuation Task and Test-Takers' Cognitive Processes: An Eye-tracking Study

Yang Zhang, Lin Shi, Lianzhen He
Zhejiang University, People's Republic of China

The continuation task is an integrated writing task which requires test-takers to complete an unfinished story in a coherent way. While numerous studies suggest that this task can improve L2 learners' writing performance and overall English proficiency, its underlying construct remains underexplored. Previous research indicates that the continuation task is not simply a combination of reading and writing, but rather possesses a distinct construct centered on test-takers' alignment with the source text. However, there is limited empirical evidence on how alignment contributes to writing performance, and the cognitive processes involved in the task are not well understood.

To fill this gap, this study employed eye-tracking and stimulated recall interviews to investigate test-takers' cognitive processes. 35 L2 English learners of different proficiency levels participated in the study. Their eye movements were recorded when they completed the task. The findings indicated a significant positive correlation between writing scores and the alignment index, reaching a medium effect size. Eye-tracking data and interviews further suggested that: (1) Test-takers strategically align with the source text during the continuation task, and the strategies differ across proficiency levels; (2) Test-takers exhibit varying degrees of attention to different parts of the text, and this variance is correlated with their writing scores.

Results suggested that alignment is a key component of the construct, enhancing our understanding of test-takers' cognitive processes and the mechanism of alignment that occurs in the continuation task. Additionally, this study also provides empirical evidence for future validation work and the development of rating criteria.

Research Papers

Time: **Saturday, 07/June/2025: 8:30am - 10:30am** *Location:* **Room 405**

Differential Item Functioning in Audio-Visual English Listening Comprehension Assessments Among Young Learners

Sun-Young Shin[1], Senyung Lee[2]
[1]Indiana University, United States of America; [2]Chonnam National University, Republic of Korea

This study aimed to detect and analyze differential item functioning (DIF) in audio-visual English listening comprehension assessments among young learners. Previous research has provided limited evidence on how different types of visual inputs impact listening test scores (Shin & Lee, 2023; Suvorov, 2015), and little is known about how these inputs may affect young learners differently at the item level. This study addresses that gap by investigating which test items display DIF and how these items are characterized in relation to stimulus and item type. A total of 50 English Language Learners (ELLs) and 78 non-ELLs from Grades 3, 4, and 5 across three U.S. public elementary schools participated. They listened to four types of stimuli: speaker-only, visual-only, speaker-and-visual, and audio-only, each varying in listenability. Two DIF detection methods, Mantel-Haenszel and logistic regression, were used. The results showed that DIF items were more frequently linked to video input than audio input, particularly with main idea and inference questions. No significant effect of different visual types on DIF items was found, and no items were identified as DIF across both video and audio stimuli, indicating that input modality plays a critical role in DIF. The implications of these findings for designing audiovisual listening tests and constructing comprehension questions will be discussed.

Assessing Business English Competence: The Role of Linguaskill Business Test in Multicultural Contexts

Aynur Ismayilli Karakoc
Cambridge University Press and Assessment

In today’s globalised world, the demand for English as a lingua franca, particularly in Business English as a Lingua Franca (BELF) contexts, has significantly increased. English is the primary medium for international business communication (Seidlhofer, 2011). However, while some BELF proponents prioritise message conveyance over proficiency, concerns about the importance of language accuracy and clarity remain, as miscommunication can lead to misunderstandings and financial repercussions for businesses (Hinner, 2005).

This study addresses the gap in assessing Business English proficiency in multicultural settings by examining the use of the Linguaskill Business test with university students in India (n=283) and English language lecturers (n=10). Employing a mixed-methods design, the research utilises student questionnaires, including closed and open-ended questions, along with focus groups with lecturers. Quantitative data were analysed using descriptive statistics, while qualitative data were assessed through thematic analysis (Braun & Clarke, 2006).

Findings reveal that test preparation enhances students' proficiency, valuing their awareness of proficiency levels. Lecturers noted that the test raises awareness of English-speaking practices, particularly among students from diverse backgrounds. Furthermore, the provided learning materials support effective curriculum development, aligning teaching with learning goals. The test not only boosts students' confidence in business communication but also bridges language learning with real-world workplace demands.

By situating BELF within India's evolving business environment, this study highlights the need for integrating language assessments with EFL curricula. This approach enhances students' business communication skills and career readiness; while ensuring they are well-prepared for the test.

Charting the Landscape of K-12 Educators' Views on Automated Writing Scoring & Feedback

Jason A. Kemp, Lynn Shafer Willner
WIDA at the University of Wisconsin-Madison, United States of America

The use of automated writing scoring and feedback (AWSF) in assessments is not new (Elliot & Willamson, 2013), but its relevance has grown with the advent of tools like ChatGPT. As these digital tools become more common in K-12 classrooms, it is crucial for test developers to understand how educators are using them. Educators' perspectives are vital, especially when considering AWSF for large-scale assessments of multilingual learners' English language proficiency. Ignoring educator input would violate the WIDA value of Collaboration, which is essential in a consortium of 41 states and territories.

Previous AWSF research in U.S. K-12 settings did not focus on multilingual learners. This study explored K-12 educators' views on AWSF for a diverse group of multilingual learners. The study began with focus groups of 14 educators, whose feedback was coded based on fairness and justice principles (Kunnan, 2018). These findings informed a consortium-wide survey, with 739 educators from 32 states responding. Educators' familiarity with AWSF varied. Some believed AWSF could reliably evaluate multilingual learners' writing, while others were concerned about barriers like students' digital literacy and disabilities. They emphasized the need to combine AWSF with human scoring.

Overall, educators supported the exploratory analysis of AWSF, with conditions: keeping educators involved, addressing bias and accessibility, and ensuring consistency and reliability. They appreciated being included in the study, and we will continue to collaborate with educators on this AWSF project. This study encourages other researchers to involve educators in similar efforts.

Expanding the Interactive Academic Listening Construct: Addressing the Gap Between Academic Lectures and Proficiency Tests

Burak Senel, Ananda Senel
Iowa State University, United States of America

Academic lectures in North American higher education are increasingly interactive, requiring both aural and oral second language competencies. Student questioning, such as confirmation checks and requests for repetition, supports aural comprehension and is central to knowledge construction in these lectures. However, traditional academic English proficiency tests, typically used in international students' admission decision-making, often rely on non-interactive listening tasks that fail to reflect this interactivity, leading to construct underrepresentation and reduced authenticity. This research addresses this gap by defining an interactive and integrated academic listening construct that integrates both aural abilities for lecture comprehension and oral skills for student questioning. Additionally, a model is introduced for more comprehensive academic listening assessments based on existing listening models (e.g., Bejar, et al., Buck, 2001; Vandergift & Goh, 2022, Wagner, 2004) and Dillon's (1988) student questioning model. As recent advancements in language models, text-to-speech, and speech-to-text technologies make it possible to develop tasks that better reflect real-world academic interaction, it is anticipated that an interactive and integrated academic listening model will be of timely use to test developers. The presentation delineates and demonstrates an example interactive and integrated academic lecture listening task developed with the above technologies and discusses some potential positive washback effects, including heavier pedagogical emphasis on oral student questioning skills, more realistic test-taker expectations about participating in future academic lectures in North American contexts that often require both listening and questioning, and facilitating the pre-arrival acculturation of Asian candidate students into North American academic lecture discourse.

Research Papers

Time: **Saturday, 07/June/2025: 3:30pm - 5:00pm** *Location:* **Ampai**

The Role of Prompt Engineering in Ensuring the Consistency Between Instructor and LLM Checklist Ratings on Written Summary Content

Yasuyo Sawaki[1], Yutaka Ishii[2], Hiroaki Yamada[3], Takenobu Tokunaga[3]
[1]Waseda University, Japan; [2]Chiba University; [3]Institute of Science Tokyo

The rapidly-growing large language model (LLM) applications to L2 instruction and assessment in recent years informs explorations of options for timely provision of fine-grained feedback on traditionally underexplored, complex task types such as summary writing. Yet, key validity issues such as the consistency between LLM ratings and instructor ratings and effects of different prompts for automated scoring and feedback on the consistency require careful examination. The present study addressed exactly these issues, specifically focusing on checklist-based rating on main idea representation in written summaries (Kintsch & van Dijk, 1978; van Dijk & Kintsch, 1983). Ninety-seven summaries written in English by undergraduates in Japan were analyzed. Two writing course instructors rated all summaries

with partial double rating. We then developed six prompts by manipulating two features: the amount of information included (three types including few-shot); and the order in which rating and its explanation were generated in the LLM output (two types). By employing OpenAI GPT-4 turbo through Open AI API, we examined the instructor vs. LLM rating consistency based on agreement indices and confusion matrices. Results showed satisfactory levels of agreement for low-stakes purposes for certain prompt type-by-item combinations, with a notable effect of the amount of information included in the prompt. LLM ratings were also found to be generally harsher than human ratings. Key results will be discussed along with qualitative analysis results of the LLM output as well as study implications for LLM analysis of checklists and their applications to granular feedback provision.

Performance of Generative AI models on Academic Writing Tasks: A Systematic Review

Livija Jakaite, Vitaly Schetinin, Chihiro Inoue, Stanislav Selitskiy
University of Bedfordshire, United Kingdom

With the rapid advancement of Large Language Models (LLMs), underlying ChatGPT, With the rapid advancement of Large Language Models (LLMs), such as ChatGPT, concerns have arisen regarding their misuse in academic writing tasks. These models generate human-written-like text, allowing students to use them for written assignments, multiple-choice, and short-answer questions, undermining academic integrity. This study conducts a systematic review to explore detection methods for AI-generated content in assessments. Seven databases, including ACM Digital Library and Science Direct, were searched, yielding 1193 abstracts. After screening, 106 abstracts were reviewed, with 41 articles selected for inclusion in the full-text review. Results show that ChatGPT performs well on restricted tasks but struggles with open-ended, advanced questions requiring in-depth knowledge and analysis. The linguistic features that may differentiate AI-written texts from human-written texts include register awareness, syntactic simplicity, deep cohesion (how well ideas are connected), use of modals and epistemic markers, and redundancy. As AI rapidly evolves, detecting AI-generated responses becomes increasingly difficult, highlighting the need to redefine the skills required in education. The study contributes to the discussion of the shift towards developing human abilities like creativity, critical thinking, and problem-solving, which AI cannot easily replicate.

