



ILTA

INTERNATIONAL LANGUAGE
TESTING ASSOCIATION



47th Annual Language Testing Research Colloquium

Global to Local Language Assessment
June 2 - 6, 2026 | Montréal, Canada

IELTS

Research funding programmes

Our research programmes support innovative projects by researchers and educational institutions exploring real-world IELTS applications.



Learn more



CAMBRIDGE
English

IELTS is jointly owned by the British Council; IDP IELTS;
and Cambridge University Press & Assessment

Platinum Sponsors



CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT



duolingo
english test

IELTS

Gold Sponsors



Montreal
BUSINESS
events

Silver Sponsors



Bronze Sponsors



your measure of success



Table of Contents

<u>Message from the ILTA President.....</u>	<u>1</u>
<u>Message from the LTRC 2026 Co-Chairs.....</u>	<u>3</u>
<u>LTRC Organising Committee.....</u>	<u>4</u>
<u>ILTA Executive Board and Committee Members.....</u>	<u>6</u>
<u>Awards and Grants Committees.....</u>	<u>7</u>
<u>Award Winners.....</u>	<u>8</u>
<u>Housekeeping and Important Information.....</u>	<u>12</u>
<u>Map and Floor Plans.....</u>	<u>13</u>
<u>Schedule (Preconference Days).....</u>	<u>14</u>
<u>Schedule (Day 1: Thursday, June 4).....</u>	<u>15</u>
<u>Schedule (Day 2: Friday, June 5).....</u>	<u>21</u>
<u>Schedule (Day 3: Saturday, June 6).....</u>	<u>26</u>
<u>The 2026 Samuel J. Messick Memorial Lecture Award, Sponsored by ETS.....</u>	<u>32</u>
<u>The 2026 Alan Davies Lecture Award, Sponsored by the British Council.....</u>	<u>33</u>
<u>The 2026 Cambridge/ILTA Distinguished Achievement Award.....</u>	<u>34</u>
<u>Pre-Conference Workshops.....</u>	<u>35</u>
<u>Opening Session: Featured Speakers.....</u>	<u>39</u>
<u>Symposia.....</u>	<u>41</u>
<u>Special Sessions.....</u>	<u>78</u>
<u>Summaries of Paper Session 1.1 (Day 1: 9:40 a.m. - 10:10 a.m.).....</u>	<u>82</u>
<u>Summaries of Paper Session 1.2 (Day 1: 10:15 a.m. - 10:45 a.m.).....</u>	<u>84</u>
<u>Summaries of Paper Session 2.1 (Day 1: 2:20 p.m. - 2:50 p.m.).....</u>	<u>87</u>
<u>Summaries of Paper Session 2.2 (Day 1: 2:55 p.m. - 3:25 p.m.).....</u>	<u>90</u>
<u>Summaries of Paper Session 3.1 (Day 1: 4:00 p.m. - 4:30 p.m.).....</u>	<u>92</u>
<u>Summaries of Paper Session 3.2 (Day 1: 4:35 p.m. - 5:05 p.m.).....</u>	<u>95</u>
<u>Summaries of Paper Session 3.3 (Day 1: 5:10 p.m. - 5:40 p.m.).....</u>	<u>98</u>
<u>Summaries of Paper Session 4.1 (Day 2: 9:40 a.m. - 10:10 a.m.).....</u>	<u>102</u>
<u>Summaries of Paper Session 4.2 (Day 2: 10:15 a.m. - 10:45 a.m.).....</u>	<u>105</u>
<u>Summaries of Paper Session 4.3 (Day 2: 10:50 a.m. - 11:20 a.m.).....</u>	<u>108</u>

(The table of content continues on the next page)

Table of Contents

<u>Summaries of Paper Session 5.1 (Day 2: 3:30 p.m. - 4:00 p.m.).....</u>	<u>111</u>
<u>Summaries of Paper Session 5.2 (Day 2: 4:05 p.m. - 4:35 p.m.).....</u>	<u>114</u>
<u>Summaries of Paper Session 5.3 (Day 2: 4:40 p.m. - 5:10 p.m.).....</u>	<u>117</u>
<u>Summaries of Paper Session 5.4 (Day 2: 5:15 p.m. - 5:45 p.m.).....</u>	<u>120</u>
<u>Summaries of Paper Session 6.1 (Day 3: 9:40 a.m. - 10:10 a.m.).....</u>	<u>123</u>
<u>Summaries of Paper Session 6.2 (Day 3: 10:15 a.m. - 10:45 a.m.).....</u>	<u>126</u>
<u>Summaries of Paper Session 6.3 (Day 3: 10:50 a.m. - 11:20 a.m.).....</u>	<u>129</u>
<u>Summaries of Paper Session 7.1 (Day 3: 12:30 p.m. - 1:00 p.m.).....</u>	<u>132</u>
<u>Summaries of Paper Session 7.2 (Day 3: 1:05 p.m. - 1:35 p.m.).....</u>	<u>135</u>
<u>Summaries of Paper Session 8.1 (Day 3: 4:25 p.m. - 4:55 p.m.).....</u>	<u>138</u>
<u>Summaries of Paper Session 8.2 (Day 3: 5:00 p.m. - 5:30 p.m.).....</u>	<u>141</u>
<u>Demos Summaries.....</u>	<u>144</u>
<u>WIPs Summaries.....</u>	<u>147</u>
<u>Posters Summaries.....</u>	<u>158</u>

Message from the ILTA President

It is wonderful to see LTRC returning to Canada for our 47th annual conference: *Bienvenue à Montréal!* This diverse, multicultural and multilingual environment provides the perfect setting to discuss language testing and assessment research in all its rich variety. It also provides abundant opportunity for social activities and fun, with access to great food and a lively cosmopolitan city, as well as areas of natural beauty.

I would like to thank the conference organisers on behalf of ILTA. We are very grateful to the two conference chairs – Heike Neumann and Beverly Baker – as well as the wider organizing committee: Angel Arias, Louis-David Bibeau, Christophe Chénier, Shahrzad Saif, and Suzanne Springer. It has been a pleasure to work with this committee and to see the conference take shape over the past three years. I am aware of the immense effort that has gone into planning this event, both in terms of the logistics and in developing a rigorous academic program. The end result will, I am sure, be impressive and memorable!

A glance through the program book reveals that the range and depth of research we will encounter during this LTRC is truly extraordinary, taking in a range of topics and methods. Part of this range and depth is down to the very well-chosen conference theme: *Global to Local Language Assessment*. This theme provides broad scope for research that highlights context. Judging by the presentation titles, we will see this theme addressed through a focus on contexts of language use, contexts of test/assessment administration, local contexts of test/assessment use, and the wider social and political contexts which influence all aspects of our field.

Context runs through the themes of the main plenary speakers, too. I am excited to see Christopher DeLuca deliver this year's Samuel J. Messick Memorial Lecture (sponsored by ETS) on the topic "From literacy to legitimacy: Validity in the assessment literacy ecosystem". This talk promises to expand on the concept of assessment literacy, highlighting the need for teachers and assessment developers to adapt to shifting contexts. The Alan Davies Lecture Award (sponsored by the British Council) will be delivered by Paula Winke. The title is "The local social life of language tests: Insight from recent qualitative and mixed-methods studies", and the focus will be on the value of qualitative and mixed-methods research in revealing context-specific issues among assessment stakeholders. Finally, Barry O'Sullivan, the recipient of this year's Distinguished Achievement Award (co-sponsored by Cambridge University Press & Assessment and ILTA) will discuss the connection between context and validity (and much else besides!) in his talk "... and no Bob, the answer isn't blowin' in the wind!".

Before the program of papers, plenaries, symposia, works-in-progress, and posters, the conference begins with four workshops. I thank all of the workshop presenters who will be providing an excellent opportunity for delegates to expand their knowledge of test development practices, new methodologies, and professional practice: Sonja Zimmermann and Katharina Karges (Assessing Languages Beyond English: Global Principles and Local Practice); Andrew Runge and Yena Park (Using LLMs for content generation and evaluation in assessment development); Karen Dunn and Dylan Burton (Using GLMMs to model rater-mediated assessment outcomes); and Slobodanka Dimova, April Ginther, and Xun Yan (Bridging Global Technological Innovations and Local Language Testing Realities).

I would also like to encourage conference delegates to explore a range of other ILTA-related sessions that will run throughout the program. Our “Newcomer session” on Wednesday 3 June at 4:15pm is highly recommended for anyone attending LTRC for the first time. We warmly welcome veteran LTRC attendees too! There will be opportunities to discuss ILTA’s recently revised Code of Ethics (Thursday 4 June at 1:20pm) or join other professional sessions running at the same time. Special Interest Group (SIG) meetings for a selected number of SIGs will run on Friday 5 June (12:20-1:25).

I extend a particularly warm invitation to all ILTA members to attend our ILTA Annual Business Meeting (ABM) on Saturday 6 June (11:20-12:30). The ABM might sound like a very official event, but it is always a fun gathering, and it is open to *every* ILTA member. We (the ILTA Executive Board) typically run through some updates on business matters (finances, membership patterns, etc.), but we will leave a substantial proportion of the time for members to discuss openly some of the exciting initiatives that ILTA is currently working on. We will also reveal the bids for LTRC 2029 during the meeting. We want to have a big crowd, so please come along! Did I mention there is also a free lunch?

On the topic of the ILTA Executive Board, I would like to say thank you to the wider team who work tirelessly to keep ILTA running. ILTA is a vibrant community, which means that ILTA office holders are frequently very busy. That busyness typically peaks in the weeks and months prior to LTRC. So, I am grateful for the hard work of all Executive Board members in preparation for LTRC, and for the support of our management company, Nardone.

I would, finally, like to express deep gratitude on behalf of the ILTA community to the generous sponsors who have supported different events, sessions, food, drink, and products at this year’s conference. Sponsorship makes LTRC viable and affordable for members, and we really are appreciative of the different organisations that contribute. This year, I acknowledge our platinum sponsors: Cambridge University Press & Assessment; Duolingo English Test, and IELTS; our gold sponsors: Goethe Institut and Montréal Business Events; our silver sponsors: ETS TOEFL and LTTC (Language Testing and Training Centre); our bronze sponsors: ACTFL, LanguageCert, Pearson, Precise Testing Solutions, and le Centre d’évaluation de compétences linguistiques (CECL) à l’Université de Québec à Montréal (UQAM); as well as other sponsorship support from the British Council, the Canadian Association of Language Assessment, IDP Education, Lancaster University, the University of Chicago Language Centre, and WIDA.

I look forward to seeing many of you at the newcomer’s session, the reception, or in the coffee or lunch breaks.

Profitez bien de LTRC et de Montréal !

Luke Harding
ILTA President
2026



Message from the LTRC 2026 Co-Chairs

Bienvenue à Montréal !

After three years of planning, we are delighted to finally welcome you to Canada and to Montreal for LTRC 2026.

LTRC 2026 offers a rich and exciting academic and professional program, thanks to you all. We had over 370 abstract submissions to consider—an embarrassment of riches. Interestingly, we have 11 symposia this time—perhaps a record for LTRC, and a testament to the wide-ranging international collaborations now so common in our field. To undertake the evaluation of more than 250 paper abstracts, we introduced a fair grade procedure using many facet Rasch analysis. We believe that this innovation supports greater transparency in our decision making (and as we all know, it also represents best practice in high stakes evaluation). We will share details of our use of this procedure—and the challenges we encountered—in our welcome session.

Given the theme of the conference, “Global to Local Language Assessment,” during our welcome session we also want to spend some time to situate the conference in its local context. To that end, we are privileged to welcome Natasha MacDonald and Amélie-Hélène Rheault to speak with us. Natasha will introduce our local context from an Indigenous perspective. This area is historically a gathering place for many First Nations--the custodians of the place where we gather for this conference. Amélie-Hélène will delve into some of the unique sociolinguistic characteristics of life in Quebec and in Montreal.

Thanks are in order to the rest of our LTRC organising team: in addition to our regular team decision-making, Suzanne Springer helped with social programs and merchandise production and Shahrzad Saif worked on sponsorships. Angel Arias and Christophe Chénier were a particular support during the abstract review process. Louis-David Bibeau has organised the volunteers, coordinated the production of the conference program, and generally helped us to keep the wheels from falling off! We would also like to thank the ILTA EB for their unfailing support and confidence in us, and our intrepid association manager Valerie Smith and our association assistant Audrey Jones for their responsiveness and proactive problem solving. We hope you will be indulgent with us, ignoring any mistakes or slips that you may notice.

Have a wonderful, stimulating and productive time with us this week.

Beverly Baker and Heike Neumann



LTRC Organising Committee

Co-Chairs

Heike Neumann, Concordia University
Beverly Baker, University of Ottawa

Organising Committee

Angel Arias, Carleton University
Louis-David Bibeau, Université de Montréal
Christophe Chénier, Université de Montréal
Shahrazad Saif, Université Laval
Suzanne Springer, Université de Québec de Montréal

Logo

Néstor José Mejía Padilla

Cover Design

Mary Catherine Heard, Nardone Consulting

Volunteers

Faramarz Amin-Lari, Université Laval
Youssef Bouaddi, Carleton University
Aliz Bodjona, Université de Montréal
Amira Ben Hmida, Université de Montréal
Solène Grenier
Lin Lu, Concordia University
Ivana Penelska, Université de Montréal
Ada Luna Salita
Luis Ochoa
Parisa Safaei, University of Ottawa
Anne-Marie Sénécal, Université de Sherbrooke
Ryuichi Suzuki, Concordia University
Jennifer Sycamore, Concordia University
Cesar Telo, Concordia University

Abstract Reviewers

Jayanti Banerjee
Aaron Olaf Batty
Emma Bruce
John Dylan Burton
Yuko Butler
Ramsey Lee Cardwell
Nathan T. Carr
Micheline Chalhoub-Deville
Sathena Chan
Carol Chapelle
Mark Derek Chapman
Jee Wha Dakin
Bart Deygers
Slobodanka Dimova
Kathrin Eberharter
Jason Fan
Kellie Frost
Evelina Galaczi
April Ginther
Elsa Fernanda Gonzalez
Tony Green
Ching-Ni Hsieh
Ari Huhta
Yo In'nami
Ofra Inbar-Lourie
Talia Isaacs
Daniel Isbell
Noriko Iwashita
Eunice Eunhee Jang
Hyunah Kim
Ute Knoch
Rie Koizumi
Eva Konrad
Judit Kormos
Benjamin Kremmel
Yong-Won Lee

Constant Leung
Ray J. T. Liao
Gad Lim
Lorena Llosa
Susy Macqueen
Meg Malone
Isadora Teixeira Moraes
Doris Moser-Froetscher
Fumiyo Nakatsuhara
Saerhim Oh
John Pill
Geoffrey Pinchbeck
Coral Yiwei Qin
Gladys Quevedo
John A. Read
Yasuyo Sawaki
Jamie Lee Schissel
Jonathan Schmidgall
Sun-Young Shin
Richard Spiby
Ruslan Suvorov
Lynda Brigid Taylor
Veronika Timpe-Laughlin
Carolyn Turner
Salomé Villa Larenas
Karin Vogt
Erik Voss
Elvis Wagner
Gordon Blaine West
Paula Winke
Mikyung Kim Wolf
Xun Yan
Cecilia Guanfang Zhao
Ying Zheng
Sonja Zimmermann

ILTA Executive Board and Committee Members

ILTA Executive Board 2026

President: Luke Harding, Lancaster University

Vice-President: Lia Plakans, University of Iowa

Secretary: Benjamin Kremmel, University of Innsbruck

Treasurer: Beverly Baker, University of Ottawa

Communications Committee Chair: Erik Voss, Columbia University

Members at Large

Bart Deygers, Ghent University

Kathrin Eberharter, University of Innsbruck

Cecilia Guanfang Zhao, University of Macau

Ruslan Suvorov, University of Western Ontario

ILTA Staff

Valerie Smith

Audrey Jones

ILTA Nominating Committee

Chair: Heidi Liu Banerjee, PSI Services LLC

Salomé Villa-Larenas, Universidad Alberto Hurtado

Jing Xu, Cambridge University Press & Assessment

Niles Zhao, National Institute of Education, Nanyang Technological University

Graduate Student Assembly

Chair: Duong Nguyen, Iowa State University

Vice Chair: Shanshan He, University of Western Ontario

Communications Chair: Chengyuan Yu, Kent 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 Grande 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 Lúcia Barbosa de Carvalho e Silva, Federal University of Sao Carlos – UFSCar

Language Assessment for Young Learners (Young Learners SIG)

Jeanne Beck, Educational Testing Service

Wenjun (Elyse) Ding, Oxford University Press

Haeun Hannah Kim, University of Melbourne

Test-taker Insights in Language Assessment (TILASIG)

Andy Jiahao LIU, University of Arizona

Ray Jui-Teng Liao, National Taiwan Ocean University

Accessibility in Language Assessment (Access SIG)

Ananda Senel, Oxford University Press

Ruolin Hu, University College London

Chihiro Inoue, University of Bedfordshire

Awards and Grants Committees

ILTA Best Article Award

Chair: Fumiyo Nakatsuhara, University of Bedfordshire
Aaron Batty, Keio University
Gad Lim, Cambridge Boxhill Language Assessment
Lyn May, Queensland University of Technology
Paula Winke, Michigan State University

2026 Robert Lado Memorial Award

Chair: Dan Isbell, University of Hawaii
Ping-Lin Chuang, Duolingo Assessments
Nazlinur Gokturk, Ministry of National Education,
Turkey
Shungo Suzuki, Nagoya University, Japan
Punchalee Wasanasomsithi, Chulalongkorn University
Language Institute, Thailand
Christine Doe, Mount Saint Vincent University, Canada
Constant Leung, King's College London
Ari Huhta, University of Jyväskylä, Finland

2026 Cambridge / ILTA Distinguished Achievement Award

Chair: Lynda Taylor, University of Bedfordshire
Lia Plakans, University of Iowa
Yong-Won Lee, Seoul National University
Nick Saville, Cambridge University Press & Assessment

ILTA / Duolingo Collaboration and Outreach Grant

Chair: Luke Harding, Lancaster University
Kathrin Eberharder, University of Innsbruck
Geoff LaFlair, Duolingo
Ruslan Suvorov, University of Western Ontario

TOEFL / ILTA Student Travel Grants

Chair: Lia Plakans, University of Iowa
Bart Deygers, Ghent University
Cecilia Guanfang Zhao, University of Macau

Pearson / ILTA Scholar Travel Grants

Chair: Luke Harding, Lancaster University
Benjamin Kremmel, University of Innsbruck
Lia Plakans, University of Iowa

SAGE / ILTA Book Award

Chair: Dylan Burton, Georgia State University
Bart Deygers, Ghent University
Daniel Lam, University of Glasgow
Susy Macqueen, Australian National University
Lianzhen He, Zhejiang University

Award Winners

Cambridge / ILTA Distinguished Achievement Award

Barry O’Sullivan OBE, British Council

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

Christopher DeLuca, Queen’s University

The Davies Lecture Award (Sponsored by British Council)

Paula Winke, Michigan State University

TOEFL / ILTA Student Travel Grants

Shishi Zhang, University College London

Dellannia Segreti, University of Toronto

Olivia Seow, Harvard University

Fahri Yilmaz, Middle East Technical University

Jennifer Morris, Lancaster University

Che-Han Chen, Georgia State University

Jinjie Wu, Shanghai International Studies University

Evelyn Díaz-Iturriaga, Georgetown University

Reza Monfared, Université Laval

Hadeel Arqbi, Georgetown University

Minh Nhut Nguyen, Macquarie University

Shuting Zhang, Zhejiang University

Soo Hyoung Joo, Teachers College Columbia

Yijia Luo, Zhejiang University

Rui Zheng, Zhejiang University

Chia-Hsin Yin, Ohio State University

Shuhui Yin, Iowa State University

Inyoung Na, Iowa State University

Lin Shi, Zhejiang University

Pearson / ILTA Scholar Travel Grants

Gladys Quevedo-Camargo, Universidade de Brasilia

Ramy Shabara, University of Hertfordshire, hosted by the Global Academic Foundation, Egypt

ILTA Best Article Award 2024

Assessing Receptive Skills in an Adult L2 Learner Population with Diverse Educational Backgrounds

Marieke Vanbuel, Ghent University

Bart Deygers, Ghent University

Published in: *Language Assessment Quarterly*, 21(4–5), 379–403. <https://doi.org/10.1080/15434303.2024.2438151>

Commendation:

Effects of the type of audio-visual input and listenability on young L2 learners' listening comprehension

Sun-Young Shin, Indiana University

Senyung Lee, Chonnam National University

Published in: *TESOL Quarterly*, 58(4), 1580–1608. <https://doi.org/10.1002/tesq.3298>

ILTA / Duolingo Collaboration and Outreach Grant

Language Assessment Intensive workshops in Chile

Salomé Villa Larenas, Universidad Alberto Hurtado

Dylan Burton, Georgia State University

SAGE / ILTA Award for Best Edited Volume on Language Testing

Fundamental Considerations in Technology Mediated Language Assessment by Karim Sadeghi and Dan Douglas, published by Routledge in 2023

and

Language Assessment Literacy and Competence Volume 1 and 2 by Beverly Baker and Lynda Taylor, published by Cambridge in 2024

***Language Testing* 2026 Reviewer of the Year Award**

Yeu-Ting Liu, National Taiwan Normal University

Jin Yan, Shanghai Jiaotong University, China

Robert Lado Memorial Award

To be announced at the LTRC Banquet

Jacqueline Ross TOEFL Dissertation Award

Laura Schildt, Ghent University

Caroline Clapham IELTS Masters Award

Elena Gorshkova, Michigan State University

Duolingo Doctoral Dissertation Awards

Erika Latham, Iowa State University

I-Chun Vera Hsiao, University of Iowa

Kelly Kendro, Northern Arizona University

Mi-Jing Chung, Lancaster University

Minkyung Kim, Seoul National University

Muhamad Firdaus Mohd Noh, Universiti Kebangsaan Malaysia

Sanshiroh Ogawa, The University of Maryland, College Park

Shanshan He, University of Western Ontario

Yue Li, University of Melbourne

Zhuohan Chen, University of Oxford

TOEFL Grants for Doctoral Research in Language Assessment

Gina Ward, Concordia University

I-Chun Hsiao, University of Iowa

Inyoung Na, Iowa State University

Kuan leong Mak, Queenie, The University of Wollongong

Mi-Jin Chung, Lancaster University

Qi Qi, University of Western Ontario

Shanshan He, Iowa State University

Shuhui Yin, Iowa State University

Soo Hyoung Joo, Columbia University

Wiktoria Allan, Lancaster University

Xiaoxiao Kong, The University of Melbourne

Yali Feng, Georgia State University

Yue Li, The University of Melbourne

TIRF's 2025 Doctoral Dissertation Grantees

Chia-Hsin Yin, the Ohio State University

Dissertation Title: Assessing Classroom-Based Writing of Adolescent English as a Foreign Language (EFL) Learners: A Comparative Study of Generative AI and Human Teachers

Funding Co-Sponsor: Duolingo

Advisor: Prof. Peter Sayer

Huiying Cai, University of Illinois Urbana-Champaign

Dissertation Title: Exploring a Feature-Based Explainable Automated Essay Scoring System for Local L2 Integrated Argumentative Writing Assessment

Funding Co-Sponsor: Duolingo

Advisor: Prof. Xun Yan

I-Chun Hsiao, University of Iowa

Dissertation Title: Fairness, Validity, and Impact of a Standardized English Proficiency Test in EMI Contexts and Beyond: A Mixed-Methods Study of Test Takers' Perceptions

Funding Co-Sponsor: Duolingo

Advisor: Prof. Lia Plakans

Inyoung Na (her), Iowa State University

Dissertation Title: Assessing Interactional Competence: Leveraging Generative AI for the English Placement Test at a U.S. University

Funding Co-Sponsor: Duolingo

Advisors: Prof. Gary Ockey

Housekeeping and Important Information



We have excellent drinking water here in Canada. Please just refill your water bottles from the tap and do not buy plastic bottles.



Please recycle your waste by putting it into the garbage cans provided. Please make sure your waste is sorted.



If any questions come up, please feel free to ask our staff at the registration desk (in Opera Foyer) at any time.

If you are feeling tired or overstimulated and need a place to take a break on your own, we have a quiet room available. Please ask at the registration desk.



Local information can be found on the conference website: <https://www.iltaonline.com/page/LTRC2026>

The full program can also be accessed using this link: <https://www.iltaonline.com/page/Schedule>



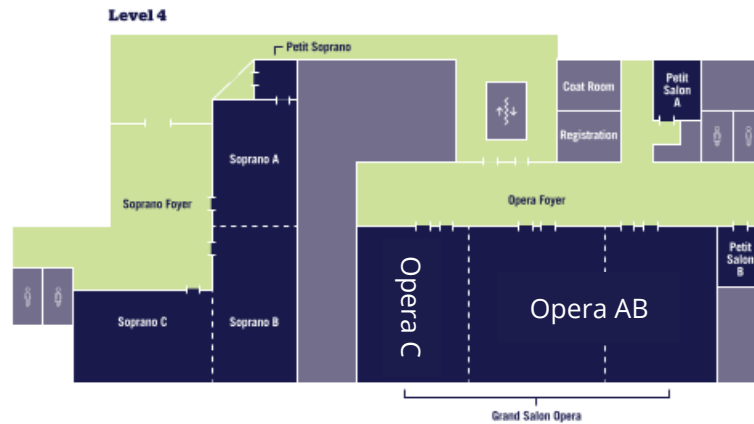
WiFi network: Hilton Honors Meeting
Wifi password: hiltonclient



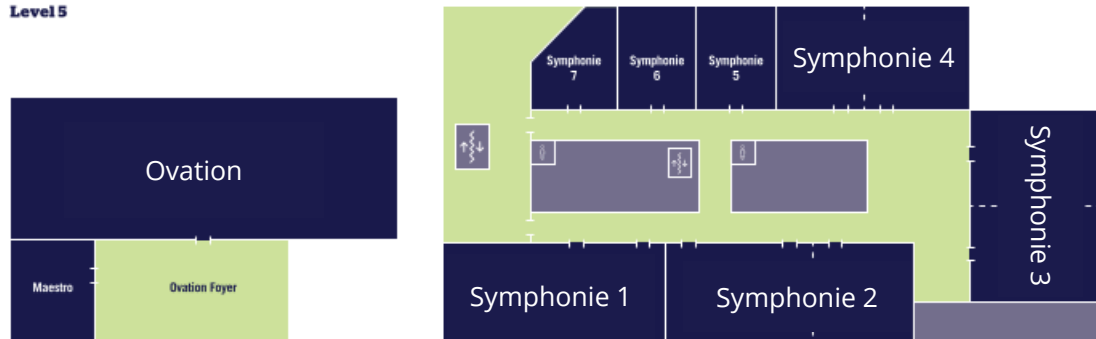
Map and Floor Plans

Floor Map Key

-  Elevators
-  Meeting/Conference Rooms
-  Amenities/Service Areas
-  Foyer Space/Public Areas



Level 5



Level 6



LTRC 2026 – Montréal
 6/2/2026 to 6/6/2026
 DoubleTree by Hilton, Montréal (Canada)
Schedule



Preconference Days Tuesday and Wednesday, June 2 - 3

Tuesday, June 2 (8:00 a.m. to 4:00 p.m.)