A Diagnostic Classification Model Analysis of Thai EFL Learners' Academic English Writing

Apichat Khamboonruang
Chulalongkorn University, Thailand

While diagnostic classification modelling (DCM) has attracted growing interest in second language (L2) assessment research, its application in L2 writing assessment remains relatively underexplored. This study applied a diagnostic classification modelling approach to diagnose Thai EFL learners' academic English writing. Writing samples were expository and argumentative essays written by Thai EFL English-major undergraduate students. Three Thai EFL university teachers rated the essays using a binary diagnostic checklist which consisted of 30 descriptors measuring five writing attributes: content, organisation, grammar, vocabulary, and mechanics. The raters also developed a Q-matrix which specifies the relationship between the descriptors and writing attributes, and which was employed in a subsequent DCM analysis. The DCM analysis of the data was performed using the Q-matrix and the generalised deterministic inputs noisy-and-gate (GDINA) model in order to estimate the learners' mastery of the writing attributes. The findings revealed that the GDINA model provided meaningful diagnostic insights into the students' writing strengths and weaknesses. Overall, learners exhibited high mastery probabilities across all writing attributes, with the strongest performance in organisation and the weakest in mechanics. Interestingly, students with similar or identical overall scores displayed varied mastery profiles for individual attributes. This highlights the nuanced diagnostic information provided by the DCM model, offering valuable formative feedback to support individualised learning. The results of this study have significant implications for L2 writing instruction and assessment, demonstrating the potential of diagnostic classification modelling to offer more detailed, targeted feedback that can enhance instructional practices and learner outcomes.

Research Papers

Time: **Saturday, 07/June/2025: 3:30pm - 5:00pm** *Location:* **Phramingkwan**

"Beyond Anxiety": Unveiling the Emotional Washback of the High-Stakes Hanyu Shuiping Kaoshi (HSK) on L2 Chinese Learners through Control-Value Theory (CVT)

Yang Yang
City University of Macau, Macau S.A.R. (China)

This study investigates the less-explored impact of high-stakes testing on learners' emotions beyond negative feelings like anxiety. Focusing on the Hanyu Shuiping Kaoshi (HSK), a prominent high-stakes Chinese proficiency test, the research examines its washback effects on the emotional experiences of students learning Chinese as a Second Language (CSL). By employing Control-Value Theory (CVT) from Positive Psychology (PP), the study explores the mechanisms through which test perceptions affect learning emotions. An exploratory sequential mixed-methods design was utilized. Initially, focus group discussions with 20 CSL students were analyzed using inductive coding to develop the Students' Learning Emotions questionnaire. Subsequently, a conceptual model based on CVT was proposed and tested using Structural Equation Modeling (SEM). This model assessed how test-takers' perceptions of the HSK's design, validity, and importance interact with their emotions and influence test performance. Data were collected from over 300 HSK test-takers in China. Participants completed a test perception questionnaire at the beginning of their preparation period and a learning emotions questionnaire after completing the test. The significance of this study lies in its comprehensive approach that incorporates emotional factors into understanding the washback phenomenon. By focusing on the under-researched group of CSL learners, the findings highlight the critical role of emotional well-being in test design and implementation. The research offers valuable insights for enhancing language assessment practices and fostering more supportive testing environments.

Human-AI Collaboration Patterns of EFL Learners in AI-Assisted Academic Writing

Shasha Xu, Xiang Xu, Fanxi Shen
Zhejiang University of Finance & Economics, People's Republic of China

The rapid development of artificial intelligence (AI) has catalyzed a substantial evolution in technology-assisted writing, especially through the advent of intelligent writing assistants. AI-powered tools are reshaping academic writing practices, aiding users in drafting, revising, conducting literature reviews, and synthesizing information—tasks integral to scholarly work. Despite the advantages these tools offer, notable challenges remain, including potential overreliance on AI-generated content, which may compromise academic integrity and weaken fundamental writing skills. This study examines the impact of human-AI collaboration on the writing process within an EFL academic context. This research involved twelve Chinese master's students enrolled in an Applied Linguistics program. Participants were tasked with composing a 500-word AI-assisted essay on second language acquisition hypotheses within a 45-minute. Screen recordings of each participant's writing process were coded via Elane software, identifying 37 distinct micro-level processes and resulting in a total of 789 coded events. The study applies an AI-driven learning analytics framework with two analytical layers: multichannel sequence analysis and pattern analysis. In the first layer, Hidden Markov Models (HMM) and sequence clustering were employed to detect and categorize underlying patterns in EFL learners' writing behaviors. The second layer utilized process mining techniques to provide a deeper analysis of these patterns. HMM analysis identified three distinct hidden states in students' writing strategies when interacting with the generative AI tool, showing a regular cyclical transition between state copying and pasting, as well as between pasting and component shaping. These hidden states were further clustered through Agglomerative Hierarchical Clustering, revealing significant variations in writing tactics. By examining the writing behaviors and tactics of these EFL master's students in technology-assisted academic writing, this study provides insights into the complex dynamics of human-AI collaboration and contributes to an understanding of AI-mediated language construct.

Digitalizing Language Tests for Migrants: Investigating a Multicultural Test Population's Digital Literacy and Target Language Use

Benjamin Kremmel[1], Eva Konrad[1], Doris Moser-Froetscher[1], Keri Hartman[2]
[1]University of Innsbruck; [2]Österreichischer Integrationsfonds

The majority of language tests conducted in the context of migration and citizenship are currently administered in a pen-and-paper format (Liana, 2022). Transitioning to digital test delivery offers benefits such as easier administration and enhanced authenticity (Ockey, 2009). However, this change may disadvantage test takers who lack experience with or access to digital devices. Previous research suggests that smartphones are highly important for refugees and migrants (Kaufmann, 2018; Alencar, 2020). Conversely, other studies have also shown a lack of access to digital devices and digital literacy among this group (Moran, 2023; IMEC, 2024). Additionally, transitioning to digital administration may also necessitate changes in test design and content.

Austria is planning to digitalize its national integration tests for adult migrants and refugees. The test population is characterized by high linguistic and sociodemographic heterogeneity. A needs analysis was conducted to investigate three main research questions and inform future test development for this national context:

a) How digitally literate is the target test population?
b) What is the (digital) target language use for the test population?
c) Are there any salient differences between population sub-groups in terms of digital literacy and language use?

An online survey was developed in six languages and administered to future test takers (N=400) enrolled in preparatory language and integration courses (A1-B2 levels). The results provide valuable insights into the digital skills and authentic language use of this multicultural test population and offer recommendations for adapting the test for digital delivery.

Research Papers

Time: **Saturday, 07/June/2025: 3:30pm - 5:00pm** *Location:* **Room 401**

Expanding Construct in EAP Speaking Assessment: Defining And Operationalizing a Critical Thinking Perspective

Shengkai Yin
[1]Shanghai Jiao Tong University; [2]The University of Melbourne

Critical thinking (CT) is one of the crucial skills of the 21st century, hence a topic of considerable interest within the domain of assessing English for academic purposes (EAP). Despite its importance, the ability to think critically has not been clearly defined, nor explicitly taught or assessed in the extant EAP speaking instruction and assessments. Given that an effective rating scale represents the de facto construct of language assessments, this study aims to conceptualize and operationalize the construct of CT in College English Test – Spoken English Test Band 6 (CET-SET6), a computer-based online EAP speaking test.

Framed in the argument-based validation framework (Knoch & Chapelle, 2018), this study was conducted in two phases, each focusing on different validity arguments. In the first phase, we described the domain of the CT construct in CET-SET6 and developed a CT rating scale for the speaking test. In the second phase, we collected validity evidence for the evaluation, generalization, and explanation inferences. The results indicated that raters achieved satisfactory inter-rater reliability at task-level and rater consistency between task types, and the categories can be reliably distinguished across different levels of difficulty, which was congruent with the statistical results. Quantitative results were triangulated with qualitative rater comments suggesting that the rubric can effectively capture variations of CT features in student performance. This study contributes to a nuanced understanding of the construct of CT in the EAP speaking context, and provides insights into how the construct of EAP speaking assessments could be expanded to incorporate CT.

Validating an Assessment of L2 Interactional Competence in Online Text Chat

Xingcheng Wang
University of Melbourne, Australia

This study developed and validated a computer-mediated assessment of interactional competence (IC) for L2 English users. The assessment comprised six role-plays featuring common social actions (e.g., requests, refusals, invitations, apologies, and complaints) administered through WeChat. 88 test-takers engaged in dyadic text chat, and their performances were analyzed using Conversational Analysis (CA), which, together with established theoretical frameworks and rating scales of IC, informed the rating scale development. The scale encompasses four categories: Role enactment (RE), epistemics management (EM), interactional infrastructure (II), and (dis)affiliation management (AM), each operationalized across five rating steps.

A many-facet Rasch measurement (MFRM) analysis was conducted on 4,536 ratings generated by three trained raters following a fully-crossed design. Results showed that test-takers were meaningfully distributed across a 6-logit range (-2 to 4) and all raters demonstrated high internal consistency. Regarding rating criteria and task difficulty, AM and EM proved more challenging than RE and II, and disaffiliative, school-oriented social actions were more difficult than affiliative, daily-life scenarios. Fit statistics indicated appropriate model fit across all facets, with only three test-takers showing misfit. The rating scale performed as intended, demonstrating monotonic advancement in scores and effective discrimination of test-taker abilities across rating steps. The Rasch model explained over 66% of the variance, and no secondary dimensions emerged, supporting the unidimensionality of the construct. Additionally, the moderate correlations between IC and general language proficiency support IC as a distinct construct, demonstrating the value of the developed assessment in measuring L2 learners' written interactional abilities crucial for digital interaction.

Research Papers

Time: **Saturday, 07/June/2025: 3:30pm - 5:00pm** *Location:* **Room 405**

Enhancing Inter-Coder Reliability in Online Think-Aloud Protocols Through Visual Behavior Analysis

Ananda Senel, Nathaniel Owen, Oliver Bigland
Oxford University Press, United Kingdom

This study reports on an innovative screen-recorded think-aloud data coding protocol, designed to investigate test-takers' cognitive processes during an intertextual reading-into-writing summary task for academic admissions. The summary task was piloted and pretested with a largely European population. This study investigates the cognitive validity of the same task with an Asian test taker sample.

Fifteen university students from India, China, Kazakhstan and Oman participated in online continuous think-aloud interviews, screen-recorded and auto-transcribed in Microsoft Teams. Development of the coding framework followed three phases: (1) Two coders developed an initial coding scheme from one interview and established models of reading and writing; (2) Coders independently applied the scheme to a second interview, discussed challenges and refined the scheme; (3) Coders independently applied the scheme to a third interview and repeated the process. The finalized scheme identified four parent codes: task management and strategies, reading for writing, writing processes (inclusive of reviewing/revising), and visual behavior. The scheme was applied independently to three interviews by both coders to calculate inter-coder reliability.

Visual behavior codes emerged as crucial, directing coders to match on-screen behaviors with processing codes (e.g., 'deleting/moving text' matched with 'reviewing/revising'). Cohen's Kappa calculations revealed substantial overall agreement (κ = 0.638), across test-takers (mean κ = 0.653), and by code, with ‘task management and strategies’ showing almost perfect agreement (κ = 0.820). The high agreement level is attributed to the additional visual information provided to coders, enhancing the authenticity of data interpretation by providing a comprehensive view of test-takers’ integrated cognitive processes.

An Evidence- and Consensus-Based Approach to Ethical AI for Language Assessment

Carla Pastorino-Campos, Evelina Galaczi
Cambridge University Press and Assessment

The rapid integration of Artificial Intelligence (AI) technology has prompted society to consider its risks and benefits. Frameworks like United Nations’ Governing AI for Humanity and the European AI Act summarise and codify key concerns while providing guidelines to address them. These classify many language assessments, especially high-stakes ones, as high-risk due to their significant impact on educational, employment, and social opportunities. High-risk AI applications are expected to meet strict requirements as regulatory frameworks evolve. In language education and assessment, this includes AI-generated test content, feedback, scoring, and malpractice identification.