8:00-9:00	Workshop registration and breakfast (Symphonie Foyer)	
9:00-12:00	Preconference workshop A (Symphonie 3)	Preconference workshop B (Symphonie 2)
12:00-1:00	Lunch (Symphonie Foyer)	
1:00-4:00	Preconference workshop A (Symphonie 3)	Preconference workshop B (Symphonie 2)

Wednesday, June 3 (8:00 a.m. to 9 p.m.)

9:00-4:00	ILTA Executive Board meeting (room and lunch sponsored by UQAM Centre d'évaluation de compétences linguistiques)
-----------	---

8:00-9:00	Workshop registration and breakfast (Symphonie Foyer)	
9:00-12:00	Preconference workshop C (Symphonie 2)	Preconference workshop D (Symphonie 3)
12:00-1:00	Lunch (Symphonie Foyer)	
1:00-4:00	Preconference workshop C (Symphonie 2)	Preconference workshop D (Symphonie 3)
4:15-4:45	Newcomer session (Opera AB) Facilitators: Dylan Burton and Meg Malone	
5:00-6:30	Opening ceremony (Opera AB) Featured speaker : Amélie-Hélène Rheault	
6:30-9:00	Opening reception (Opera C and Grand Foyer) Cocktails sponsored by Duolingo, reception sponsored by Cambridge University Press & Assessment	

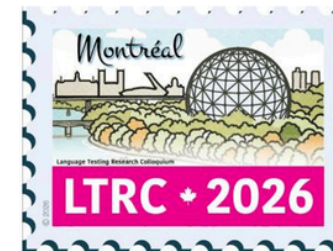
LTRC 2026 – Montréal
6/2/2026 to 6/6/2026
DoubleTree by Hilton, Montréal (Canada)
Schedule



Day 1 – Thursday, June 4
(7:30 a.m.-10:45 a.m.)

		Opera AB	Symphonie 1	Symphonie 2	Symphonie 3	Symphonie 4	Ovation
7:30-8:30	Registration and Breakfast (Opera C and Grand Foyer) Sponsored by Goethe Institut						
All-day	Exhibitors: ACTFL, Canadian Association of Language Assessment, Duolingo, ETS TOEFL, Goethe Institut & Pearson (Opera C)						
8:30-9:30	The 2026 Samuel J. Messick Memorial Lecture Award (Opera AB) Sponsored by ETS From Literacy to Legitimacy: Validity in the Assessment Literacy Ecosystem Christopher DeLuca						
9:40-10:10	Paper session 1.1	Assessing with AI: How Raters Negotiate LLM Suggestions in Writing Scoring Hongyi Yang, Gabrielle Gaudeau, Andrew Caines, Andreas Säuberli	Global Frameworks, Local Contexts: Evaluating the Cultural Appropriateness of English Language Tests for Vietnamese Candidates Brigita Seguis, Quynh Nguyen, Nadir Zanini	The Impact of Rater Cognition on the Assessment of Interactive Listening in Vietnamese Higher Education Context Huong Tram Anh Nguyen	Integrating Criticality in Language Assessment Education Lia Plakans, Kwangmin Lee, Marina Durinova	Reverse-Engineered Validity: Retrofitting a Legacy Multiple-Choice Test Through Assessment Engineering Principles Petia Alexevia, Troy L. Cox, Matthew Wilcox	Demo session 1
							Demo session 2
10:15-10:45	Paper session 1.2	Cancelled A GenAI Writing Examiner for ESL writing: Scoring quality and revision engagement in test-preparation and classroom contexts Daniel Yu-Sheng Chang, Chun-Chia Chang	Global Measures, Local Outcomes: ELP Test Predictive Validity in Graduate Admissions Allison Michalowski, Ramsey Lee Cardwell, Jill Burstein, Steven Nydick, Anthony Verardi	Integrated Speaking Tasks: Investigating Content Across CEFR Levels Nazlinur Gokturk, Sermin Vardal Ocakli, Gokce Fidan	Pre-service teachers' knowledge, practices and conceptions of writing assessment: A Comparative Study of Anglophone and Francophone Sectors in Quebec Amira Ben Hmida	The Role of Language Proficiency and Background Factors in Predicting International Students' Academic Success Okim Kang, Yuna Bae, Masha Kostromitina, Harriet Denta	Demo session 3

LTRC 2026 – Montréal
 6/2/2026 to 6/6/2026
 DoubleTree by Hilton, Montréal (Canada)
Schedule



Day 1 – Thursday, June 4

Details of Demos (Ovation Room 9:40-10:45) Facilitator: Claudia Harsch				
1. Assessing multilingual learners in kindergarten: incorporating age-appropriate design features into a high-stakes test Mark Derek Chapman, Gordon Blaine West, Marcy Olson, Beth DeVito, Christy Shepherd, Regina DiBella	2. Context-Aware Adaptive Assessment Model (CAAM): Bridging Global Frameworks and Local Realities in LLM-Driven Language Testing Kamaledine Tabine	3. Improving Proctoring Consistency in Remote English Language Assessment: A Feedback and Training Tool for Proctor Decision Making William Belzak, Yigal Attali	Cancelled 4. LLM-Simulated Student Reasoning Plausibility: A Novel Metric for Evaluating Distractor Quality in English Reading Comprehension Assessments Ruixue Liang, Han Bao, Yang Zhao, Shichen Hong, Lianzhen He	5. A Fully AI-mediated Learning-Oriented Assessment of Interactional Competence: A Demonstration from an Operational Test in Japan Anastasia Ramjag, Masaki Eguchi, Elliott Bruce, Mao Saeki, Hiroaki Takatsu, Max Thomson, Daniel Travin, Yuta Ishida, Sadahiro Yoshikawa, Yoichi Matsuyama

LTRC 2026 – Montréal
 6/2/2026 to 6/6/2026
 DoubleTree by Hilton, Montréal (Canada)
Schedule



Day 1 – Thursday, June 4
(10:50 a.m.-12:20 p.m.)

		Ovation Room - Facilitator: Jee Wha Dakin
10:50-12:20	WIPs	1. Localizing the “global” CEFR framework for a national assessment project in Cuba Yoan Martínez Márquez, Claudia Harsch
		2. From Humans to Chatbots: Emotional Mediation between Task Mode and Performance in L2 Speaking Assessment Lin Shi, Shuting Zhang, Yang Zhang
		3. Cheating is in the eye of the beholder: exploring views on academic integrity and cheating in contemporary language assessment Anna Soltyska
		4. Comparing BERT, LLM, and Human Ratings of Elementary Students’ Writing Strategies Angelie Ignacio, Hajung Kim, Bosco Zeng, Han Lai, Eunice Eunhee Jang
		5. Language Classroom Assessment and its Psychological Consequences: Examining ESL Learners’ Perceptions of Assessment, Mental Health, and Subjective Wellbeing Wei Yan, Glenda Christou, Peiyu Wang
		6. From Global Models to Local Uptake: Examining AI-Mediated Feedback and Instructor Mediation in Japanese Students’ Academic English Writing Johanathan Woodworth, Darlene Yamauchi
		7. L2 Test Takers’ Cognitive Engagement in Video-Based Listening Assessments: An EEG Study Shanshan He, Ruslan Suvorov
		8. Tracking L2 writing development with China’s Standards of English Language Ability: Application of longitudinal cognitive diagnosis models Hui Wang, Shangchao Min
		9. Comparing Dependency-Based Collocation Indices and LLM-Derived Surprisals and Semantic Similarity Measures in Predicting L2 Writing Scores Rui Zheng, Hang Li
		10. Examining the Reliability of LLMs Scoring for Integrated Writing Tasks Yijia Luo
		11. To what extent do computer-based speaking performances relate to interactive performances in group oral discussions? Yujia Zhou, Masashi Negishi, Asako Yoshitomi, James Carpenter
		12. Exploring Raters’ Cognitive Processes in Summary Content Evaluation: A Preliminary Think-Aloud Study Yutaka Ishii, Yasuyo Sawaki, Yumi Fujita
		13. Leveraging LLM to Identify Writing Strengths and Weaknesses Across Achievement Levels for Instructional Support Maggie Dunlop, Hyunah Kim, Zhimei Gu, Jennifer Hove
		14. Aligning the content and scores of a large-scale K-12 test to an international framework Meg Malone, Camelot Marshall, Reuben Vyn, Abbie Finnegan, Carly Favero, Aliza Price
		15. Contextualizing AI Ethics for Local Language Assessment: A Mixed-Methods Study of University Policies and Stakeholders’ Ethical Awareness in China Qi Wang, Yan Jin

LTRC 2026 – Montréal
 6/2/2026 to 6/6/2026
 DoubleTree by Hilton, Montréal (Canada)
Schedule



Day 1 – Thursday, June 4
(10:50 a.m.-2:20 p.m.)

		Ovation Room - Facilitator: Jee Wha Dakin
10:50-12:20	WIPs	16. Examining L2 Learners' Cognitive Processing During GenAI-Mediated Multimodal Speaking Tasks Sha Liu, Richard Spiby, Carolyn Westbrook 17. Globalizing Language Assessment: Designing an Integrated Performance Assessment for Montreal's Sociolinguistic Reality Fanny Mace 18. Developing a Partial Interpretation Argument for a Learning-Oriented, AI-Enhanced, Scenario-Based Language Assessment for Business Case Study Daniel Eskin 19. The Alignment Between Language Frameworks and Curriculum Standards: The Case of English Speaking Standards in China Shuai Zou, Yan Jin 20. Customising GPT for learning-oriented test preparation in second language writing Carrie Xin Peng, Xingcheng Wang 21. Validating a Standardized Mandarin Chinese Reading Assessment for Children as Heritage Language Speakers: A Cognitive and Learner-Centered Approach - Cancelled Chia Hsin Yin 22. Needs and Content Analysis for the Brigham Young University (BYU) English Language Admissions Test: Listening and Speaking Shantona Mir 23. How Do Interactional Patterns Differ in Human–AI and Human–Human Speaking Tasks? An Alignment Perspective Yang Zhang, Lin Shi, Shuting Zhang
12:20-1:20	Lunch (explore local options)	

		Opera AB	Symphonie 1	Symphonie 2	Symphonie 3
1:20-2:20	Professional sessions	Graduate Student Assembly (GSA)	ILTA Code of Ethics Discussion	ILTA Equity and Diversity Working Group Meeting	CALA Annual General Meeting (AGM)

LTRC 2026 – Montréal
 6/2/2026 to 6/6/2026
 DoubleTree by Hilton, Montréal (Canada)
Schedule

Day 1 – Thursday, June 4
(2:20 p.m.-4:00 p.m.)



		Opera AB	Symphonie 1	Symphonie 2	Symphonie 3	Symphonie 4	Ovation
2:20-2:50	Paper session 2.1	Investigating the implementation of automated writing scoring for adolescent multilingual learners in grades 6-12 Mark Derek Chapman, Katie Schultz, Ahyoung Alicia Kim, Kristin Erin Graw, Heather Elliott	Global Frameworks, Local Classrooms: Aligning the AAPPL with the CEFR Scale Jennifer Dobberfuhr Quinlan, Young-A Son	Self-reported strategic planning and avatar-character comprehension in a computerized-adaptive speaking assessment Sarah Goodwin	Exploring the link between linguistic features and psychometric difficulty in vocabulary assessment Ping-Lin Chuang, Ben Naismith, Allison Michalowski, Masha Kostromitina, Geoff LaFlair	Global AI-Scoring and Feedback, Local Fairness: TESOL Teachers' AI Literacy and Identity in Assessment Practices Johanathan Woodworth	Rethinking Item Calibration: Hybrid Approaches for Alternative Item Calibration Ying Zheng, David Booth
2:55-3:25	Paper session 2.2	Rating Scale Adaptation for an AI-mediated L2 Speaking Test Yasin Karatay, Jing Xu, Nicholas Glasson, Masaki Eguchi, Mao Saeki, Yoichi Matsuyama	From Global Validity to Local Gatekeeping: Language Proficiency in South Korea's Marriage-Immigrant Naturalization Interviews Jiyoon Lee	Eliciting Interactional Competence with Generative AI: A Mixed-Methods Study in a U.S. English Placement Test Context Inyoung Na	Comparing signaling features in the stimuli of an academic listening test and authentic lectures Zhi Li	Rubrics for classroom-based language assessment: Teachers' use of GenAI Meg Montee	
3:30-4:00	Coffee break (Opera C and Grand Foyer)						

LTRC 2026 – Montréal
 6/2/2026 to 6/6/2026
 DoubleTree by Hilton, Montréal (Canada)
Schedule



Day 1 – Thursday, June 4
(4:00 p.m.-5:40 p.m.)