This research supports the development of ethical principles for AI in language assessment by analysing current evidence and stakeholder views. Our presentation will share findings from a review of publications on the ethical use of AI in education and language assessment, focusing on key actors like intergovernmental organizations and language testing providers. We will address recurring concerns such as fairness, transparency, and bias in AI applications, and the impact of AI on educational content quality.

Our review indicates that, despite numerous guidelines on the ethical use of AI in education, there is less focus on language assessment. However, general guidelines can be the blueprint for the development of domain-specific guidelines for language assessment. We will discuss the importance of principles like intelligibility, human-centered AI, and proportionality, and highlight specific considerations for our field, such as construct representation and validity.
We will conclude with recommendations for integrating ethical AI use in language assessment, emphasizing its role in ensuring equitable and valid assessments.

Adapting and Evaluating Formative L2 Comprehensibility Assessment Scale

Aki Tsunemoto[1], Rie Koizumi[2], Makoto Fukazawa[3], Yo In’nami[4], Ryo Maie[5], Mariko Abe[6]
[1]Kansai University, Japan; [2]University of Tsukuba, Japan; [3]University of the Ryukyus, Japan; [4]Chuo University, Japan; [5]Tohoku University, Japan; [6]Okayama University, Japan

Comprehensibility (the ease of understanding speech) is a key component of second language (L2) oral proficiency, playing an important role in effective communication (Saito & Plonsky, 2019). Research indicates that comprehensibility is influenced by various linguistic factors, such as phonological, temporal, lexical, and grammatical features of L2 speech (e.g., Isaacs & Trofimovich, 2012). Isaacs et al. (2018) developed an L2 comprehensibility scale to help teachers and learners identify the linguistic factors that contribute to comprehensibility and improve its instruction. However, no studies have yet explored the scale's applicability in classroom settings or its reliability in different contexts, such as Japan, where exposure to and interaction with the target language are limited, which was the study’s focus.

The original scale (Isaacs et al., 2018) included both holistic (comprehensibility) and analytic (pronunciation, fluency, vocabulary, grammar) rubrics with 5 levels. We selected 50 speech samples, each 45 seconds long, from a speech corpus (Abe & Kondo, 2019) containing extemporaneous speaking tasks performed by Japanese secondary school students (e.g., describing a favorite meal). Five university instructors, experienced in assessing Japanese students' L2 speech, rated the samples using the scale. After completing their ratings, the raters took part in an online semi-structured interview to discuss their experiences with the scale. A multi-faceted Rasch analysis revealed consistent scoring yet significant variations among raters, suggesting a need for rater training or calibration to ensure consistent application of the rating criteria. Implications for refining language assessment rubrics and their potential use in classroom settings will be discussed.

Research Papers

Time: **Saturday, 07/June/2025: 3:30pm - 5:00pm** *Location:* **Poonsapaya**

Are Proctors of High-Stakes Language Assessments Fair?

Will Belzak, Alina von Davier
Duolingo, United States of America

In high-stakes language assessments, fairness is essential to ensure valid results and maintain trust among stakeholders. Proctors, who oversee standardized testing conditions, can unintentionally introduce bias into the testing process (Isbell, Kremmel, & Kim, 2023). This study empirically examines the potential for bias and inconsistency in proctors of a large-scale remote English language assessment. Specifically, we evaluate how characteristics of both proctors and test-takers, such as gender, age, and country of origin, may influence proctors' decisions about score certifications and rule violations.

This research analyzes data from nearly one million test takers and hundreds of proctors who reviewed recorded test sessions asynchronously. Using random-effects modeling, we measure consistency in proctor decision making and examine whether demographic variables of proctors and test takers lead to systematic biases in decision-making.

The findings reveal considerable variability in proctoring decisions, with evidence of subtle biases related to proctors' and test-takers' countries of origin. This potentially reflects an “in-group, out-group” dynamic (Tajfel & Turner, 1986). These results inform new procedures for minimizing bias, including specialized proctor training and ongoing decision calibration. Ultimately, the study offers recommendations for enhancing fairness and ensuring that language assessment outcomes are based solely on test-taker performance and adherence to the testing rules, free from undue influence by proctors.

Investigating Standard Setter Cognition

Doris Moser-Froetscher, Stefanie Hollenstein, Robert Hilbe
St.Gallen University of Teacher Education, Switzerland

Decisions made by standard-setters directly affect student attainment thresholds. It is therefore crucial to better understand the black box of standard-setting (McGinty, 2010). While research into standard-setters' decision-making has examined group discussions (Papageorgiou, 2010), individual thought processes have not been investigated. Furthermore, decision research postulates that complex decisions are shaped by individuals' decision-making style and their preferential mode of processing (Newell & Bröder, 2008), a focus with relevance to language assessment (Baker, 2012; Eberharter, 2021).

Our research questions are:
RQ1: What are standard-setters' thought processes while deciding on the CEFR level of English reading items and level boundaries?
RQ2: What is the relationship between standard setters' decision-making style and preferred processing mode, and their severity and fit statistics?

This study was part of an alignment project linking computer-adaptive English and French L2 reading and listening tests to the CEFR, employing the descriptor-matching method (Ferrara & Lewis, 2012).

For RQ1, seven English reading standard-setters participated in a think-aloud study. Transcripts were coded deductively and inductively using MAXQDA. To address RQ2, standard-setters (n=45) completed the Rational-Experiential Inventory 40 (Pacini & Epstein, 1999) and the General Decision Making Style Inventory (Scott & Bruce, 1995), validated questionnaires gauging decision-making style and preferred processing mode. Questionnaire subscores were correlated with MFRM rating metrics.

Findings show that judges relied on CEFR descriptors, mainly using rational decision-making but with some intuitive elements, highlighting individual differences. This study provides insights into individual decision-making in standard-setting, with implications for training and managing standard-setting processes.

Intercultural Considerations When Developing Materials and Assessments for Minoritised Languages

Lynda Brigid Taylor[1], Jill Wigglesworth[2], Rosalie Grant[3]
[1]CRELLA, University of Bedfordshire, United Kingdom; [2]University of Melbourne, Australia; [3]WIDA, University of Wisconsin-Madison, US

The theme for LTRC 2025 addresses language assessment in multicultural contexts where divergent norms and traditions intersect to shape the context for language test development, sometimes raising issues and challenges. This can be true for the learning and assessment of minoritised languages, especially when collaborating with Indigenous communities seeking localised language education programmes.
Cross-culturally, there may be practical and logistical issues, as well as challenges associated with differing epistemological or theoretical perspectives.

This paper discusses the importance of developing relevant understanding and expertise among professional language testing specialists to promote learning and assessment practices that are 'culturally responsive' (Montenegro & Jankowski, 2017). It highlights the value of localised case studies in which Western approaches intersect with Indigenous ways of knowing, being and doing (Williams & Perrone, 2018).

The presentation explores two case studies relating to the development of learning materials and assessments for Indigenous languages in different parts of the world. The first concerns a project in Australia's Northern Territory to develop a learning resource for children whose first language is Dhuwaya (a Yolŋu Matha koine) (Wigglesworth et al, 2021). In the second case study, test developers partnered with school districts in Alaska to create a language proficiency assessment for children learning the local Yup'ik language (Grant et al, 2023).

The presentation draws out theoretical and practical principles that can be applied to other cross-cultural contexts, and reflects on the contribution and value of community-based, participatory and collaborative research when co-designing the development of linguistically and culturally sustaining assessment.

Poster Presentations – Saturday, June 7, 2025

Time: **Saturday, 07/June/2025: 1:30pm - 3:00pm** *Location:* **Room 104**

Exploring The Appropriateness of the IELTS Academic for Australian Teacher Registration: Insights fom Domain Insiders

Xiaoxiao Kong
University of Melbourne, Australia

Since 2011, the IELTS Academic has served as an English language proficiency test for teacher registration in Australia at early childhood, primary, and secondary levels. The validity and appropriateness of this practice, however, have rarely been investigated.

Guided by an argument-based validation framework (e.g., Knoch & Chapelle, 2018) and needs analysis for language assessments for professional purposes (LAPPs; Knoch & Macqueen, 2020; Long, 2005), this study investigates the linguistic and communicative demands of early childhood and school teachers in Australia, as well as the appropriateness and adequacy of the IELTS Academic for assessing English language proficiency for teacher registration. Document analysis (n = 106), focus groups (n = 37), survey (n = 123) and interviews (n = 15) were conducted in three sequential stages to explore language use characteristics of important and frequently occurring workplace communication tasks, as well as teachers' views on the domain relevance of the IELTS Academic test tasks. This poster showcases findings regarding differences in teachers' language demands across education levels, which translated to differences in the degree of domain representativeness of the IELTS Academic within the three education contexts. Additionally, a mismatch was found between the IELTS Academic and teachers' workplace communication in terms of task format and characteristics of expected response, raising concerns over the validity and appropriateness of the IELTS Academic for teacher registration purposes. Such investigations provide implications for policy formulation as well as the design and implementation of language assessment for teacher registration, which could in turn contribute to student outcomes.

Development of an English as a Second Language Proficiency Test for Spanish-speaking migrant children in Trinidad & Tobago

Romulo Guedez Fernandez
The University of the West Indies, Trinidad and Tobago

This study focused on developing an English as a Second Language Proficiency Test (ESLPT) specifically designed for Spanish-speaking migrant children aged 5 to 11+ in Trinidad and Tobago. The ESLPT aimed to assess reading, writing, and oral language skills based on the Common European Framework of Reference (CEFR) guidelines for young learners. A total of 366 children were assessed, with younger children (5-7 years old) taking only the oral test and older children (8-11+ years old) taking both written and oral tests. The written test included short-answer questions, a reading comprehension task, and a creative writing task. The oral test was conducted as an interview between the child and a bilingual examiner.

Comprehensive instructions for every section were presented in both Spanish and English. In instances where clarification was needed, bilingual invigilators interacted with candidates in Spanish, taking meticulous care to ensure their clear understanding of the instructions. The study found that while many children (77.52%) demonstrated strong oral English skills, a significant number (42.81%) struggled with reading and writing, particularly creative writing. Some children even exhibited difficulties in reading and writing in their native Spanish language. Based on these findings, the study recommends early intervention literacy programs for younger children, remedial teaching programs for older children, and professional development for teachers to enhance English language teaching in diverse classrooms.

Inclusive and Equitable Test Development: Stakeholder Involvement of the Jewish Community in Antwerp (Flanders, Belgium)

Mieke De Latter[1], Fauve De Backer[2]
[1]Artevelde University of Applied Sciences; [2]Ghent University, Belgium

In this poster presentation, we explore how involving stakeholders, particularly the Jewish community in Antwerp (Belgium, Flanders), enhances the fairness and validity of standardized assessments. In 2024, standardized Dutch reading tests were administered for the first time in Flanders, where the 'Flemish research center for assessment in education' promotes accessible and reliable testing based on the Universal Design for Assessment (UDA). This approach ensures that tests are accessible and free from bias, especially important in Flanders' super-diverse society. The Jewish community in Antwerp, with approximately 20,000 members, is a unique multilingual group with limited interaction with non-Jewish people. This distinct cultural context can influence students' familiarity with certain test content. In order to investigate what aspects of the reading tasks might inadvertently disadvantage or offend students from the Jewish community due to cultural or religious sensitivities, we adopted a Participatory Action Research (PAR) approach. By consulting directly with stakeholders, including Jewish educators, to understand their perspectives, we identified potential biases.

Our findings reveal how specific task contexts, such as social settings and leisure activities unfamiliar to this community, could affect test results. Additionally, we identified sensitive content related to religious beliefs that might provoke emotional reactions and influence test performance. By removing these barriers, we aim to ensure that assessments are fair and inclusive for all students, regardless of their cultural background. This study illustrates the importance of stakeholder involvement in creating culturally sensitive tests, with broader implications for improving the fairness of educational assessments in diverse settings worldwide.

Teacher-Student Collaborative Assessment in the Production-Oriented Approach to Improve English Writing Proficiency and Perceived Self-Efficacy of Chinese Undergraduate Students

Xi Li[1,2], Punchalee Wasanasomsithi[1]
[1]English as an International Language Program, Graduate School, Chulalongkorn University, Bangkok, 10330, Thailand; [2]College of Foreign Studies, Guangxi Normal University, Guilin, 541004, Guangxi, P.R.China

The production-oriented approach (POA) has emerged as a pedagogical approach aimed at overcome weaknesses in English language instruction at the tertiary level in mainland China which integrates strengths of Western instructional approaches into Eastern educational contexts (Wen, 2007, 2012, 2015). The POA comprises three core components: teaching principles, teaching hypotheses, and teacher-mediated processes. In the POA, assessment occurs after the motivating and enabling phases (Wen, 2015, 2016b). Despite the pedagogical benefits yielded by the POA, instructors encounter significant challenges in deciding effective assessment strategies, which inevitably affects the success of POA implementation.

Recent empirical studies on teacher-student collaborative assessment (TSCA) highlight its potential to mitigate challenges faced by instructors in university-level English writing classes in China (e.g., Fan, 2020; Sun, 2017, 2020; Sun & Wen, 2018). TSCA aims to enhance the efficiency and effectiveness of feedback on student work. By engaging in collaborative activities, such as, co-creating assessment criteria and providing peer feedback, TSCA empowers students in their writing development. This approach not only promotes writing proficiency but also significantly enhances students' perceived self-efficacy. This study seeks to explore the effects of TSCA within the POA on improving English writing proficiency and self-efficacy among Chinese undergraduate students, employing a mixed-methods approach with 30 participants. Quantitative data are gathered through pretests, posttests, and self-efficacy questionnaires, complemented by qualitative insights from semi-structured interviews. The results provide valuable guidance for educators and policymakers in refining assessment practices and suggest avenues for further research on TSCA's role in enhancing learning outcomes.

Impact of Corpus-Assisted Self-Assessment on Enhancing Speaking Proficiency and Reducing Anxiety in EFL Learners

Pei-Ju Hsiung, Po Han Wu, Wei-Ting Wu
National University of Tainan, Taiwan

This study examines how corpus-assisted self-assessment influences junior college EFL learners' English-speaking proficiency and foreign language anxiety. Forty students from a technological university in Taiwan participated in the survey, with proficiency levels between CEFR A1 and B1. The participants were divided into two groups: an experimental group that used corpus tools for speaking practice and self-assessment and a control group that followed traditional speaking instruction. The study aimed to explore how corpus-based activities could provide learners with authentic language input and how self-assessment might foster metacognitive skills and reduce anxiety.

Over six weeks, both groups received weekly 90-minute lessons, but only the experimental group engaged in corpus-based speaking practice and self-assessment. The research employed a speaking proficiency test and the Foreign Language Classroom Anxiety Scale (FLCAS) to measure the outcomes.

Results demonstrated that the experimental group improved their speaking abilities, particularly in interactive competence. Additionally, their foreign language anxiety decreased significantly compared to the control group. Qualitative analysis showed that students using corpus tools gained confidence in their speaking abilities, while self-assessment helped them become more aware of their progress and areas for improvement.

The study concludes that corpus-assisted self-assessment is an effective approach for improving speaking skills and reducing anxiety among EFL learners, especially for lower proficiency students.

Training Needs of Middle School EFL Teachers on Language Assessment Literacy: A Study Based on Quantitative Ethnography

Hui Liu[1,2], Xiaomei Ma[1]
[1]Xi'an Jiaotong University, China, People's Republic of; [2]Xi'an Jiaotong University City College, People's Republic of

This study examines language assessment literacy among middle school English teachers in Northwest China by using methods of quantitative ethnography. Through interviews and surveys with 40 teachers, it was found that most teachers have received assessment training but feel their skills need improvement, particularly in theoretical framework. The epistemic network structure of teachers' assessment literacy is mainly supported by theoretical knowledge, that is, knowledge of linguistics theories and fundamentals of testing theories, and practical operations, including statistics literacy, testing operation ability, testing context flexibility, and digital literacy. Junior middle school teachers focus on basic testing knowledge, while senior middle school teachers consider factors related to testing context. Teachers in Key middle school have a better understanding of linguistic theories. Teachers' training needs prioritize testing skills, linguistic theories, and testing theories, with a preference for online or blended training. The study suggests enhancing practical assessment abilities and offering diverse training to meet teachers' needs, which could help improve professional literacy and teaching quality, and reform educational evaluation systems. Future research should explore factors affecting assessment literacy and develop training programs for teachers at different career stages.

Language Assessment Practices from Public School Teachers in Chile: Balancing Contextual Factors

Salomé Villa Larenas
Universidad Alberto Hurtado, Chile

In the realm of language testing, the literature underscores the significance of language assessment literacy (LAL) among language teachers, advocating for focused attention on this area beyond general educational assessment. Inbar-Lourie (2008) asserts that "[l]anguage assessors need to be familiar with contemporary theories about the learning, teaching and assessment of grammar" to design assessment tasks that accurately reflect language constructs (p. 392). Recent research by Levi and Inbar-Lourie (2019) indicates that while

language teachers can apply some generic assessment knowledge from training courses, the "multi-componential complexity of language teachers' assessment literacies," highlighting their distinct needs (p. 13).

Within the Chilean context, there is limited understanding of English teachers' assessment practices. One notable study by Tagle et al. (2021) explored the assessment practices of pre-service and novice teachers, revealing an emphasis on assessing grammatical features in isolation, contrary to the communicative approach endorsed by the national curriculum. Despite Tagle et al.'s findings, the assessment competencies of in-service English teachers remain largely unexamined.

This poster presentation is part of a larger LAL study aiming to investigate the LAL of Chilean in-service public school English teachers using a mixed-methods approach. It seeks to identify the components of their LAL, including assessment knowledge, practices, beliefs, and contextual factors. Preliminary findings from interviews with 12 in-service teachers and six school administrators will be shared, focusing on contextual influences on assessment practices, thus contributing to a deeper understanding of the tensions teachers navigate in their language assessment work.

Redesigning a Kindergarten to Grade 12 ELP Assessment Score Report: Gathering Evidence from Multiple Perspectives

Ahyoung Alicia Kim[1], Jason A. Kemp[1], Fujiuju Daisy Chang[1], Kerry Pusey[2], Fabiana MacMillan[1]
[1]WIDA, Univ. of Wisconsin-Madison, United States of America; [2]University of Pennsylvania

In kindergarten to grade 12 (K-12) settings in the United States, English learners (ELs) are required to take an annual English language proficiency (ELP) assessment according to federal law. These assessments are aligned with English Language Development (ELD) Standards which influence instruction of ELs. Once ELs complete the test, score reports are provided to ELs, and their families and educators. Score report information is used to monitor ELs' language progress and determine instructional supports.

We present the process of redesigning a K-12 ELP assessment score report to better reflect the recently updated ELD Standards. The test in question is ACCESS for ELLs (hereafter ACCESS) which is aligned to the WIDA ELD Standards. ACCESS is currently used across 41 U.S. states and territories to measure students' development in their ELP in the four language domains of listening, speaking, reading, and writing. The Standards were updated in 2020, which will be reflected on the ACCESS test beginning test administration year 2025-2026.

This study describes a three-phase iterative process of redesigning a score report that is intended to be used by multiple audiences. Interviews were conducted with varying user groups. Three major groups include ELs, ELs' families, and educators (at the state-, district-, and school-levels)–who might represent different socio-cultural backgrounds. Therefore, they may have varying perspectives on the score reports. We discuss the necessity and challenges of gathering input on score reports from multiple perspectives. Findings provide practical implications for K-12 ELP assessment score report development.

Assessing Listening Skills in Vietnamese: A Case of Track Separation in Beginning Vietnamese

Hanh Nguyen
University of Pennsylvania, United States of America

Vietnamese is classified as a Less Commonly Taught Language, leading to limited contemporary teaching and testing materials. Traditionally, assessments were based on prior student knowledge. However, following the Southeast Asian Language Council's training series on applying backward design onto test design and material developments, the Vietnamese program at the University of Pennsylvania has begun prioritizing test creation prior to instructional activities using the GRASP model. This change significantly impacts the assessment of Vietnamese listening skills.

Most students in Penn's Vietnamese classes are heritage learners with varying exposure to the language; while sharing the struggle with tonal writing, they have vast disparity in listening proficiency. To accommodate diverse learner needs and create an equitable environment, the program will introduce two tracks for beginning students: one for heritage speakers and another for general beginners, starting in Spring 2025. These tracks will merge into a single Intermediate course the following year.

This initiative aims to provide balanced support for students facing similar reading and writing challenges while addressing their different listening proficiencies based on their backgrounds. The presentation will reflect on past assessment practices, introduce the new tracking system, and analyze the challenges students face in listening, discuss the development of assessment tools and scoring rubrics, interpretation and application of scores, and the language test validation process, ultimately evaluating the effectiveness of the new assessment design.

Raters' Professional Background as a Potential Source of Distinct Behaviors: A Mixed-Methods Investigation to Interpreting Quality Assessment

Xini Liao, Jinjie Wu, Mingwei Pan
Shanghai International Studies University, China

Rater effects have long been discussed as a major topic in language performance tests. Presumably, raters' professional background variables give rise to various types of rater effects. However, our review of current interpreting quality assessment (IQA) literature indicated that the effects of raters' professional backgrounds had not been borne out by empirical evidence yet. In view of the potential influences triggered by such backgrounds on test scores, the study adopted mix-methods to probe into rater effects on IQA. Overall, such backgrounds as interpreters and teachers produced significant effects on rater behaviors. Many-facet Rasch measurement (MFRM) results unveiled differential rater severity/leniency, self-(in)consistency. In what followed, qualitative introspection specified individual attributes indicative of raters' professional backgrounds. This poster will present some discussions centering on the source of differences between teacher raters (i.e., interpreting trainer) (TR) and interpreter raters (IR).