		Opera AB	Symphonie 1	Symphonie 2	Symphonie 3	Symphonie 4	Ovation
4:00-4:30	Paper session 3.1	University staff's language assessment literacy development: cross-contextual comparisons Angela Gayton, Daniel MK Lam, Beverly Baker, Coral Yiwei Qin, Emma Bruce, Mojtaba Heydari	Evaluating Consequences in Dynamic/Diagnostic Language Assessment: An Integrated Framework for Macro-Micro and Global-Local Impact Yong-Won Lee	Introducing Behavioral Differential Item Functioning in Language Testing Sébastien Béland, Sandip Sinharay, Yikai Lu	Can Paired Tests Assess the Interactional Competence of Beginner Learners in EFL Classroom Settings? Kazuyo Kawamura	Contextualizing Automated Reading Assessment with Dynamic Human Mediation Melissa R Hunte, Eunice Eunhee Jang, Angelie Ignacio	Investigating the impact of calm music on L2 listening test anxiety through electrodermal activity (EDA) sensors and self-reports Ga Eul You, Ruslan Suvorov
4:35-5:05	Paper session 3.2	Putting measurement to the test. Assessing the construct validity of implicit and explicit measurement instruments for vulnerable learners. Emma Maes, Bart Deygers, Marieke Vanbuel	An ethnographic approach to domain analysis in LSP assessment design: Exploring local and global dimensions of English in a Turkish Tech Start-up Jennifer Kay Morris	Cancelled Exploring the Impact of Learning Journals as a Formative Assessment Tool on Chinese-English Major Students' Self-Regulated Listening Strategy Use Qiaozhen Yan, Shiyao He	AI Feedback for Speaking Assessment Sherry Hsin-Ying Li, Jessica Row-Whei Wu, Rachel Yi-Fen Wu, Berlin Chen, Fu-An Chao	Harnessing Generative AI to Develop L2 Reading Comprehension Tests: A Human-AI Comparison Ray J. T. Liao, GoMee Park, I-Chun Hsiao	Cancelled AI or No AI: Contextualizing Listening Assessment Across Languages and Proficiency Levels Ruth Maria Baptista, Jennifer Hansen, Matthew Wilcox
5:10-5:40	Paper session 3.3	How Raters with Different Backgrounds Judge West-African Test Takers' Video Interviews Kadidja Kone, Paula Winke	Bridging Global Assessments and Local Contexts: How Learner Background Shapes Young Learners' English Proficiency Mohammadreza Dalman, Okim Kang, Maria Nelly Gutierrez Arvizu, Francesca Grixoni	Supplementing calibration with linguistic and cognitive item features: lessons from large-scale English language tests Nadir Zanini, Mark Elliott, Yan Huang, Lina Loxley, Siyuan Tang	Cancelled Tailoring the Learning Journey: Unveiling Personalized Learning Paths in L2 Speaking through Cognitive Diagnosis Shuting Zhang	Design and Validation of a Language Placement Test: Insights from Class Performances, CEFR-Based Self-Assessment, and Test Perceptions Punchalee Wasanasomsithi	Investigating the cognitive validity of an integrated listen-to-summarize test: A mixed-methods analysis of test-taker notes Rebecca Yeager, GoMee Park, Junhee Park

LTRC 2026 – Montréal
 6/2/2026 to 6/6/2026
 DoubleTree by Hilton, Montréal (Canada)
Schedule



Day 2 – Friday, June 5

(7:30 a.m.-10:45 a.m.)

		Opera AB	Symphonie 1	Symphonie 2	Symphonie 3	Symphonie 4	Ovation
7:30-8:30		Registration and Breakfast (Opera C and Grand Foyer)					
All-day		Exhibitors: ACTFL, Canadian Association of Language Assessment, Duolingo, ETS TOEFL, Goethe Institut & Pearson (Opera C)					
8:30-9:30		The 2026 Alan Davies Lecture Award (Opera AB) Sponsored by the British Council The Local Social Life of Language Tests: Insights from Recent Qualitative and Mixed-Methods Studies Paula Winke					
9:40-10:10	Paper session 4.1	LLM surprisal as a meaningful indicator of writing proficiency in L2 French Guillaume Loignon, Angel Arias, Émilie Courteau	Predictive roles of nonverbal behavioral patterns for analytic scores in L2 speaking assessment Jiayi Wu, Shangchao Min, Zhuohan Hou	Investigating the Role of Rater Profiles in Applying Rating Rubrics for Second Language Comprehensibility Aki Tsunemoto, Rie Koizumi, Ryo Maie, Yo In'nami, Makoto Fukazawa, Mariko Abe	Does word class matter? Investigating the relationship between word frequency and item difficulty across parts of speech in L2 French vocabulary testing Ji-young Shin, Jeffrey Steele, Magda Tigchelaar	Applicability of Dynamic Assessment in EFL Classrooms: An Investigation of Teachers' Perceptions Reza Monfared	Evaluating a Simulation-Based German for Medical Purposes Test: A Many-Facet Rasch Model Analysis Laura Mair
10:15-10:45	Paper session 4.2	Cancelled Defining and Operationalizing GenAI Application Skills in Academic Writing: An AI-Mediated Performance Task Shuhui Yin	Meaningful Measures: How well can computer-based speaking assessments measure real-world communicative language proficiency? Anastasia Ulicheva, Ellen Barrow, Sarah R. Hughes	Multi-stage development of an L2 prosody scoring rubric for Chinese learners of English Yuanyue Hao	Derivational Morphology Gaps Among EFL Learners: Implications for Population-Specific Vocabulary Assessment Mastoor Al-Kaboody, Geoffrey G. Pinchbeck	Emerging Language Assessment Literacy in Preservice Teacher Practicum: Challenges and Insights from a Brazilian Context Vanessa Borges-Almeida	Reducing Avoidable Research Waste in AI-Driven Language Assessment: Is Radical Transparency the Panacea? Talia Isaacs, Hamish Chalmers

LTRC 2026 – Montréal
 6/2/2026 to 6/6/2026
 DoubleTree by Hilton, Montréal (Canada)
Schedule



Day 2 – Friday, June 5
(10:50 a.m.-1:25 p.m.)

		Opera AB	Symphonie 1	Symphonie 2	Symphonie 3	Symphonie 4	Ovation
10:50-11:20	Paper session 4.3	Diagnosing rater effects in writing assessment: Comparing three analytical frameworks Mark Elliott, Nadir Zanini, Madeleine Auste	When Emotion Meets Cognition: An Eye-Tracking Investigation in L2 Summary Writing Yang Zhang, Lin Shi, Lianzhen He	Uncovering Panelist Decision Strategies: Insights from Mapping CELPIP Test Scores to CEFR Standards You-Min Lin, Michelle Chen	Where measurement meets machine learning: Explaining item difficulty with examples from L2 vocabulary assessment Karen Dunn, Lucy Skidmore, Thomas Rogers	Participants as Co-Researchers: A Co-Analysis Approach to Language Assessment for Immigration Coral Yiwei Qin	A Meta-Analysis of Reported Agreement Between Large Language Models and Human Raters in Essay Scoring Hongli Li, Che Han Chen, Chiho Young-Johnson, Soyoung Lim, Yali Feng, Kevin Fan
11:20-12:20	Lunch (explore local options)						
12:20-1:25	Special Interest Groups (SIGs) meetings		Language assessment literacy (LAL SIG)	Language assessment for young learners (YL SIG)	Integrated assessment (IA SIG)	Automated language assessment (ALA SIG)	

LTRC 2026 – Montréal
 6/2/2026 to 6/6/2026
 DoubleTree by Hilton, Montréal (Canada)
Schedule



Day 2 – Friday, June 5
Symposia (1:30 p.m.-3:00 p.m.)

	Opera AB	Symphonie 1	Symphonie 2	Symphonie 3	Symphonie 4	Ovation
1:30-3:00	From Policy to Practice: Local and Global Co-operation Nahal Khabbazzbashi Discussant: Yasuyo Sawaki - Policy Formation and Implementation Yoshihiro Takizawa - Research and Development Barry O'Sullivan, Richard Spiby - Delivery and Rating Akihiro Matsuura - Performance and Impact Karen Dunn, Barry O'Sullivan	Are Humans Needed? The Use of Spoken Dialogue Systems in Assessing L2 Oral Communication Gary Ockey & Elvis Wagner Discussant: Elvis Wagner - Spoken Dialogue Systems and Potential Interactions with Test-Taker Personal Characteristics Gary Ockey - Conversational AI as Interlocutor Erik Voss, Hansun Zhang Waring - Revisiting the Socio-Cognitive Framework for AI-Mediated Oral Communication Assessment: Validity Considerations Evelina Galaczi, Yasin Karatay	Multilingual assessment - toward a principle of localization Saskia Van Viegen & Sunny Lau Discussant: Constant Leung - Localizing Assessment in Tanzania's Language-in-Education Policy Context: A Systematic Review Jamie Schissel, Donather Magabe - Exploring Multilingual Primary Students' Cross-Lingusitic Awareness in French-English Education Susan Ballinger, Beverly Baker, Claude Quevillon Lacasse - Creating multilingual assessment ecologies in Canadian classrooms Saskia Van Viegen, Sunny Lau - Translanguaging for Equity and Justice in Assessment: A Systematic Review Zhongfeng Tian, Chia-Hsin Yin, Jessica McConnell	Contextualizing Standard Setting and Alignment in Canada: Local Practices, Global Frameworks, and Policy Relevance Angel Arias & Eunice Eunhee Jang Discussant: Stephen Sireci - Landscape of Canadian Standard Setting and Alignment Research and Practice Eunice Eunhee Jang - A Triangulated Approach to Standard Setting for Language-Specific Purposes (LSP) Maryam Wagner - Renewing Language Proficiency Standards for Teacher Certification in Canada: A Multi-Phase Standard-Setting Initiative Angel Arias - Aligning the Canadian Language Benchmarks with the AI-enhanced, adaptive Duolingo English Test using Item-Descriptor Matching Jeanne Sinclair	Reframing Language Assessment Literacy in Latin America: The Multiple Cases of Brazil, Colombia and Mexico Elsa Fernanda Gonzalez Discussant: Gladys Quevedo-Camargo - Presenting a Brazilian Language Assessment Literacy Course: Syllabus, Pre-Service Teachers' Perceptions and Short-Term Impact Isadora Teixeira Moraes - Developing Language Assessment Literacy in Early English Education: Insights from Latin America Juliana Reichert Assunção Tonelli - Links between teaching and assessment in context: a workshop to develop in-service teachers' (L)AL Sonia Patricia Hernández-Ocampo - Mexican young learner EFL teachers and training instructors' perspectives: How useful is online LA training? Elsa Fernanda Gonzalez	Connecting Global Assessment Systems and Local Practices: Trust, Affect, and AI in Language Education Canadian Association of Language Assessment Discussant: Samira ElAtia - Beyond Language Proficiency Scores: Exploring the Socio-Emotional Learning Dimension of Study Abroad Suzanne Springer, Martyna Kozłowska, Jaime Demperio - Reframing Teacher Identity and Relationality in the AI Era: A Mixed-Methods Study of TESL Educators' Role Perceptions Johanathan Woodworth - Taking a positive turn for the affective dimension of the Learning-Oriented Assessment Framework: "These things give us courage" Christine Doe, Krista Ritchie - Responsive Human Mediation and Higher-Order Language Skills Development: A Bayesian Time-Series Case Study Melissa R. Hunte, Nanda Warren

LTRC 2026 – Montréal
6/2/2026 to 6/6/2026
DoubleTree by Hilton, Montréal (Canada)
Schedule

Day 2 – Friday, June 5
(3:00 p.m.-4:35 p.m.)



		Opera AB	Symphonie 1	Symphonie 2	Symphonie 3	Symphonie 4	Ovation
3:00-3:30	Coffee break (Opera C and Grand Foyer) and Group photo (Festival Terrasse; Opera AB in case of rain)						
3:30-4:00	Paper session 5.1	Diagnosing Language Ability with China's Standards of English Language Ability: A Cross-sectional Comparison across Educational Stages Hui Wang, Shangchao Min	Cancelled Language Assessment Literacy among Colloquial Arabic Instructors: Convergence and Divergence between Local Practices and Global Assessments Hadeel Arqbi	Text Ease and Readability Assessment of Human-versus ChatGPT-Generated Listen-to-Summarize Conversation Scripts Lixia Cheng, Rodrigo A. Rodríguez-Fuentes	Global Standards, Local Practices: Rethinking Note-Taking Modes in Language Testing Jieun Kim	Language Assessment Literacy In Practice: EFL Teachers' Implementation of Classroom-Based Assessments in Chile Evelyn Diaz-Iturriaga	Cancelled Developing a rating scale for a large language model driven spoken dialogue system delivered paired discussion tasks Zhuohan Hou, Shangchao Min
4:05-4:35	Paper session 5.2	Validating the Elicited Imitation Task through Eye-Tracking Evidence Yulin Pan, Huiying Cai, Amanda Li, Angie Lee	Ethical Dimension in the Language Assessment Practices of University Language Teachers in Mexico Hugo A. Andrade-Mayer, Virna Velazquez, Georgia Grondin	Which AI should we use? Using corpus linguistics to compare and evaluate ChatGPT- and Claude-generated texts for TOEFL iBT's Writing for an Academic Discussion task Che-Han Chen	Exploring the relationship between self-assessment and objective measures of reading and listening abilities among young Spanish learners of English Renka Ohta, Ching-Ni Hsieh	Cancelled From interactional data to measurement modelling: Investigating cognitive processing of L2 pragmatic competence in the design of a UK pre-sessional speaking test Shishi Zhang	Can an LLM-powered spoken dialogue system accommodate L2 learners of varying proficiency levels? Jing Xu, Masaki Eguchi, Yasin Karatay, Mao Saeki, Nicholas Glasson, Anastasia Ramjag, Annabelle Pinnington, Elliott Bruce, Evelina Galaczi, Max Thomson, Martin Robinson, Yoichi Matsuyama

LTRC 2026 – Montréal
 6/2/2026 to 6/6/2026
 DoubleTree by Hilton, Montréal (Canada)
Schedule



Day 2 – Friday, June 5
(4:40 p.m.-9:00 p.m.)