Empowering Teachers and Students with AI: Developing a Vocabulary Assessment for Korean EFL Learners on the CAFA Platform

Jung-Hee Byun[1], Kyung-Il Yoon[2]
[1]Gyeongsang National University HS, South Korea; [2]Notre Dame of Maryland University, US

This research project introduces the development process of an AI-powered vocabulary assessment based on the CAFA (Collective AI on the Foundation AI; Choi, 2024) framework and its digital platform (Choi, Kim, & Yoon, 2012). The project addresses the vocabulary knowledge gap derived from the mismatch between curriculum expectations and learners' actual vocabulary proficiency in Korean EFL context. It guides teachers in developing formative assessment tools using the CAFA framework. These tools help evaluate and enhance students' vocabulary knowledge and usage. Aligned with the 2025 Revised National English curriculum standards, the system assesses vocabulary skills across form, definition, and contextual usage through recognition and recall tasks.

Key features of the application include multiple aspects of lexical competence assessment, detailed feedback with explanations and additional vocabulary information, and score interpretation to help learners understand performance implications. The development process involved setting assessment goals, selecting vocabulary items and test constructs, and using generative AI for multi-turn assessments.

A preliminary study with 30 Korean students will be reported regarding vocabulary gains and positive feedback, supporting the system's potential. The poster will showcase the development process, system features, and study results alongside a live demo to engage attendees with the CAFA platform's capabilities in AI-driven language assessment.

The Use of AI to Generate Picture Prompts for Story Writing Tasks

Haeun {Hannah} Kim
The University of Melbourne, Australia

Language tests designed for young learners often use picture prompts to elicit written narrative responses from examinees. Picture prompts provide topical support as well as organizational support when presented as a sequence of images. This lessens examinees' cognitive effort required to create the story, allowing them to concentrate more on the language elements of their writing. However, creating a test task with picture prompts is resource-intensive because this process often entails commissioning illustrators. Beyond the monetary costs, there is also the time investment in collaborating with artists to ensure picture prompts meet the task specifications and maintain a consistent style. Given these challenges, this study explores the feasibility of leveraging gen-AI models, specifically DALL·E, DreamStudio, and Midjourney, to produce these picture prompts for narrative writing tasks. First, a picture prompt titled "Carlos's Rainy Day" from a sample ACCESS for ELLs test (Grades 9-12) was reproduced using each AI model. Various AI prompting techniques were compared to see which method would lead to an image that most closely mirrors the original picture prompt in terms of story elements (i.e., setting, characters, plot) and artistic style (i.e., color, scaling, positioning, dimensionality). The same procedures were applied to the picture prompt "A New Locker" which was also from a sample ACCESS for ELLs test (Grades 9-12) but had a different artistic style. The pictures generated with these methods will be presented in the poster presentation to show the potential applications and limitations of using gen-AI in designing picture prompts for story writing tasks.

Investigating the Impact of Languages Other Than English Subjects in the National Matriculation Test: A Poststructuralist Perspective

Chenyang Zhang
University of Melbourne, Australia

After the promulgation of the Belt & Road Initiative in 2013 in China, the role of languages other than English (LOTE) has been increasingly recognised in recent years. The national assessment policy encourages senior secondary students to consider a LOTE subject, in addition to English, in the National Matriculation Test (NMT), the most significant high-stakes test in China. This policy shift has captured scant attention on the impact of LOTE subjects in the NMT. Thus, the study from a poststructuralist perspective aims to investigate how policy actors at the different levels exercise their agency – “a discursively mobilized capacity to act” (Miller, 2010, p.495) – to navigate the impact of LOTE subjects in the NMT. To understand policy levels, this study viewed language policy as the process of “creation, interpretation, and appropriation” (Johnson 2013, p. 224), which can occur within different policy levels.

Data was collected through one-on-one semi-structured interviews with a local testing project coordinator, school administrators (n=2), LOTE teachers (n = 4), and students (n = 30). The constructivist grounded theory (Tweed & Charmaz, 2011) was applied to analyse all the data. Findings show that language tests can lead to dynamic consequences by providing possibilities and constraints for stakeholders to negotiate and exercise their agency in their sociopolitical space. Based on the findings, a test impact model was developed, arguing that test impact should be understood within the discursive attribute of the sociopolitical context and its dynamics and indeterminacy shape and are shaped by stakeholders’ agency at different policy levels.

Mapping Language Needs Through a Proficiency Framework: A Case Study

Troy L Cox
Brigham Young University, United States of America

This study highlights the effectiveness of using a proficiency framework, specifically the ACTFL guidelines, in conducting a Language Needs Analysis (LNA) within a multicultural Language for Specific Purposes (LSP) context. Though some scholars suggest analyzing language needs without overarching proficiency levels, this research demonstrates how proficiency scales provide a structured foundation for cataloging and assessing language tasks, particularly in diverse and multilingual settings. Proficiency scales act as an “altimeter,” mapping the complexity of tasks, while task frequency and criticality are charted as “latitude and longitude.”

The research involved a structured approach: (1) reviewing missionary training documentation, categorized by proficiency levels; (2) conducting focus groups with recently returned missionaries to gather qualitative data; and (3) consulting language training experts.

These steps informed a 41-item survey covering four core language skills and four proficiency levels, aimed at eliciting the frequency and complexity of tasks.

Administered to 1,419 native speakers across five languages, the survey minimized second-language limitations, with Rasch analysis confirming high reliability (Cronbach's alpha = .95). Findings revealed that even Novice-level speakers could complete some daily tasks, though Advanced proficiency was often essential for complex interactions. Variations across languages indicated a need for tailored language support.

This study underscores the utility of proficiency frameworks in LNAs, providing actionable insights for designing language programs suited to diverse, real-world settings

---

A Comparison of Writing Assessment between Japanese High School Teachers and Aptis English Test Professional raters

Chiho Young-Johnson
Georgia State University, United States of America

As more second language learners take international English proficiency tests, concerns arise about potential discrepancies between local teachers' evaluations and those of trained international raters. However, there is little research comparing the raters' evaluation of Japanese K-12 teachers in writing assessment to those of trained raters in international English proficiency tests.

In this mixed-methods study, 10 novice, untrained Japanese raters, who were all teachers in K-12 settings, and 10 professionally trained Aptis test raters participated. Each rater assessed 20 Aptis essays, post-training, using an unguided holistic scale. These essays, part of the Aptis test, were written by Japanese teenagers. Holistic writing scores were analyzed using many-facet Rasch Measurement (MFRM). Additionally, raters provided written and spoken justifications for their scores through questionnaires and semi-structured interviews, which were analyzed for content and themes using NVivo.

Quantitative analysis using MFRM showed no statistically significant differences in overall scores for rater severity or agreement. However, qualitative analysis of written and spoken justifications revealed that Japanese teachers focused more on content (task achievement and organization), while Aptis raters emphasized language use (grammar, sentence structure, and vocabulary). Japanese teachers tended to focus on specific errors, whereas Aptis raters evaluated overall language use and complexity.

This research contributes to understanding how raters' backgrounds shape writing assessment and has important implications for aligning classroom and exam rating practices, developing culturally responsive rater training, and enhancing teachers' assessment literacy. The findings can inform efforts to bridge the gap between local and international assessment practices in second language writing evaluation.

My B1 is not Necessarily Your B1 - Evidence from Germany

Hella Klemmert, Juliane Braeutigam
German Federal Employment Agency, Germany

In Germany, the CEFR plays a central role in language learning and testing. Although the CEFR is deliberately designed as an adaptable framework, tests certifying the same skills at the same level should set comparable requirements.

Irritations caused by a lack of consistency between language tests arise during the integration of migrants into the German labour market. The procedure for placement in training or work depends heavily on the person's knowledge of German. Decisions become difficult if, for example, the person has a B1 certificate but very limited language skills. As such inconsistencies have recurred over time, an empirical study was initiated at the German Federal Employment Agency. The study investigated whether different tests of general German language proficiency lead to essentially the same CEFR classifications.
The study involved n > 1000 clients of the Vocational Psychological Service of the Federal Employment Agency with German as their L2. All participants completed a German placement test and also reported the CEFR level of a previous language test. All language tests were produced by established providers.

The CEFR classifications of both tests (placement test and self-reported previous test) differ systematically with higher levels for the self-reported test, regardless of the time since this test was taken.

One explanation is that the self-reports were too optimistic. But they may also be broadly valid. Then there is a strong case for further investigation of the inconsistency, preferably by directly comparing the classifications from different language tests.

Analyzing Written Proficiency Levels in Portuguese as an Additional Language: Corpus-Driven Results from the CorCel Corpus

Elisa Marchioro Stumpf, Juliana Roquele Schoffen, Deise Amaral, Isadora Dahmer Hanauer
Federal University of Rio Grande do Sul, Brazil

This paper analyzes CorCel, a corpus of over 15,000 texts produced under exam conditions for the Celpe-Bras, Brazil's official proficiency exam in Portuguese as an Additional Language (PAL) evaluated within the exam’s score range. The written exam comprises four integrated tasks assessing both comprehension and production, with criteria based on audience awareness, genre, and effective use of input material. The corpus allows for detailed examinations of proficiency levels, focusing on the description of linguistic patterns.

Using Corpus Linguistics tools available on Sketch Engine, the study evaluates lexical diversity, genre adequacy, and coherence among different proficiency levels. Key findings reveal significant lexical differences: higher-rated texts (grade 5) are longer and more complex than lower-rated ones (grade 2). For instance, grade 5 letters incorporate more greetings and address the audience effectively, while lower-grade texts tend to be more fragmented and less coherent. Additionally, grade 5 writers exhibit a broader vocabulary and demonstrate better paraphrasing skills compared to grade 2 writers, who often rely on copying fragmented information from input texts.

The study emphasizes that higher-proficiency writers utilize linguistic resources more efficiently, contributing to more coherent and varied texts. Given the limited quantitative research on PAL proficiency, the findings from CorCel can enhance the assessment criteria and improve the validation of the Celpe-Bras exam, offering valuable insights for both teaching and assessment in the PAL field.

Rubric Co-construction in Language for Specific Purposes Assessments

Qiaona Yu
Wake Forest University, United States of America

Language for Specific Purposes (LSP) assessments should consider both technical/practical aspects and human/social elements, as Winke (2023) advocated for all language assessments, with the goal of fostering positive washback. Cognitively, the interdisciplinary nature of LSP requires collaboration between domain experts and language instructors (Nekrasova-Beker & Becker, 2017). Socioemotionally, learner agency needs to be cultivated and spatially–temporally situated in the assessments from a Complex Dynamic System Theory perspective (Larsen-Freeman, 2019).

This study takes student-engaged rubric co-construction as an organic foundation for cognitive and socioemotional learning interactions. Situated in a semester-long Business Chinese course offered in an American university, this study investigated how rubric co-constructed assessments incorporating the voices of different stakeholders may affect students' agency and identity. It designed an interdisciplinary rubric-referenced performance-based assessment series. For the four rubrics used, students sequentially acted as a recipient, an editor, a creator, and a role of their choice. In addition, students' self-assessment and peer-review referring to the rubric also contributed as 30% of their grades. Qualitative data (i.e., rubric construction screen-and-audio recording of four students, reflections by the four students, one instructor, and three domain experts) were collected, transcribed, and analyzed using the Process Tracing methodology.