		Opera AB	Symphonie 1	Symphonie 2	Symphonie 3	Symphonie 4	Ovation
4:40-5:10	Paper session 5.3	When Tests Meet Reality: Cognitive and Context Validity of the PTE Core for Canadian Immigration and Employment Peiyu Wang, Liying Cheng	English Proficiency Test Exemption Policies in US Research-Intensive Universities Nicholas Coney, Daniel Isbell	Surprisal as an Indicator of Linguistic Accuracy in L2 Writing Xuan Hu	Cancelled Beyond the 'Benefit of the Doubt': An Eye-Tracking Study of Assessment-modality Effects on Rater Behaviour and Cognition Daniel Yu Sheng Chang	Contextualising constructs: Applying Systemic Functional Linguistics to language assessment design and validation Mark Elliott, Hye-won Lee	Revisiting the Social Dimension of Foreign Language Tests: The Case of English Gaokao from 1950-1966 in China Jinjie Wu, Mingwei Pan, Xini Liao
5:15-5:45	Paper session 5.4	Bridging Proficiency and Pedagogy: Rethinking the validity of Chinese Placement Tests in Canada Shujiao Wang, Zhi Li, Min Sun	Building a Policy Literacy Survey Laura Schildt, Bart Deygers	Modeling Human-AI Collaboration Dynamics in EFL Writing: Hidden Markov Model and Process Mining Approach Shasha Xu	Binary and CEFR Classification Consistency of the Duolingo English Test Ramsey Lee Cardwell, Allison Michalowski	"Brevity is the soul of wit" but why is it untrue in writing assessment? Bosco Zeng, Angelie Ignacio, Hajung Kim	Picture Prompt Design and Interactional Competence in Paired Speaking Assessment: Evidence from a Low-Stakes Online English Speaking Test in China Miao Gao
6:45-9:00	Networking dinners (meet at Opera Grand Foyer at 6:45 near registration desk to meet your host, if you signed up)						

LTRC 2026 – Montréal
6/2/2026 to 6/6/2026
DoubleTree by Hilton, Montréal (Canada)
Schedule



Day 3 – Saturday, June 6
(7:30 a.m.-10:45 a.m.)

		Opera AB	Symphonie 1	Symphonie 2	Symphonie 3	Symphonie 4	Ovation
7:30-8:30	Registration and Breakfast (Opera C and Grand Foyer) Sponsored by the Language Training and Testing Centre						
All-day	Exhibitors: ACTFL, Canadian Association of Language Assessment, Duolingo, ETS TOEFL, Goethe Institut & Pearson (Opera C)						
8:30-9:30	2026 Cambridge/ILTA Distinguished Achievement Award (Opera AB) ... and no Bob, the answer isn't blowin' in the wind! Barry O'Sullivan						
9:40-10:10	Paper session 6.1	Cancelled Assessing metapragmatic awareness for intercultural communication: Test localisation targeting UK pre-sessional students Shishi Zhang	Test Taker Perceptions of Fairness, Validity, and Impact in a Local Standardized English Proficiency Test I-Chun Vera Hsiao	Attention Allocation in L2 Summary Writing: Insights from Eye-Movement Data Lin Shi, Yang Zhang, Lianzhen He	Global to Local: Examining the Predictive Validity of the DET Interactive Listening Tasks Noriko Iwashita, Hong Giang Nguyen, Anh Nguyen, Van-Trao Nguyen, Trung Quang Doan, Hanh Hong Kieu	Does the Rating Make the Score? Comparing how Human Ratings affect Automated Essay Scoring in a Large-Scale Dutch Writing Assessment Joni Kruijsbergen, Marieke Vanbuel, Amaury Van Parys, Bart Deygers, Orphée De Clercq	Invisible Assessment & Transparent Evidence with LACES Olivia Seow
10:15-10:45	Paper session 6.2	A mixed methods study of rater judgements of pronunciation in a paired discussion task Liberato Silva dos Santos	Prompt design based on communication strategies to facilitate spoken interaction between generative AI and EFL learners, and validation of its effectiveness Yutaka Yamauchi	Effects of task type on the variability pattern of syntactic complexity in second language writing Minh Nhut Nguyen	Global Standards, Local Impact: Educational and Social Consequences of EXAVER, Mexico's English Certification Marion Alain Meunier Colula	A Language-for-Specific-Purposes Approach to University Admission: Meeting Local Needs through Needs Analysis and Test Innovation Ashlee Farnsworth, Shantu Mir, Troy L. Cox	Response Patterns of Native and Non-Native English Speakers on PIAAC Literacy Items: A Latent Profile Analysis Hongli Li, Boshi Wang, Natalie Pruitt

LTRC 2026 – Montréal
 6/2/20206 to 6/6/2026
 DoubleTree by Hilton, Montréal (Canada)
Schedule



Day 3 – Saturday, June 6
(10:50 a.m.-1:35 p.m.)

		Opera AB	Symphonie 1	Symphonie 2	Symphonie 3	Symphonie 4	Ovation
10:50-11:20	Paper session 6.3	Cancelled Human Judgment or Algorithmic Ease? Score Comparability and Student Perceptions in Adolescent EFL Writing Assessment Chia-Hsin Yin	Supporting the Validity Argument of a High-Stakes Language Assessment Through Evidence of Test-Taker Identity Verification Ramsey Lee Cardwell, William Belzak, Manqian Liao, Geoffrey T. LaFlair	Cancelled Does Input Modality Affect L2 Integrated Writing Consistently? Examining the Moderating Effect of English Proficiency, Modality Preference, and Visual-spatial Working Memory Jiayi Dong	From Global Standards to Local Realities: Rethinking Vocabulary Testing for Chinese and Arabic Learners. Geoffrey Pinchbeck, Mastoor Al Kaboody	Assessing a Localized Construct of Writing Ability: A Validity Argument for an NGSS-Aligned Scenario-Based Language Assessment Soo Hyoung Joo	Yup'ik Students' Voices: Alaska Native Language (Yugtun) assessment Catherine Moses, Rosalie Lincoln, Sally Samson, Rosalie Grant
11:20-12:25	ILTA Annual Business Meeting (lunch included for first 100 participants – Ovation Room)						
12:30-1:00	Paper session 7.1	A Framework for Validating Hybrid-Marking Models of L2 Speech & Writing Mark Brenchley, Trevor Breakspear, Yan Huang	Local writing placement tests and generative AI assistance: A domain analysis of an evolving academic construct Rebecca Yeager, Rurik Tywoniw, Ha Ram Kim, Melissa Meisterheim, Shireen Baghestani, Jim Ranalli, Sara Cushing	Test-taker agency in writing assessment: Balancing universal proficiency constructs with context-sensitive design through topic choice Hye-won Lee, Carla Pastorino-Campos	Can you prototype in the field?: Revisiting the tripartite model of preoperational testing Yena Park, Xiaowan Zhang	Bridging Policy and Practice: A multi-method evaluation of a high-stakes language test's fitness-for-purpose in the Canadian migration context Ellen Barrow, Sarah Hughes	Exploring the potential of AI-generated reformulation and model texts in a resource-constrained assessment context: An eye tracking and stimulated recall study Atta Gebril, Khaled ElEbyary, Ramy Shabara
1:05-1:35	Paper session 7.2	Qualitative Patterns of Change in Argument Development in a Test of L2 Argumentative Speaking Ability Jorge Luis Beltran Zuniga	Bridging Traditional and AI-Assisted Writing Constructs in the Chinese EFL Context Jianda Liu, Xuan Zhou	Assessing the Impact of Note-Taking Methods on Specialized Vocabulary Knowledge and Lecture Comprehension in EMI Graduate Classroom Nataliya Borkovska, Pelin Irgin	Failing Forward in Local Contexts: Marginal Gains and the Infinite Game of Language Learning Troy L. Cox	Context Matters: Proctor Language and Procedural Equity in Remote English Testing Jieun Kim	Revision and learning transfer effects of actionable high-level suggestions on writing drafts Yigal Attali, Kai-Ling Lo, Yena Park, Jacqueline Church, Andrew Runge

LTRC 2026 – Montréal
 6/2/2026 to 6/6/2026
 DoubleTree by Hilton, Montréal (Canada)
Schedule



Day 3 – Saturday, June 6
Symposia (1:40 p.m.-3:10 p.m.)

	Opera AB	Symphonie 1	Symphonie 2	Symphonie 3	Ovation
1:40-3:10	Situating Language Proficiencies: Assessment From An Ecological Linguaging Perspective Constant Leung, Eunice Eunhee Jang & Angel M. Y. Lin - Language Proficiency – From the Global to the Situated Constant Leung - Ecological Linguaging Competencies (ELC) – Localizing Assessment Angel M.Y. Lin, Paul J. Thibault - Glocalizing Language Assessment: Redesigning EAP Speaking Rubrics Through The Ecological Linguaging Competencies (ELC) Lens Peichang He, Paul J. Thibault, Wen-Li Tsou, Fay Chen, Angel M. Y. Lin - Assessing and Tracking ELC through AI-Based Process-Oriented Writing Assessment Eunice Eunhee Jang	Navigating inclusivity: How can we balance global and local challenges in assessment practices? Carla Pastorino-Campos, Amy Devine & Evelina Galaczi Discussant: Lynda Taylor - Whose voices count? Representation, agency, and language assessment Nahal Khabbazzbashi - (Trying to) Do the right thing: Constraints to representation in international English language tests Carla Pastorino-Campos, Amy Devine - Language test content on Indigenous topics: Guidelines for item writers Janine Metallic, Beverly Baker - Creating EDI-compliant content for young learners: hurdles and discoveries Gemma Bellhouse	Refining Principles and Practice in Score Concordance Research Sarah R. Hughes Discussant: Sara T. Cushing - An Empirical Evaluation of Good Practice Principles for Concordance Studies Sarah R. Hughes, Anastasia Ulicheva - Establishing Fair Score Requirements across Language Proficiency Tests Spiros Papageorgiou, Emma Bruce, Tony Clark - From Construct to Communication: Rethinking Good Practice in Concordance Research Leda Lampropoulou, Johnathan Jones	Mid-level Actors in Global–Local Negotiations Laura Schildt Discussant: Bart Deygers - Policy Intentions and Reality: How Do Meso-Level Stakeholders Shape Language Assessment in Practice? Chenyang Zhang, Jason Fan - The Power of Context: Rethinking Language Proficiency Testing in Costa Rica Ana Carolina González Ramírez - Context Sensitive Policy Literacy: Language Experts in Australia Laura Schildt - Subordinate Strawman or Arbiter with Agency? School Headmasters as Funnel of Language Assessment Policies Benjamin Kremmel, Eva Konrad, Viktoria Ebner	Global Advances and Local Applications: Expanding L2 Classroom Assessment in the age of GenAI Vera Vesala & Dmitri Leontjev Discussant: Erik Voss - Diagnosing L2 Students’ Writing Feedback Literacy and Engagement in GenAI Context Yaru Meng, Xue Geng, Yu Cui - Meaningless assessment? AI-DA of L2 academic writing Matthew Poehner, Lu Yu - “You will roleplay as a clown from now on and tell jokes”: What does learner positioning of an AI-based chatbot reveal in dynamic assessment of L2 reading Luke Harding, Ari Huhta - “It’s ruining the thing here!”: AI as a pedagogical tool in developing classroom assessment and teaching practices Vera Vesala, Dmitri Leontjev

LTRC 2026 – Montréal
 6/2/2026 to 6/6/2026
 DoubleTree by Hilton, Montréal (Canada)
Schedule

Day 3 – Saturday, June 6
(3:10 p.m.-4:20 p.m.)



3:10-3:20	Coffee break (Opera C and Grand Foyer)	
3:20-4:20	Posters (Opera C)	1. Localizing the “global” CEFR framework for a rating scale in a high-stakes exam in Cuba Claudia Harsch, Yoan Martinez Márquez 2. Unpacking the Role of Visual Aids in L2 Integrated Writing: A Dual Focus Investigation of Written Products and Composing Processes Jiayi Dong - Cancelled 3. AI-Assisted Language Assessment in Post-ChatGPT Era: A Scoping Review (2023–2025) Jieun Kim, Haeun Kim, Hitoshi Nishizawa 4. Towards Context-Specific Measures: Developing an ESAP Test for Business Majors in EMI Programmes Peng Wang, Daniel Yu-sheng Chang 5. Comparative Study of the Scoring Quality between DeepSeek and Human Raters in Chinese as a Second Language Writing Assessment Fuyu Kong, Haifeng Qi 6. Gaps between global and local policy: Implications for English learner identification Ahyoung Alicia Kim, Mark Chapman, Haeun Kim, Jeanne Beck, Pauline Ho 7. Global Standards, Local Profiles: Implementing the Updated Vocabulary Level Test for Diagnostic Assessment of Thai Graduate Students Sumanee Pinweha, Sutthirak Sapsirin 8. Formative Assessment in Practice: Exploring How Chinese Teachers Support Learning in Elementary ESL Classroom Liying Cheng, Peiyu Wang, Liyan Huang 9. Commonalities and Localized Differences in How Language Testers Teach Language Assessment around the World Benjamin Kremmel, Paula Winke 10. Agile for Equity: Leveraging Agile Project Management Methodologies to Advance Fairness in Local Language Assessment Sharareh Taghizadeh Vahed 11. Assessing second language pragmatic competence for intercultural communication: The case of pre-sessional students in UK higher education Shishi Zhang - Cancelled 12. Trends, Methods, and Practices in Rater Training: A Scoping Review of Rater Training Research in L2 Performance Testing Ryuhei Mizoguchi, Chiho Young-Johnson, Yumi Fujita, Dylan Burton 13. Listening and Reading to Write: Corpus-Based Evidence on Task Type and Proficiency in Portuguese as an Additional Language Elisa Marchioro Stumpf, Juliana Roquele Schoffen, Isadora Dahmer Hanauer, Luiza Sarmiento Divino 14. From Global Standards to Local Policy: Contextualizing GEPT and BESTEP in Taiwan’s Language Assessment Landscape Jessica Row Whei Wu, Rachel Yifen Wu, Anita Chun-Wen Lin, Anne Liu

LTRC 2026 – Montréal
 6/2/2026 to 6/6/2026
 DoubleTree by Hilton, Montréal (Canada)
Schedule

Day 3 – Saturday, June 6
(3:20 p.m.-4:20 p.m.)