The results show that learner agency was achieved and enhanced through the rubric co-constructed assessments. Phylogenetically, the agency-assessment complementarity developed from students taking control on grades over an opening structure, to improving performance over a guided structure. Ontogenically, learner agency revealed individualized trajectories dependent on students' initial states and varying spaciotemporal adaptation to the context.

Enhancing Reliability in Writing Assessment: Investigating the Use of Customised ChatGPT 4.0 and Human Raters

Turgay Han[1], Özgür Şahan[2], Doğan Saltaş[3]
[1]Ordu University, Turkiye; [2]University of Southampton, UK; [3]Ardahan University, Turkiye

Recent research has explored the use of generative Artificial Intelligence (AI), such as ChatGPT, for scoring writing assignments. While studies have examined ChatGPT 3.5 and 4.0 models, few have investigated a customized version of ChatGPT 4.0, and none have simulated human scoring behaviors in AI-based assessments. This study addresses this gap by using generalizability theory (G-theory) to examine the variability and reliability of holistic scores assigned by human raters and a customized ChatGPT 4.0 model, "MyChatGPT," as an automated scoring tool. The study focuses on postgraduate-level assignments submitted by international students at a UK university, aiming to evaluate whether integrating AI can enhance assessment reliability. Twenty-four research-based assignments were scored using a departmental holistic rubric by both the AI model and four trained human raters. To simulate human behaviors, the AI rater scored the assignments under four different conditions, varying the degree to which the rubric was referenced during scoring. The results revealed significant differences between the scores assigned by human raters and the AI rater across different conditions. G-theory analyses showed that human raters displayed inconsistencies, while the AI rater produced more reliable results. Additionally, integrating AI scores with human scores improved generalizability and dependability coefficients. The study concludes that incorporating AI into the scoring process can reduce variability in human scores, offering a more reliable and cost-effective assessment method. However, it emphasizes the need for ongoing training and professional development to improve human rater consistency.

Adapting to AI: A Qualitative Study of Teachers' Formative Assessment Practices and Perceptions of Generative AI

Angelie Ignacio
University of Toronto, Canada

Classroom formative assessment is crucial for enhancing student achievement (Bennett, 2011; Black & Wiliam, 1998), yet teachers encounter challenges when implementing it. Thus, this study investigates the challenges and barriers to implementing formative assessment for K-6 teachers and explores how technologies, including generative artificial intelligence (GenAI), can or have been leveraged to address the challenges and obstacles K-6 teachers face.

We interviewed eight teachers. The interview guide focused on three themes: (1) challenges and barriers to implementing formative assessment, (2) how teachers have integrated technology, including GenAI, as part of their teaching and assessment practices, and (3) teachers' overall impressions of the impact of GenAI on student learning, their teaching and assessment practices, and their profession.

Analysis was conducted using reflexive thematic analysis (Braun & Clarke, 2019). Preliminary results showed that major barriers to formative assessment included time constraints, issues with differentiating assessment practices for students with different skill levels, and structural barriers such as lack of educational support personnel and funding. Furthermore, teachers were resilient and adept at using online resources and technology as part of their teaching and assessment practices. Teachers use online tools and resources to provide immediate and ongoing feedback to students and continuously monitor student work. Finally, all teachers expressed positive opinions regarding GenAI and considered it to be a valuable resource

when developing learning tools, such as rubrics or guiding questions. However, most teachers have not fully explored using GenAI and have no plans to integrate it as part of their teaching and assessment practices.

Stepping Stones or Stumbling Blocks: Oral Proficiency Level Descriptors and Their Effects on Rater Confidence

Birgitte Grande[1], Clayton D. Leishman[2]
[1]Norwegian Defence University College, Norway; [2]US Department of Defense (Retired)

This study addresses an under-explored aspect of oral proficiency assessment: the influence of level descriptor language on rater confidence and assessment reliability. While previous research has investigated factors like rater training and bias, less attention has been paid to the tools used by raters—namely, proficiency descriptors. This presentation outlines a study examining how clarity and structure of descriptor language in proficiency scales affect rater confidence. The research focuses on three sets of oral proficiency descriptors: the Interagency Language Roundtable 1986 (ILR-86), NATO STANAG 6001 (STANAG), and the revised Interagency Language Roundtable 2021 (ILR-21).

Using a modified Delphi method, 18 expert raters, consisting of NATO and U.S. government language testers, evaluated their confidence in applying these descriptors across seven sub-categories of spoken language production. The study's three rounds involved confidence ratings, ranking influential factors, and agreement with thematic statements. Results showed that revisions to the ILR-21 scale notably improved rater confidence, while ambiguous descriptor language often decreased confidence, particularly when linked to native-speaker norms.

The findings suggest that clarity in descriptor language is key to enhancing rater confidence and, in turn, the reliability of assessments. However, rater experience and familiarity with a scale also play crucial roles in mitigating inconsistencies. This study advocates for continuous revision of language assessment tools to adapt to evolving language use and ensure valid, reliable proficiency assessments across diverse contexts.

Enhancing Students' Feedback Literacy Using Generative Artificial Intelligence (GAI)

Limei Zhang
Nanyang Technological Unversity, Singapore

Students' feedback literacy involves students' comprehension, evaluation of feedback, and Students' feedback literacy involves students' comprehension, evaluation of feedback, and active engagement in feedback processes. Research shows that developing students' competency in feedback literacy not only boosts their academic performance but also enhances their confidence in learning (Carless & Boud, 2018; Sutton, 2012; Winstone et al., 2021). In the era of large language model (LLM) and general artificial intelligence (GAI), pedagogical AI agents can give timely and personalised feedback for students (Lan & Chen, 2024), which provides optimal opportunities to promote students' feedback literacy.

The purpose of this study is to investigate how pedagogical GAI agent enhance students' feedback literacy in Chinese speaking. One hundred and six Singapore secondary students who learn Chinese as a Second Language were invited to participate in the study. They were engaged with a designed pedagogical AI agent to improve their Chinese oral ability for one school term while human teachers guided the design of learning activities. Their feedback literacy and oral ability were examined before and after the school term quantitatively and qualitatively. Results showed that their feedback literacy improved affectively, behaviourally and cognitively. At the same time, their Chinese oral ability was also enhanced. The study provides important implications for the design and use of GAI tools in improving students' feedback literacy and consequently their self-directed learning ability.

Washback of Multilingual Assessment

Karin Vogt[1], Dina Tsagari[2], Lucilla Lopriore[3]
[1]Heidelberg University of Education, Germany; [2]Oslo Metropolitan University - OsloMet, Norway; [3]Roma Tre University, Italy

Language assessment along 'monolingual' lines has been the dominant tendency in the field for years. However, due to the recent advocacy of the 'multilingual turn' in language education (May, 2014), assessment that can tap on the multilingual competence of learners and test-takers (De Angelis, 2021) has received more attention. There have been attempts either locally or nationally to provide assessments that can accommodate the needs of multilingual students (Schissel et al, 2018) often describing challenges and compromises (Seed & Holland, 2020; Reierstam 2020; Tsagari et al, 2022). Despite such developments, the field has not yet investigated the effect that the provision of multilingual assessment might have on language learning, teaching and classroom assessment ('washback' effect). The washback effect has been profound in the context of language education, influencing not only what is taught but also how languages are learned and assessed within and beyond the classroom walls (Hughes, 2002; Spratt, 2005).

The poster will offer an overview of the theoretical underpinnings of what is currently meant by 'multilingual assessment' as well as its affordances and challenges. It also presents a critical review of existing assessment research and current practices that taps on the multilingual competence of test-takers in a variety of settings and discusses positive and negative washback effects from such assessment practices. Suggestions will be made for future research and practice particularly in raising stakeholders' multilingual assessment literacy and responding to calls for the implementation of principles of JEDI (Justice, Equity, Diversity and Inclusion) within the field (Shohamy, 2022).

The Use of Automated Feedback in Turkish EFL Students' Writing Classes

Turgay Han[1], Elif Sari[2]
[1]Ordu University, Turkiye; [2]Karadeniz Technical University, Turkiye

This study examines the impact of feedback methods on Turkish EFL students' writing by comparing automated feedback from an Automated Writing Evaluation (AWE) system with conventional teacher feedback. A comparative design with pre-test and post-test measures was employed, involving university students receiving English-medium instruction. Participants were divided into two groups: the experimental group received combined feedback from the AWE system and their teacher, while the control group received teacher feedback only. Over an academic term, both groups completed six writing assignments, and data were collected through pre- and post-test writing tasks, AWE error analysis reports, and student reflections. The analysis focused on the effects of feedback on students' analytic writing scores and error frequencies in grammar, usage, and mechanics. Quantitative data, analyzed using the inferential statistics, showed significant reductions in grammar and mechanics errors in the experimental group, while qualitative reflections highlighted the immediacy and detail of automated feedback. Despite these improvements, no significant difference was observed in overall writing proficiency scores between the groups. Students valued automated feedback for its precision in form-related issues but noted its limitations in addressing content and organization. The findings suggest that automated feedback effectively complements teacher feedback by reducing errors, but human input remains essential for writing skills. The study underscores the importance of integrating automated tools with conventional methods in EFL writing instruction and calls for further research into their long-term effects on writing development.

The Impact of Exposure to Different English Accents on Chinese Children’s Learning of EFL --- Listening, Speaking, and Attitude

Zhuohan Chen
University of Oxford, United Kingdom

Incorporating diverse English accents in assessments is common practice, but its impact on test-takers—particularly young EFL learners—remains unclear. These learners frequently encounter varied L2 accents but may not be cognitively equipped to detect them due to limited exposure and input.

Previous research indicates that exposure to a wider range of English accents might enable adult learners to understand and to be understood by a broader range of English speakers. However, little research has focused on younger learners. The few studies conducted have yielded inconclusive results. Some found a facilitative effect on children’s generalisation of phonemic contrasts and recognition of words, while some indicated that accents slow them down. Likewise, little is known about how children react to different L2 accents (Dai & Roever, 2019; Lee, 2020), despite evidence from L1 research showing children befriending those with a similar accent.

Thus, this ongoing project examines how exposure to varied English accents impacts children's L2 English learning regarding listening comprehension, oral production, and attitudinal preferences.

The project consists of three studies:
1.Study 1 (systematic review) synthesizes existing research on L2 accent exposure in children, analyzing 32 articles from 804 screened sources.
2.Study 2 (cross-sectional) investigates how 54 L1 Mandarin children in China (aged 7-8) recognize and respond to different accents, selecting preferred English-speaking cartoon characters as potential teachers, followed by interviews.
3.Study 3 (experiment) assesses how varied vs. single accent exposure influences 108 children’s recognition and production of familiar and unfamiliar L2 words, using pre- and post-tests.