3:20-4:20	Posters (Opera C)	15. AI literacy as part of student teachers' Language Assessment Literacy - a university teacher students' perspective on the use of generative AI tools for writing assessments in EMI contexts Karin Vogt, Alexander Mayer
		16. Local practices for developing language assessment literacy (LAL) in Brazil: a comparison between LAL courses in three public universities Isadora Teixeira Moraes, Estogildo Gledson Batista, Renan William Silva de Deus
		17. The Language Assessment Accessibility Resource Centre (LaARC): Supporting Inclusive English Assessment in Latin America Juliana Reichert Assunção Tonelli, Gladys Quevedo Camargo
		18. Automated Topic Detection in Spoken Language Assessment: Comparing Human-Labeled Training and Prompt-Based Category Prediction Young-A Son, Erik Voss, Kim Sallee, Meg Malone
		19. Comparing Input Prompt Visuals Using Virtual Reality in L2 English Speaking Task Yilin Zhang, Erik Voss
		20. Integrating Collocations into Predicting L2 Oral Proficiency: Evidence from the TEM8 Speaking Test in China Jinjie Wu, Xini Liao
		21. Bridging global standards and local needs: a conceptual reflection from the Italian scenario-based assessment case study Giulia Peri, Sabrina Machetti
		22. Bridging Local Curriculum and Large-Scale Testing: Evaluating AI-Generated MCQs from Textbook Passages in Preparation for Large-Scale Assessment Jean-Young Chun, Jameson Moore - Cancelled
		23. Flexibility Without Compromise: Adapting the Delphi Method to Language Assessment Studies Birgitte Grande, Clayton David Leishman
		24. The development of an audio-visual listening test for the Canadian Armed Forces Military Training and Cooperation Programme Nancy Powers, May Tan
		25. Bridging Cognitive Evidence and Test Validity: A Scoping Review of Biometric Research in L2 Listening Assessment Shanshan He
		26. Bridging Global and Local: Collaborative Reflections on Developing Language Assessment Literacy in Canadian TESOL Teacher Education Shanshan He, Chuan Liu
		27. The Development and Validation of an Innovative, Fully-automated Interactive Speaking Task and its Implications for Contextualized Assessments Yena Park, Yigal Attali, Xiaowan Zhang, Jacqueline Church, Kai-Ling Lo, Andrew Runge, Geoff LaFlair
		28. Writing and Identity Development through Bilingual Writing Portfolios in a Dual Language Bilingual Program in Texas GoMee Park, Jennifer Burden

29. Investigating writing behaviours, neuro-cognitive processes, and their relationships to test scores in two integrated writing formats
 Andrea Revesz, Hyeonjeong Jeong, Sathena Chan, Haining Cui, Atsushi Doi, Dominique Jiang, Sewon Kim, Nicola Latimer, Motoaki Sugiura

30. The Impact of Exposure to Global Englishes Accents on Chinese Children's Learning of EFL (English as A Foreign Language)
 Zhuohan Chen

LTRC 2026 – Montréal
 6/2/2026 to 6/6/2026
 DoubleTree by Hilton, Montréal (Canada)
Schedule

Day 3 – Saturday, June 6
(4:25 p.m.-10:00 p.m.)



		Opera AB	Symphonie 1	Symphonie 2	Symphonie 3	Symphonie 4	Ovation
4:25-4:55	Paper session 8.1	Can a process-oriented approach to listening assessment and instruction meet the needs of standards-based L2 listening assessment? Jianling Liao	The Role of Linguistic Knowledge and Input Text Modality in an Integrated Reading-to-Speaking Task Shungo Suzuki, Judit Kormos	Picture-based narrative writing tasks in language assessments for young learners: Effects of task design on test performance and implications for validity Haeun {Hannah} Kim	Insights from Primary Reading Assessments in Southeast Asia: The Tiger and the Tapir. Louise Courtney	AI-Facilitated Evaluation of Chinese as a Second Language Integrated Speaking Performance: Measurement Features, Human Agreement, and Criterion-Relatedness Xiaozhu Wang	Cognitive Diagnostic Modelling (CDM) for Construct Validity Evidence in e-T&P Reading Comprehension Tasks: A Mixed-Methods Approach to Test Validation Fahri Yilmaz, Secil Alaca, Merve Selçuk, Zeynep Urkun
5:00-5:30	Paper session 8.2	Accent Effects of Text-to-speech Voices in Listening Assessment Xiaowan Zhang, Jacqueline Church, Andrew Runge, Geoff LaFlair	Systematizing local concerns about a large-scale, standardized speaking test for young language learners David MacGregor, Mark Chapman, Fabiana MacMillan, Meg Montee, Burak Senel	Human versus Machine Rater: Investigating the Influence of Response Length in Essay Scoring Dellannia Segreti, Jia Guo, Athena Tassis	Transparent to Whom? Localizing the Principles of Language Assessment Transparency Nathaniel Owen, Ananda Senel	Test impact on teachers' wellbeing—a case study of NMET in China Xiao Cheng, Xunyi Pan, Feier Ji	Modeling Classical Item Difficulty in a Medical English Test Using Semantic Embeddings, Retrieval–Entailment Features, and Answer-Choice Interactions Peter Kim
6:30-10:00	Banquet at the Île de France – Le 9^e restaurant (1500 Boulevard Robert-Bourassa, Suite 900) <i>Ticket-holders only</i>						

The 2026 Samuel J. Messick Memorial Lecture Award, Sponsored by ETS

Day 1 – Thursday, June 4

Time: 8:30-9:30 a.m.

Location: Opera AB

Dr. Christopher DeLuca, Queen's University, Canada.

Title: "From Literacy to Legitimacy: Validity in the Assessment Literacy Ecosystem."



By shifting validity from a purely technical concern to one that encompasses social consequences and value implications, Messick laid the foundation for the central challenge now before us: the legitimacy of assessment when learning is increasingly mediated, if not obscured, by artificial intelligence. Situated at the heart of this challenge, educators must navigate a way forward that balances academic integrity while embracing new forms of learning, assessment, and digital resources. What conception and practice of assessment literacy will guide educators in navigating a productive future for assessment? In this keynote talk, I call for an expanded conception of assessment literacy that pushes beyond the application of established methods or technical proficiency and instead equips educators and assessment leaders with the capacity to adapt to uncertainty through reflexivity, learning, and professional judgment. Drawing on a multi-jurisdictional study of teachers' assessment literacy across Australia, Canada, England, and New Zealand, I identify four interrelated dimensions—epistemological, ethical, experiential, and embodied—that enable educators to confront the future of assessment and re-examine what counts as validity today. This keynote invites educators and assessment leaders to rethink assessment not only as a set of practices, but as set of interconnected capacities nested within a broader educational ecosystem. In doing so, it positions assessment literacy as a cornerstone of educational resilience and innovation in response to a rapidly changing world.

Dr. Christopher DeLuca is a Professor of Educational Assessment at Queen's University, where he serves as Director of the Queen's Assessment and Evaluation Group. His research examines the alignment of assessment, curriculum, and pedagogy through socio-cultural frameworks, with a focus on supporting teachers in navigating these interconnected areas of practice to enhance student learning for all. Dr. DeLuca's work has been recognized with numerous awards, including the AERA Outstanding Paper in Classroom Assessment Award, the R.W. Jackson Award, and the National Mentorship Award. He has served as Chair of the AERA's Classroom Assessment SIG, President of the Canadian Educational Researchers' Association, and Editor of the Canadian Journal of Education. Dr. DeLuca is currently Lead Editor of *Assessment in Education: Principles, Policy and Practice*.

The 2026 Alan Davies Lecture Award, Sponsored by the British Council

Day 2 – Friday, June 5

Time: 8:30-9:30 a.m.

Location: Opera AB

Professor Paula Winke, Michigan State University.

Title: "The Local Social Life of Language Tests: Insights from Recent Qualitative and Mixed-Methods Studies."



Alan Davies stressed the uncertainty in language testing. He showed us how a test's cut scores, criteria, and proficiency definitions are not absolute, but rather subject to interpretation. To set them appropriately, researchers must pay attention to a test's stakeholders and to society at large. Meanwhile, language testing research has been overwhelmingly quantitative. That research has prioritized controlled, structured data collection designs, large samples of numerical data, and robust statistical analyses to produce generalizable results. Qualitative studies are comparatively rare. That is unfortunate, because by examining individual test stakeholders, qualitative research can uncover the uncertainties stakeholders face. Indeed, qualitative studies use uncertainty as the starting point. Exploratory, qualitative research is based on constructivism, which assumes there is no single truth and that people construct meaning through their personal experiences and local interactions. In this talk, I will look with you at our field today as I imagine Alan Davies might have viewed it, using a lens that is not numerical, but social. With research assistants at Michigan State University, I sorted the last seven years of research articles published across five major journals in language testing (Assessing Writing, Language Assessment Quarterly, Language Testing, Language Testing in Asia, Studies in Language Testing) into quantitative, qualitative, and mixed-methods research bins. My assistants and I uploaded the qualitative and mixed-methods articles to MAXQDA and coded them to explore themes. I will discuss the major trends in qualitative and mixed methods research within our field. At the end, I will summarize how qualitative research can help all language-testing research programs, including large-scale and global ones, be more attentive to the contexts and needs of local stakeholders.

Paula Winke is a professor at Michigan State University, USA. She investigates the experiences of test takers and designs test-creation toolkits for teachers. She was the co-editor of the journal "Language Testing" 2019 through 2024. She has won "Best Article" awards from the International Language Testing Association (in 2025), the American Association of Applied Linguistics (in 2020), and the CALICO Journal (in 2009). She has additionally received the Paul Pimsleur Award for Research in World Language Education (in 2020) and the International TESOL Award for Distinguished Research (in 2012). Paula is currently co-editing the Applied Linguistic Press open-access book series "Syllabi in Applied Linguistics" with Benjamin Kremmel. Outside of language testing, Paula likes to run, play tennis, and hang out with her family.

The 2026 Cambridge/ILTA Distinguished Achievement Award

Day 3 – Saturday, June 6

Time: 8:30-9:30 a.m.

Location: Opera AB

Professor Barry O’Sullivan.

Title: "... and no Bob, the answer isn't blowin' in the wind!"



CAMBRIDGE



ILTA
INTERNATIONAL LANGUAGE
TESTING ASSOCIATION

I’ve had a pretty good innings over the past five decades in the world of life and work and in this talk, I will reflect on the long and winding journey that has taken me from the Christian Brothers in Thurles, County Tipperary to the DAA Lecture in Montreal in 2026. I will also meander through the equally winding journey that took me from an English language classroom in Peru to the corridors of language testing and assessment at the British Council. In doing this, I will discuss the experiences and influences that have shaped my thinking around test policy, design, and use. I will also share my thoughts on where language testing sits currently and where it might be going, highlighting what I see as the opportunities and risks associated with the increasingly ubiquitous usage of artificial intelligence. Just like everything else in testing and assessment, whatever the question or issue highlighted, the answer always comes down to validity and validation in the end. With this in mind, I will re-visit the work I was fortunate enough to have done with my friends Cyril Weir and Micheline Chalhoub-Deville focusing on how a validity model which more fully reflects the context-of-use can be seen as offering a solution to the current position where validation is essentially a mythological entity we all talk about but don’t really engage with – a bit like faeries or leprechauns. This requires us to broaden our thinking around stakeholders and the associated issues of impact/washback and assessment and policy literacy.

At the end of my last plenary to the LTRC (in Seoul, 2015) I asked that we put the language back into language testing. In Montreal, I will ask for something quite different, but related (I think). You’ll hear what that is on the day...

In over three decades of work on language testing Barry O’Sullivan has written extensively on the subject and presented his work at conferences around the world on hundreds of occasions. His work includes the British Council’s Aptis test, the socio-cognitive model of test development and validation and the Comprehensive Learning System. He currently leads a major project at the British Council exploring the application of AI and machine learning to teaching, learning and assessment. He was granted fellowship of the UK’s Academy of Social Sciences (2016); elected fellow of the Asian Association for Language Assessment (2017); and awarded an OBE by the UK government (2019). In addition, he received the e-Assessment Association’s Lifetime Contribution Award (2023), and the English Language Testing Society’s inaugural Lifetime Achievement Award (2025).

Pre-Conference Workshops

Workshop A: Assessing Languages Beyond English – Global Principles and Local Practices

Day: Tuesday, June 2 (9 a.m. to 4 p.m.)

Location: Symphonie 3

Sonja Zimmermann & Katharina Karges



The field of language testing and assessment has largely been shaped by research, theory, and practice conducted for English as a first, second or foreign language. Yet assessment in Languages Other Than English (LOTEs), explicitly including sign language, is common in many parts of the world and takes place within diverse linguistic, educational, and sociocultural contexts. Exchange of knowledge and experience across languages and with the English-speaking language assessment research community, are rare, however.