Rating Performance and Quality of Novice and Expert Raters of Integrated Writing: Findings from a Mixed Methods Study in a German Higher Education Context

Valeriia Koval[1], Ximena Delgado Osorio[2], Claudia Harsch[1], Johannes Hartig[2]
[1]University of Bremen, Germany; [2]DIPF | Leibniz Institute for Research and Information in Education, Frankfurt am Main, Germany

It has been broadly acknowledged that assessing source-based writing (e.g., reading-into-writing) is challenging for raters (Gebril & Plakans, 2014; Plakans & Ohta, 2021). During the rating process, raters draw upon their prior experience and training to consider a range of significant elements of language usage and source integration. Despite the existence of a substantial body of research comparing the performance of expert and novice raters (Barkaoui, 2010; Cumming, 1990; Lim, 2011), no consensus has been reached regarding the impact of expertise on assessment quality. The objective of this mixed-method study is to examine the differences between these two groups when rating integrated writing tasks by combining qualitative and quantitative findings. For this purpose, we compared five novice and four expert raters, who rated the same 72 text products from integrated writing tasks using an analytic rating scale. For the qualitative strand, think-aloud protocol data were collected while raters assessed the same four text products. The coding and analysis approach were based on the findings of previous research on rating processes and strategies (e.g., Lumley, 2005; Cumming et al., 2002; Crisp, 2012). The quantitative strand examined differences between the groups in terms of rater severity and inter-rater agreement. The results of the study indicate that training may exert a higher influence on reliability than expertise. The findings enhance our understanding of the role of rater expertise and rater training for scoring approaches and rating quality in the context of integrated writing assessment.

Towards Effective CLIL Assessment: Understanding Classroom Questioning Practices in Asian EFL Settings

Wenhsien Yang
National Kaohsiung University of Hospitality and Tourism, Taiwan

Numerous Asian EFL countries have applied CLIL at various educational levels to promote bilingualism. However, researchers and practitioners find CLIL evaluation the most difficult since it evaluates language and content. This two-year study collected written and verbal classroom-based CLIL assessment questions and tasks from formative and summative exams. We then examined their language structure, intellectual demands, and cognitive discourse functions (CDFs) in soft and hard CLIL courses with different language pedagogy and content priorities in social science disciplines at the tertiary level in Taiwan, Thailand, and Japan. In addition to a student questionnaire, class observations and focus group interviews with teachers and students were used. This presentation presents the results of the first year's investigation. Our initial research examined 1,405 minutes from 39 CLIL courses. CLIL teachers use display, referential, and confirmation checks, which raise concerns about pedagogical objectives, challenge learners' cognitive development, and maintain an involved and meaningful classroom conversation. However, we found that teacher gender, learner level, and course duration significantly affect classroom questions. To foster more participatory discourse in CLIL classes, CLIL teachers must know the importance of questioning, eliciting accountable conversation, translanguaging, L1/L2 translation, and multimodalities and need enough professional development in classroom questioning skills.

Dynamic Assessment of Integrated Argumentative Writing: Diagnosing and Promoting Multilingual Writers' Development in the ZPD

Lu Yu
University of Melbourne, Australia

Argumentative writing in academic contexts often involves drawing upon source materials as supporting evidence. While L2 argumentation is commonly assessed through reading-writing integrated tasks, few studies have investigated multilingual writers' development in integrated argumentative writing over time. Reporting data from a larger dynamic assessment (DA) of L2 writing project, this case study examines the developmental trajectories of two multilingual writers as they learned to construct source-based argumentation. Amir and Xuan (pseudonyms), speaking Arabic and Vietnamese as first language respectively, were recruited from the same academic writing class at a U.S. university. In the initial DA procedure (DA1), the learners composed integrated argumentative writing independently, engaged in video recorded, interactionist DA with a mediator to review the essays, and then revised their drafts independently. With different prior learning experiences, they approached the initial writing task differently. The diagnoses of different learner ZPDs informed individualized, five-week writing instructional programs, after which DA1 procedure was repeated with a different writing topic (DA2). The study concludes with a non-DA, transfer assessment to determine writers' ability to transfer their learning to a more challenging task. Additional data sources include semi-structured interviews and blind independent ratings of the drafts. Changes in drafts were analyzed using Wingate's (2012) essay writing framework to examine the interconnection between establishing an own position, engagement with sources, and essay structure. Microgenetic and macrogenetic analysis of mediator-learner interaction within and across DA sessions revealed the learners' divergent developmental trajectories as impacted by the mediation provided through DA and DA-informed writing instruction.

Proposing and Applying a Conceptual Model of Test Fairness in a Local Context in China

Juan Zhang, Lianzhen He
Zhejiang University

Fairness has emerged as a key dimension of test evaluation (Geisinger, 2015). Existing research has identified key focuses of test fairness for international language tests (Xi, 2010). However, further research is needed to deepen the conceptual understanding of test fairness in the context of high-stakes local language tests, as test fairness is multifaceted, socially-constructed, and contextually-situated in nature. In this presentation, we will first introduce the development of a two-layer conceptual model of test fairness, structured in concentric circles. The inner circle encompasses four dimensions of test fairness: Comparability, Accessibility, Consistency, and Accountability. Surrounding these dimensions is an outer circle comprising key stakeholder groups: test developers, test-takers, test administrators, and test users.

Guided by the model, this study employed a convergent mixed-methods design to examine the stakeholders' perceived fairness of a high-stakes tertiary-level English proficiency test in China. Quantitatively, a questionnaire was administered to 1,646 test-takers. Exploratory factor analysis of the questionnaire data revealed a congruence between the empirically derived factors and the dimensions in the model. Qualitatively, one-on-one semi-structured interviews were conducted with 20 test-takers, six instructors, two administrators, and three policymakers. Thematic analysis of the interview transcripts identified four themes corresponding to the dimensions in the model and four themes representing the sociocultural, educational, institutional, and personal factors that influenced the stakeholders’ perceptions. The model was refined by incorporating an additional layer of these factors. This refined model is expected to guide evaluations on the fairness of local language tests and promote the fair use of local-, national-, and international-level language tests in China.

Paper and Demo Summaries – Sunday, June 8, 2025

Research Papers

Time: **Sunday, 08/June/2025: 8:30am - 10:30am** *Location:* **Ampai**

Exploring Profiles of Researcher-Teacher Collaboration in Language Assessment Literacy Studies

Beverly Baker[1], Lynda Taylor[2], Louis-David Bibeau[3]
[1]University of Ottawa, Canada; [2]CRELLA, University of Bedfordshire, UK; [3]Université de Montréal, Canada

In this study we examine the collaborative nature of recent LAL research projects with teachers. Our primary research question was the following: To what extent can LAL research between academic researchers and teacher practitioners be characterised as "community-engaged research" (Hacker, 2013; Israel et al., 2005; Key et al., 2019; Wallerstein et al., 2008)?

Phase 1: We collected a sample of 100 published LAL studies over the past 15 years and reviewed them with reference to the Continuum of Community Engagement in Research (Key et al., 2019).

Phase 2: We then collected surveys with the authors of 25 of the studies from Phase 1, focussing on context- and equity-related aspects of the continuum that could not be gleaned from the academic publications. Questions included initiation of the study, relationship-building, power-sharing during project decision-making, and ownership and benefits of project outcomes. This survey was designed so open-ended questions could be answered with audio-recorded rather than typed responses, to encourage more extended answers and anecdotes.

The results of these analyses were combined to create profiles of community-engaged research represented by the current body of work and to identify barriers to beneficial impact or to greater collaboration. Given the constraints of the information found in the articles and the need to contact the authors for a full picture of the nature of the project, we call for a critical re-evaluation of what information about a research study is considered valuable to include in research reports.

Unpacking Test-Wiseness Strategies: Effects on Second Language Reading Performance

Ray J. T. Liao[1], Kwangmin Lee[2]
[1]National Taiwan Ocean University, Taiwan; [2]Western Michigan University, U.S.A.

The use of test-wiseness (TW) strategies in L2 reading tests is a concern for language testers, as these strategies allow test-takers to bypass the cognitive processes necessary to arrive at correct answers (Wu & Stone, 2016). Researchers examining the role of TW strategies in L2 reading assessments have typically considered their effects holistically, rather than distinguishing between the contributions of different types of TW strategies. This study aimed to investigate the dimensionality of TW strategies and estimate the impact of different types on L2 reading performance.

Participants included 531 English learners from a Taiwanese university. They first completed a TOEIC reading task, followed by a self-assessment of TW strategy use through an inventory of 32 items. TW strategies were categorized into four types: time-using, error avoidance, deductive reasoning, and cue-using strategies. We employed Item Factor Analysis and Structural Equation Modeling (SEM) to examine the relationships between these strategy types and students' TOEIC reading scores.

Descriptive statistics revealed that deductive reasoning strategies were used most frequently, followed by cue-using, time-using, and error avoidance strategies. A model comparison approach indicated that the data were best explained by a four-factor correlated model, as demonstrated by the global model fit indices. SEM analysis showed that error avoidance was the only strategy type significantly and positively associated with reading scores. Our findings revealed that TW strategies may not necessarily compromise the accuracy of students' L2 reading test scores. The implications of these findings are discussed to inform L2 teaching practices in reading assessments.

Research Papers

Time: **Sunday, 08/June/2025: 8:30am - 10:30am** *Location:* **Phramingkwan**

Investigating the Challenges of Language Assessment Across Contexts Through an Assessment Literacy MOOC

Carolyn Westbrook, Richard Spiby, Jordan Weide
British Council, United Kingdom

Assessment literacy has long been recognised as a fundamental component of teacher education (Stiggins, 1991; Popham, 2009). While initial teacher training courses are increasingly providing modules on assessment literacy, in-service teachers are often left to educate themselves through continuing professional development. However, even when such training is available, many factors prevent teachers from undertaking such courses, for example, time, money and location. For this reason, the Language Assessment in the Classroom MOOC was created to provide free-of-charge training to large numbers of teachers in different contexts.

The 4-week, free-of-charge course attracted over 31,000 learners from over 170 countries during the 5 runs conducted with moderators between 2018 and 2021. During the course, participants completed various tasks and posted comments on the discussion boards, interacting with the moderators and each other, resulting in over 126,000 comments. These comments, the end-of-week video transcripts and post-course questionnaire responses were fed into a corpus tool and analysed both quantitatively and qualitatively to investigate a) the main assessment literacy topics that participants struggled with on the course and b) the challenges they faced when assessing their learners in different contexts. Among the results for RQ1 are difficulties with assessment literacy terms, distinguishing between different types of assessments and assessing productive and integrated skills while the results for RQ2 suggest that online assessment and alternatives to this were among the major challenges for teachers in different contexts. In this presentation, we will present the results of this research, concluding with recommendations for assessment literacy curricula.

Development And Validation of a Language Assessment Literacy Rubric: A Case for University-Level English Teachers in China

Ling Gan[1], Xun Yan[2]
[1]Beijing Technology and Business University; [2]University of Illinois at Urbana Champaign

It is crucial to understand how teachers advance their language assessment literacy (LAL) from “novice” to “expert”, yet few measures exist specifically for language teachers’ LAL levels, unlike those available in general education. This study aims to develop and validate an analytic LAL rubric for university English teachers in China, examining its construct and psychometric features using an exploratory sequential mixed-method design.

The rubric, developed inductively and deductively from a longitudinal qualitative study, includes 13 dimensions across four categories: assessment knowledge, skills, principles, and identity, with descriptors for three levels of expertise (preliminary, intermediate and advanced). Rasch analysis of 415 pilot-testing survey responses confirmed the rubric’s unidimensional construct and its effectiveness in differentiating teachers’ LAL levels as expected. However, while the 13 dimensions were reasonably designed, the three levels were not distinct enough to fully separate teachers’ LAL, indicating the need for further refinement, particularly in level distribution within each dimension.

This study provides valuable insights for researchers developing context-specific LAL rubrics and emphasises the importance of examining the psychometric quality of LAL assessment tools, ensuring they align with local educational settings and effectively assess teachers’ LAL levels and needs.

Promoting Assessment Literacy and Professionalization in Language Testing: Reflecting on the Role and Impact of the Studies in Language Testing Series

Lynda Brigid Taylor[1], Nick Saville[2]
[1]CRELLA, University of Bedfordshire, United Kingdom; [2]Cambridge University Press & Assessment

Since the early 1990s, the field of language testing and assessment (LTA) has steadily expanded and professionalized. This is evident in the creation of professional associations at national, regional and international level, and in an ever-growing volume of published material. Taylor & Green (2020) noted how such initiatives support the professionalization of the LTA field and promote assessment literacy among differing test stakeholders.

This presentation examines the role and impact of one particular series of LTA-related publications: Studies in Language Testing (SiLT). Volume 1 appeared in 1995, published jointly by UCLES EFL and CUP. Since then, 56 titles have been published addressing a wide range of topics and chronicling the growth and professionalization of LTA. The series’ 30-year history is examined to analyse the content and focus of its many volumes and what they indicate about the field’s expansion in terms of LTA research, development and validation activity. The series steadily evolved to acknowledge developments in the field as well as the widening interest in language assessment literacy in educational and societal contexts beyond academia. Over its lifetime, it published books or papers by more than 300 academics and practitioners from around 40 countries worldwide. In evaluating the global impact of the series, our retrospective will draw on data analyses from a large-scale survey of readers and users, as well as from personal accounts and interviews with those who contributed to or benefitted from the series in various ways in their professional lives and careers.

Research Papers

Time: **Sunday, 08/June/2025: 8:30am - 10:30am** *Location:* **Room 401**

Towards a Framework of Critical Thinking for Assessing EAP Speaking

Shengkai Yin[1,2]
[1]Shanghai Jiao Tong University; [2]The University of Melbourne

As one of the fundamental skills of the 21st century, critical thinking (CT) is a topic of considerable interest within the domain of assessing English for academic purposes (EAP). Recent literature on EAP instruction and assessment indicates that EAP has evolved beyond a strict focus on language improvement to encompass discourse competence and broader academic literacy development, with CT playing an important role in academic communication. A fundamental consideration in educational assessment is the construct that defines the knowledge, skills, or abilities to be assessed. However, CT has not received due attention in the research literature on EAP speaking assessment, thus raising a legitimate concern about the underrepresentation of the academic speaking construct. This study draws on Macqueen's (2022) distinction between theoretical, stated, perceived, and operationalized assessment constructs, with an aim to describe the EAP speaking assessment construct by including the concept of CT. By establishing alignment between what the literature indicates (theoretical), what assessment requires (stated), and how people understand assessment (perceived), a CT assessment framework is proposed to reflect how test takers experience it (operationalized). Our study contributes to a more nuanced conceptualization of CT in the context of EAP speaking, with implications for EAP speaking test development.

An Exploratory Study for Specifying the Q-Matrix in Cognitive Diagnostic Assessment of Chinese EFL Speaking Proficiency: Combining Theory with Data

Shuting Zhang, Lianzhen He
Institute of Applied Linguistics, Zhejiang University, China

Cognitive Diagnostic Assessment (CDA) is gaining increasing attention in language assessment for its ability to identify strengths and weaknesses and provide tailored feedback (He et al., 2021; Lee & Sawaki, 2009; Lee, 2015). Achieving accurate examinee attribute classification in CDA requires valid Q-matrix specification, yet many studies assume a Q-matrix is correct once constructed, rarely validating it or evaluating the appropriate number of attributes—posing validity concerns (DeCarlo, 2011; de la Torre & Chiu, 2016; Nájera et al., 2021). Therefore, this study adopts an exploratory approach to specify the Q-matrix for diagnosing Chinese EFL speaking proficiency.

Using Chinese Standards of English Language Ability (CSE) descriptors and rater feedback, a diagnostic checklist of 20 items was developed, covering vocabulary, grammar, pronunciation, fluency, coherence, content organization, and appropriateness. Four experienced raters assessed 400 examinee spoken responses in an in-house English Proficiency Test with this checklist. Principal component analysis (PCA) was applied to examine the internal structure of attributes assessed. Six Q-matrices informed by PCA, a purely expert-designed Q-matrix, and one modified by G-DINA package's built-in validation method (Ma & de la Torre, 2020) were compared in applying cognitive diagnostic models to the data.

Results showed that sGDINA, utilizing a Q-matrix with four attributes—pronunciation, fluency, language use, and content organization with meaningful cross-loadings—yielded the best model fit. This study proposes an exploratory approach for specifying Q-matrix, and highlights the importance of accurately representing skills in test design, thus enhancing diagnostic feedback's reliability.

Duolingo 2024 Doctoral Dissertation Awards

Wiktoria Allan

Lancaster University

Through a focus on extended time accommodations, Wiktoria Allan's research investigates how English language testing can better support students with ADHD. This study examines how learners with and without ADHD use and perceive additional time in reading and listening-to-write tasks, as well as whether these accommodations are helpful or harmful. The research addresses a critical gap in second-language testing research and the findings have the potential to shape policies that create fairer, more inclusive assessments, empowering neurodivergent students in high-stakes academic and professional settings.

Carla Consolini

University of Oregon

What makes a second-language Spanish essay stand out? Carla Consolini's research analyzes linguistic features such as lexical diversity and complexity, and it aims to identify measurable predictors of high-quality writing throughout the development of learners of Spanish as a second language. By extending automated writing evaluation tools to Spanish, this work addresses a growing demand for large-scale language assessment, offering faster, more specific feedback for learners and educators alike.

Jieun Kim

University of Hawaiʻi at Mānoa

Does the way students take notes—by hand or on a keyboard—impact their listening test performance? Jieun Kim's research explores this timely question, analyzing how note-taking modes affect test outcomes and the content of notes among second-language learners. The findings offer an opportunity for test designers to consider policies that ensure consistent test scores, accurately reflect learners' listening abilities, and promote fairness for all.

Valeriia Koval

University of Bremen

Assessing integrated academic writing poses unique challenges, especially when raters must distinguish between a writer's original language and borrowed material. Valeriia's research explores how a source-use detection tool impacts the evaluation process, examining its effects on rater perceptions, processes, and the quality of their judgments. By comparing ratings with and without the tool, the study investigates whether this technology improves fairness, consistency, and accuracy in evaluating skills like paraphrasing and source integration. The findings are expected to enhance rater training and support the use of technology for more transparent and reliable academic writing assessments.

Sebnem Kurt

Iowa State University

Could virtual reality (VR) be the future of language assessment? Sebnem Kurt's study explores how VR-based testing compares to traditional methods in evaluating oral English proficiency. By addressing issues like accessibility, security, and cost, this research could revolutionize language certification, offering a flexible and reliable alternative that meets modern testing needs, including during crises like the COVID-19 pandemic.

Jennifer Kay Morris

Lancaster University

For professionals in Türkiye's private technology sector, English proficiency is often a key to global success. Jennifer Kay Morris explores how these workplace demands can inform the design of language-for-specific-purpose (LSP) tests. Using online ethnography, this research aims to align language assessments with real-world business communication needs, offering a model that can transform how workplace language skills are evaluated globally.

Xue Nan

Beijing Language and Culture University

Incorporating authentic audio from platforms like TikTok and Bilibili into Chinese language tests presents both opportunities and challenges. Xue Nan's research uses natural language processing (NLP) to develop models that assess the difficulty of these materials, ensuring consistency in listening evaluations. This work modernizes Chinese language teaching by leveraging real-world content, helping learners build practical communication skills while improving test fairness and reliability.

Chenyang Zhang

University of Melbourne

How do policies promoting languages other than English (LOTE) in China's National Matriculation Test (NMT) influence education on the ground? Chenyang Zhang's research examines this question, exploring the complex ways policymakers, teachers, and students interact with LOTE testing amid China's push for multilingualism under initiatives like Belt & Road. By developing a model to capture test impact, this study offers valuable insights into addressing resource disparities and reshaping policies to promote equitable multilingual education.

A world ready for you, created by Cambridge

We believe that English can unlock a lifetime of experiences and, together with teachers and our partners, we help people to learn and confidently prove their skills to the world.

5.5m
assessments taken every year.

25,000+
organisations accept our exams worldwide.

2,800
exam centres in 130 countries.

50,000
preparation centres in more than 130 countries.

3m+
teachers and learners use Cambridge One for digital learning.

cambridge.org/english

CAMBRIDGE

Where your world grows

BRITISH
COUNCIL

British Council English Language Research

British Council English Language Research experts drive innovation through impactful research in product development and education reform. We provide technical expertise, advice and guidance on English teaching, learning and assessment worldwide.

We are proud supporters of the language testing community through our Assessment Research Awards and Grants and our programmes aimed at fostering emerging talent in the field.

English Language Research | British Council

Scholarship and Funding Opportunities - Lancaster University

英検 公益財団法人 日本英語検定協会
Eiken Foundation of Japan

Eiken promotes the globalization of Japan.

Eiken is one of the largest English-proficiency tests in Japan, and is taken by people of all ages.
The test has been designed with practical situations in mind and covers a variety of topics ranging from realistic everyday conversations to enlightening social topics.

Our mission is to support people's lifelong learning of English and to do our utmost to strengthen the development of Japan's global human resources.

Since 1963

Founded in 1963, Eiken has a history spanning more than 60 years.

100m+ examinees

The total number of Eiken examinees has exceeded 100 million.

400+ certifications

Approximately 400 universities and colleges, including those in North America, have certified Eiken as a proof of language proficiency for studying abroad.

EXCELLENCE IN GERMAN LANGUAGE TESTING
goethe.de/exams
GOETHE INSTITUT
Sprache. Kultur. Deutschland.

ILTA
INTERNATIONAL LANGUAGE
TESTING ASSOCIATION

Montréal

47th Language Testing Research Colloquium

2-6 June, 2026   Double Tree Hilton

duolingo english test

Built on the Latest Language Assessment Science

- ✓ Accessible by design, supporting learners wherever they are
- ✓ Built on rigorous research and industry- leading security
- ✓ Integrates the latest assessment science and AI for accurate results

The future of language assessment is here

The Duolingo English Test is a computer adaptive test powered human-in-the-loop AI and supported by rigorous validity research. The test measures speaking, writing, reading, and listening skills, providing a deeper insight into English proficiency.

englishtest.duolingo.com