This one-day workshop brings together language assessment professionals who work on or are interested in the assessment of LOTE to explore how global frameworks and concepts intersect with local practices, challenges, and innovations. Participants will engage in structured discussions and collaborative activities to share experiences, identify common challenges, and highlight the potential of local practices to contribute to global understanding. Through group work and plenary exchange, the workshop will examine how global standards such as validation models or the CEFR are interpreted and adapted in different settings, what barriers exist for visibility and transferability, and how local solutions can inspire broader perspectives.

The workshop aims to foster mutual learning and professional exchange and encourage participants to see assessment in LOTE as a dynamic field that connects global principles with local realities. Outcomes include shared insights and practical reflections about language assessment in general and language assessment in LOTE in particular, as well as stronger professional connections across linguistic and regional boundaries.

Sonja Zimmermann is the Head of Product Development, with a focus on testing at g.a.s.t. (Gesellschaft für akademische Studienvorbereitung und Testentwicklung). She has many years of experience in language assessment development and validation projects, mainly for productive skills. Sonja holds a PhD in Research into Language Learning and Teaching from the University of Bremen with a focus in integrated writing assessment and is an active member of the academic community. She regularly reviews articles for leading journals in language assessment and presents at national and international conferences.

Katharina Karges is a Research Associate at the Herder Institute of the University of Leipzig, specialising in foreign language assessment, quantitative methodologies and foreign language teaching research. She also collaborates on research projects in German as a Second Language as a data analyst. She has recently finished her PhD at the University of Fribourg/Switzerland, where she analysed item difficulty in a multilingual reading comprehension assessment.

Workshop B: Using LLMs for Content Generation and Evaluation in Assessment Development

Day: Tuesday, June 2 (9 a.m. to 4 p.m.)

Location: Symphonie 2

Andrew Runge & Yena Park



There is a growing interest in applying large language models (LLMs) to the task of automatically generating test content and questions for both research and practical applications. While popular applications such as ChatGPT can provide an easy way to interact with an LLM to produce small numbers of sample questions, creating and evaluating large amounts of test content in this way can be extremely time-consuming, difficult to scale, and limited in reproducibility. This workshop provides a hands-on demonstration of how to create repeatable and scalable test content creation processes using LLMs. Participants will use Google Colab and the OpenAI API to complete several hands-on activities that will demonstrate how to create test content such as reading passages and multiple choice questions (e.g., keys and distractors) for different assessment purposes and tasks using a multi-step approach to generating test content.

Participants will additionally learn about the strengths and weaknesses of different prompting strategies, as well as manual and automated approaches to prompt creation that can support the development of robust, automated content filtering and evaluation tools that can help ensure that generated test content meets the standards for assessment.

Experience with coding, LLMs, or assessment development is not necessary. While some activities will require making small changes to the provided Python code, no prior Python or coding experience is required.

Andrew Runge is an AI Research Engineer at Duolingo. He works in test development for the Duolingo English Test, where he specializes in automatic item generation. He holds a master's in Language Technologies from Carnegie Mellon University.

Yena Park is an assessment scientist on the test development team for the Duolingo English Test, where she specializes in using technology for test development. She holds a PhD in second language studies specializing in language assessment.

Workshop C: Using GLMMs to Model Rater-mediated Assessment Outcomes

Day: Wednesday, June 3 (9 a.m. to 4 p.m.)

Location: Symphonie 2

Karen Dunn & Dylan Burton



Generalized linear mixed effects models (GLMMs) are a family of regression models increasingly used in the social and behavioral sciences, including applied linguistics, to investigate the influences of context, person and item characteristics, and bias on measured outcomes. GLMMs can be used for a wide range of applications to model the degree and impact of particular variables, such as rater training and experience, lexical features of items, participant background variables, and environmental factors that affect ratings of language proficiency. This type of modeling can provide critical information about construct representation and bias that can be used to articulate validity arguments.

The primary goal of this workshop is to provide attendees with theoretical and practical hands-on experience with the use of GLMMs to analyze rater-mediated outcomes. The workshop will cover dataset construction and model selection, building, comparison, and evaluation, addressing statistical assumptions and reporting practices as well. A comparison is given between this modelling framework and models within the Rasch measurement family frequently used by language testers. Participants will conduct sample analyses using freely downloadable software, R Studio. Real datasets from published studies will be made available for practice analyses of phenomena, such as the impact of rater attributes and phonological features on comprehensibility ratings and the role of nonverbal behavior on language ability measures. The workshop will have a strong practical focus and all code will be available to participants to download. Participants will have time at the end of the workshop to practice modeling and exploring their own datasets.

Karen Dunn is a senior researcher and head of data for British Council English Language Research. She specializes in language test evaluation, statistical methods and measurement, and has worked extensively with latent trait models, including item response theory and mixed models. Karen has a PhD in Applied Social Statistics, in which she focused on explanatory item response theory as applied to language test data. She has taught courses in statistical methodology for analyzing linguistic and psychological data at undergraduate and postgraduate level, and has previously run workshops with a statistical focus at EALTA, the British Council's New Directions conference and UKALTA's Language Testing Forum. She has published methodological papers in *Frontiers in Psychology*, and *Studies in Second Language Acquisition*, and *Research Methods in Applied Linguistics*.

Dylan Burton is an assistant professor at Georgia State University in the Department of Applied Linguistics and ESL. He specializes in research on language assessment and second language acquisition, with an interdisciplinary focus on nonverbal behavior and affect. Dylan holds a PhD in Second Language Studies from Michigan State University. His past projects have involved various uses of multilevel data to account for the effects of raters and other sources of variance, particularly with ordinal outcomes. He has taught classes and led workshops on language testing, second language acquisition, and quantitative research methods. His research has been published in *Language Testing*, *Studies in Second Language Acquisition*, *The Modern Language Journal*, *Language Assessment Quarterly*, *System*, and *Papers in Language Testing and Assessment*.

Workshop D: Bridging Global Technological Innovations and Local Language Testing Realities

Day: Wednesday, June 3 (9 a.m. to 4 p.m.)

Location: Symphonie 3

Slobodanka Dimova, April Ginther &

Xun Yan



This interactive pre-conference workshop centers local language testing in the fast-moving conversation about artificial intelligence (AI). While most research, discourse, and tools target large-scale tests or classroom assessment, we focus on how local testing programs can responsibly evaluate, adapt, and—where appropriate—integrate AI (especially generative AI [GenAI]) into test design, administration, validation, and use. After a brief, non-technical overview of current capabilities of AI technologies in assessment, participants will map their own contextual realities (test purpose, test-taker population, data infrastructure and requirements, stakeholder network, and instructional and technical resources) to outline a sustainable, AI-enhanced assessment ecosystem. A central principle in this workshop is balancing innovation with reliability, validity, fairness, and practicality, including attention to bias, data governance, and local ownership of assessment practices. Through case studies from local programs that have and have not adopted AI, we introduce a structured decision framework to judge “AI readiness” and align potential solutions with program goals, constraints, and values. Working in small groups, participants will apply the framework to their settings, draft actionable plans, and identify community- of-practice strategies to build capacity and share resources. By the end, attendees will be able to (1) critically assess realistic opportunities and limitations of AI for local tests; (2) locate expertise and resources in their ecosystems; (3) evaluate ethical and fairness implications; and (4) produce a context-sensitive roadmap for developing or adapting AI systems or tools for local assessment. The session prioritizes agency, inclusivity, and feasibility so that local testing programs can leverage technological advances without compromising the core missions and values of local language testing.

Slobodanka Dimova is a Professor of Multilingualism and Language Testing at the University of Copenhagen, Denmark. Her research focuses on the development, validation, and implementation of local language tests in higher education and professional contexts, including university entrance and English medium instruction (EMI) lecturer certification exams. She has extensive experience directing local testing initiatives, such as the Test of Oral English Proficiency for Academic StaA (TOEPAS), which she helped design to address institutional and linguistic needs within the Danish context. In recent years, her research interests have been related to how technological innovations, including AI-driven tools, can be integrated responsibly into localized testing frameworks without compromising contextual validity or educational equity.

April Ginther is a Professor Emerita of English at Purdue University, taught graduate classes in language testing and quantitative research. She also directed two of the university’s English language support programs: The Oral English Proficiency Program (OEPP), the support program for international teaching assistants and The Purdue Language and Cultural Exchange (PLaCE), the support program for incoming international undergraduate students. As director of both programs, she was responsible for the development, maintenance, and administration of the local exams used to evaluate the English language skills of 1000 incoming students each academic year. She also served for five years as the co-editor of the journal *Language Testing*.

Xun Yan is a Professor of Linguistics, Second Language Acquisition and Teacher Education, and Educational Psychology at the University of Illinois at Urbana-Champaign (UIUC). He is a faculty member in the Beckman Institute for Advanced Science and Technology. Xun’s research interests include speaking and writing assessment, psycholinguistic and corpus/computational approaches to language testing, and language assessment literacy. Xun directs three local language testing programs, the English Placement Test and the Oral English Assessment Interview at UIUC, and the TQII (Training Qualified Interpreters in Illinois) Multilingual Proficiency Test for the state of Illinois in the US. He is a current co-editor of the journal *Language Testing*.

Opening Session: Featured Speaker

The Hidden Side of Montreal's Historical Multilingualism

Day: Wednesday, June 3 (5 p.m. to 6:30 p.m.)

Location: Opera AB

Amélie-Hélène Rheault



Les tensions linguistiques à Montréal ne datent pas d'hier. Si son passé de ville coloniale française conquise par l'Angleterre avait mis la table pour une cohabitation des francophones et des anglophones, ce sont les dynamiques sociales entre les deux groupes qui ont servi de terreau aux querelles linguistiques des années 1950 et 1960. Celles-ci ont obligé les gouvernements du Canada et du Québec à légiférer en matière de langue, espérant atténuer les conflits. Encore aujourd'hui, cependant, Montréal est toujours secouée par des conflits linguistiques et identitaires, ancrés dans l'imaginaire linguistique de tous les Québécois par leur couverture médiatique.

Sur fond d'histoire sociale, ce sont les manifestations de ces tensions linguistiques qui seront discutées ici (par ex. la crise de Saint-Léonard en 1967-1969 ou le tapage médiatique entourant le monolinguisme du président-directeur général d'Air Canada en 2021 et en 2026), de même que certaines mesures prises par les gouvernements provincial et fédéral (commissions d'enquêtes et lois linguistiques au Québec et au Canada).

(English version)

Although Montreal is known for its multilingualism, bringing many French speaking immigrants to this province to live the American dream, linguistic tensions in Montreal have shaped its history for centuries. While its past as a French colonial city conquered by England laid the groundwork for the coexistence of Francophones and Anglophones, it was the social dynamics between the two groups that fostered the linguistic disputes of the 1950s and 1960s. These conflicts compelled the Canadian and Quebec governments to legislate on language matters in the hope of easing tensions. However, Montreal continues today to be shaken by linguistic and identity-related conflicts, deeply embedded in the linguistic imagination of all Quebecers through their media coverage.

The manifestations of these linguistic tensions will be examined here (e.g., the Saint-Léonard crisis of 1967–1969 or the media uproar surrounding the unilingualism of Air Canada's CEO in 2021 and 2026), as well as certain measures implemented by the provincial and federal governments (commissions of inquiry and language laws in Quebec and Canada).

Amélie-Hélène Rheault has been teaching linguistics at the Université de Sherbrooke for over 20 years. She holds a joint PhD in sociolinguistics from the Université de Sherbrooke and the Université catholique de Louvain in Belgium : her dissertation focused on the linguistic insecurity of Quebecers and their discourse on the French language. Her interests encompass all aspects related to the relationship with language, primarily in Quebec.

Amélie-Hélène Rheault has been invited to participate on the television program *La table de Kim* to discuss language—a show broadcast on public television—and has given several public lectures, including for the Association Québec-France.

Symposia

Symposium: From Policy to Practice: Local and Global Co-operation

Chair: Nahal Khabbazzbashi

Discussant: Yasuyo Sawaki

Day: Friday, June 5th (1:30 p.m. – 3:00 p.m.)

Location: Opera AB

In 2022, Tokyo Metropolitan Government Board of Education (TMG) introduced the English Speaking Achievement Test (ESAT-J). The objective of this reform initiative was to evaluate both the achievements and the challenges associated with the development of speaking ability – one of the four core language skills (listening, reading, speaking, and writing) addressed in lower secondary education. The reform also aimed to enhance instructional practices in alignment with the objectives of the National Curriculum Guidelines. In addition, it sought to introduce a summative assessment of speaking proficiency at the third-year junior high school level as part of the entrance examination process for Tokyo Metropolitan high schools.

Following the first administration, TMG partnered with a major UK-based English language teaching learning and assessment organisation and expanded the project to include formative assessment for students from years 1 and 2.

To most readers, Tokyo is known as the world's largest city with a population of approximately 14 million living in a modern high-tech urban conurbation. The reality is that Tokyo encompasses a wide range of districts with diverse characteristics, since the metropolitan area includes a vast remote mountainous region to the west and a slew of islands that reach as far as Ogasawara, 1,000 kilometres to the south. With over 600 junior high schools spread across this geographical region the logistical challenges become obvious. This geographical diversity is compounded by the availability and reliability of the technology required to deliver the proposed tablet-based solution.

The decision to introduce the innovation was not taken lightly, with long discussions taking place at the Tokyo Metropolitan Government Board of Education. Meanwhile, it should also be noted that opposition to the implementation was voiced within the Tokyo Metropolitan Assembly and from other organizations.

The development team, made up of experts from both Japan and abroad, then undertook a series of projects analysing the national course of study (curriculum), all approved textbooks and relevant research. This resulted in a suite of three assessments that were demonstrably relevant to the Tokyo context. The team also worked with colleagues at TMG to devise communications and teacher development plans to support the roll out.

This symposium begins with a discussion by a leading policy maker at TMG of the formation and implementation of this major education reform initiative, a first for Japan. This offers attendees a unique opportunity to glimpse behind the scenes of high-level policy making and the pressures faced by policy makers and implementers. The second paper then presents an overview of the research and development efforts that supported the design and development of the resulting assessment suite. The third paper then deals with the delivery and rating of the assessments, highlighting the logistical challenges and solutions. The final paper reflects on the outcomes of the assessments, with details of achievement levels, candidate experiences of the process and early indicators of washback. Finally, the discussant reflects on the lessons learnt from the whole process and looks to generalise these reflections to other contexts.

Paper 1: Policy Formation and Implementation

Yoshihiro Takizawa (Tokyo Metropolitan Board of Education)

Paper 2: Research and Development

Barry O'Sullivan (British Council), Richard Spiby (British Council)

Paper 3: Delivery and Rating

Akihiro Matsuura (British Council, Japan)

Paper 4: Performance and Impact

Karen Dunn (British Council), Barry O'Sullivan (British Council)

Paper 1: Policy Formation and Implementation

Yoshihiro Takizawa (Tokyo Metropolitan Board of Education)

After a prolonged period of discussion and with experts, as well as preliminary research and pilot testing the Tokyo Metropolitan Government Board of Education (TMG) formally introduced the English Speaking Achievement Test (ESAT-J) in 2022. The primary objective of this initiative was to examine both the outcomes and the challenges related to the acquisition of speaking skills among the four language skills—listening, reading, speaking, and writing—studied in junior high school, to further enhance instruction toward achieving the objectives of the Course of Study, and to introduce a summative assessment of speaking proficiency at the third-year junior high school level in the Tokyo Metropolitan high school entrance examinations. TMG partnered with a major UK-based organisation and decided to expand the project to include the formative assessment of students from Years 1 and 2.

Implementing the test for approximately 240,000 junior high school students across the entire Tokyo metropolitan area also posed significant challenges. While there were supportive views, there were likewise many opposing voices; however, incremental improvements were made over time, leading to the present situation.

In this paper, the speaker, a leading policy maker at TMG, offers an insider's perspective on the pressures and processes that led to the creation of the policy, tracking the ESAT-J from the drivers behind the initial policy decisions, to the internal TMG discussions around the potential implementation of the policy. The paper concludes by reflecting on the lessons learnt from the process from policy and implementation perspectives.

Paper 2: Research and Development

Barry O'Sullivan (British Council), Richard Sipby (British Council)

Taking a Comprehensive Learning System (CLS) approach to building a fully integrated English language learning system, the research and development team undertook a number of projects. These projects are presented here in terms of how they contributed to the integration of the ESAT-J into the current TMG junior high school system and to the ultimate design of the assessments that form the system.

The first project reviewed the Japanese course of study for English. The team focused on the spoken communication elements, the types of learning activities proposed or implied by these elements and the themes and topics included. This resulted in a detailed specification of appropriate task types and a corresponding set of topics that could be used in a future test. The second project involved a formal review of the three levels (Years 1, 2 and 3) of the six textbooks formally approved by the Ministry of Education, Culture, Sports, Science and Technology. During this project, all speaking tasks were recorded and specified using the pro-forma templates developed for the Aptis test (see O'Sullivan & Dunlea, 2015). By taking this approach, the team were able to further refine the types of tasks that might be used across all three assessments.

The third research project then focused on the design of a series of three assessments, from the preliminary concept to the development work undertaken by the team. This project saw the team work with expert colleagues in the UK and Japan. In the latter case, these comprised fellow team members as well as locally and internationally based Japanese experts. Sub-projects included here involve exploration of white noise to combat the impact of external interference.

The paper finally reflects on the process of development with a focus on the collaboration between the external and internal expert teams and TMG.

Paper 3: Delivery and Rating

Akihiro Matsuura (British Council, Japan)

This paper is presented by colleagues who designed the administrative and scoring elements of the assessment system. This involved the creation of an operational plan to ensure that the assessments were downloaded onto tablet computers, then distributed across the Metropolitan area and securely delivered. The tablets then had to be collected, and the performance recordings uploaded to the scoring system. Finally, the performances were evaluated by a team of trained human raters.

Complications included:

- The Year 3 assessment (approximately 70,000 candidates) was to be delivered in a single day in November, with the first and the second session. The Years 1 and 2 sessions (approximately 140,000) were spread over six weeks in February and March.
- The turnaround time for scoring/feedback was approximately six weeks.
- The process was significantly impacted by the geographical challenges faced by the team as over 600 schools were involved in the project.
- Numerous staff were required for the Year 3 administration, these included invigilators, centre staff and supervisors.
- Raters were recruited globally (to allow for 24 hour rating). These were trained, evaluated and allocated performances.
- Rater performance was monitored by senior experienced examiners who had also been trained for ESAT-J.

The paper reflects on the various challenges, outlines the solutions arrived at and discusses the implications for similar initiatives both in Japan and internationally. This is of significant importance given the recent growth in awareness of the importance of teaching, learning and assessing the productive skills, heightened by the imminent introduction of a modern language (English) speaking test to the PISA suite.

Paper 4: Performance and Impact

Karen Dunn (British Council), Barry O'Sullivan (British Council)

The final paper in this symposium initially focuses on the process of generating test scores for Years 3. Years 1 and 2 assessments are feedback-focused and are considered as being low stakes for candidates. However, as TMG use the data for future planning and resourcing, the stakes are higher for schools. The process begins with the reconciliation of the various candidate lists, from school attendance registers (generally handwritten) to the performances stored in the assessment database post retrieval from the tablets. The Measurement team ensure version comparability and also statistically analyse the data to finalise the results output. The processes behind this are outlined and discussed.

This section of the paper ends with a presentation of the results from the 2024 and 2025 sessions for all assessments. The results demonstrate that the target levels of response were met for all assessments. These range from PreA1 to A2.2 using the CEFR-J grading system. Unlike the CEFR which recognises A1, A2.1 and A2.2, the CEFR-J subdivides the A1 level into three. The results and the underlying trends are discussed.

In the second part of the paper, we turn to the impact of the test, both on candidates and on broader society. The details of a student feedback survey are first presented and discussed, with the data showing a positive reaction to the assessments and a keenness for them to continue. This survey is presented as part of a large-scale ongoing washback study, the preliminary outcomes of which will be shared.

The paper then turns to the reaction of other stakeholder groups to the assessments. The adoption of an interactive communication model, these reactions, which consisted of a range of opinions and suggestions, has contributed to the gradual improvement of the whole system. This communication approach has seen a considerable easing of the opposition, though the work continues.

Symposium: Are Humans Needed? The Use of Spoken Dialogue Systems in Assessing L2 Oral Communication

Chairs: Gary Ockey & Elvis Wagner

Discussant: Elvis Wagner

Day: Friday, June 5th (1:30 p.m. – 3:00 p.m.)

Location: Symphonie 1

Spoken dialogue systems (SDSs) utilizing artificial intelligence (AI) are increasingly being used in L2 oral communication assessments, serving as administrators and interlocutors. In some ways, SDSs can be seen as the “holy grail” of L2 language testing, in that they enable test developers to efficiently and reliably assess L2 learners’ communicative ability without the need for expensive and error prone human interlocutors. These systems have been applauded for their potential for standardization by increasing uniformity of delivery and reliability of scores, potentially making them fairer than approaches that use human interlocutors (Karatay & Xu, 2025; Ockey, et al., 2023). While seen as a panacea for some, the use of such systems also introduces thorny validity concerns. At the center of these concerns are the following questions: Is it advisable to make inferences about a person’s ability to orally communicate with humans based on results from talking with AI? How similar is SDS interaction with a human during an assessment compared to target language use language interactions with a human partner? To what degree can SDSs provide test takers with opportunities to demonstrate their abilities to effectively interact with a human? To what extent can human/SDS speaking tests capture the range of abilities involved in oral communication? There is also the issue of individual differences; research indicates that some test takers prefer talking with machines while others prefer talking with humans when having their oral communication skills assessed (Timpe-Laughlin et al., 2022). This may result from certain individual characteristics, such as personality, proficiency level, age, experience with technology, and views about the use of AI taking a role that humans have traditionally had impacting how people perform across the two delivery modes. In this colloquium, we have a group of experts with diverse views on this issue. The three paper presenters will make a case for their views based on the rapidly emerging research on this topic, and each presentation will (try) to answer the question, “Are human interlocutors needed to deliver effective oral communication assessments?” The symposium will begin with a short introduction and overview of the topic, followed by the three papers and a discussant. Each paper presentation will take 18 minutes, and the discussant will speak for 12 minutes. The idea of using SDSs for delivering oral communication assessments is one of the most compelling and controversial issues in L2 assessment. Many individuals both within and outside of the language assessment field have strong views on this topic. Thus, we think it is important to devote a significant part of the symposium (more than 20 minutes) to questions and thoughts from the audience. Our expectation is an engaging discussion of the topic that will provide participants with a better understanding of the somewhat abstract validity issues regarding the use of these emerging technologies, while also providing some concrete guidance for using these technologies when assessing L2 oral communication, and suggestions for future research in this area.

Paper 1: Spoken Dialogue Systems and Potential Interactions with Test-Taker Personal Characteristics

Gary Ockey (Iowa State University)

Paper 2: Conversational AI as Interlocutor

Erik Voss (Teachers College, Columbia University), Hansun Zhang Waring (Teachers College, Columbia University)

Paper 3: Revisiting the Socio-Cognitive Framework for AI-Mediated Oral Communication Assessment: Validity Considerations

Evelina Galaczi (Cambridge University Press & Assessment), Yasin Karatay (Cambridge University Press & Assessment)

Paper 1: Spoken Dialogue Systems and Potential Interactions with Test-Taker Personal Characteristics

Gary Ockey (Iowa State University)

Can scores from SDS interactions be generalized to human partner interactions?

Personal characteristics, such as personality, anxiety, and AI familiarity may impact how people perform when speaking with a human or a spoken dialogue system (SDS). For instance, people who are introverts or get anxious when speaking with a human or who spend a lot of time with computers may be more successful when taking an oral communication test with an SDS than with a human partner. On the other hand, test takers who enjoy talking with other humans and believe that talking with an SDS is not engaging or meaningful may be more successful when taking the test with a human partner. As a result, using SDSs as interlocutors in an oral communication assessment could lead to inferences about oral ability when speaking with a human that are not warranted. In this talk, I will discuss an ongoing study that addresses these issues. In the research, university English learners spoke with both a spoken dialogue system and a human partner, and these discussions were video recorded. Learners then completed questionnaires about their experiences with the two systems and about their backgrounds, levels of extroversion, oral assessment anxiety, and experience with AI. Human raters provided oral communication scores based on the video recordings. A series of repeated-measures ANOVAs, with oral ability as the dependent variable, was used to determine if there was an interaction between each of these variables and presentation mode (human or computer partner). The research raises many questions, such as the degree to which we should attempt to reduce test takers' anxiety when they take a test and whether or not we should consider personality as part of the oral communication construct.

Paper 2: Conversational AI as Interlocutor

Erik Voss (Teachers College, Columbia University), Hansun Zhang Waring (Teachers College, Columbia University)

Interaction with conversational AI has shown promise for assessing second language oral proficiency (Karatay & Xu, 2025). However, research indicates that speaking with a computer is different than speaking with a human (Ockey & Chukharev-Hudilainen, 2021). It would be particularly problematic then to make inferences about a test taker's ability to conduct human interaction based solely on their talking with an AI interlocutor. As such, our answer to this symposium's question of whether it is defensible to do so is: Not yet. Clearly, efforts are needed to enhance the naturalness or "authenticity" of conversational agents (Voss, 2024). Drawing upon five decades of conversation analytic research, one recent endeavor of assessing AI naturalness highlights three AI-specific principles of natural conversation (i.e., emergence, indexicality, and recipient design) (Voss & Waring, 2025).

In this talk, we illustrate issues of naturalness in using conversational AI as an interlocutor, with a specific focus on the principle of recipient design – that of designing each turn at talk for the specific recipient in that specific moment. One might think of recipient design as the opposite of monologic renderings devoid of any audience considerations. More specifically, we demonstrate the extent to which an AI interlocutor behaves in ways that embodies, or deviates from, the principle of recipient design.

Recordings of L2 learners interacting with ChatGPT along with their conversation analytic transcripts were repeatedly examined to locate discourse features that evidence the presence or absence of recipient design. Our findings provide a partial, but nuanced and empirically based, profile of ChatGPT's current ability of enacting the role of a human interlocutor in the potential context of oral assessment – with both promises and pitfalls. We end our talk by calling for the much-needed groundwork to assess AI speaking before using AI as an interlocutor in assessing human speaking.

Paper 3: Revisiting the Socio-Cognitive Framework for AI-Mediated Oral Communication Assessment: Validity Considerations

Evelina Galaczi (Cambridge University Press & Assessment), Yasin Karatay (Cambridge University Press & Assessment)

Spoken dialogue systems (SDSs) are increasingly being investigated as interlocutors in L2 oral communication assessment. Much of this research has focused on technical feasibility, yet less attention has been paid to how SDS-mediated assessments align with established validity frameworks. This symposium presentation revisits Weir's (2005) socio-cognitive framework and calls for a systematic investigation of the use of SDSs in assessment, focusing on identifying the contexts in which such tests may—or may not—be fit for purpose.

Through the socio-cognitive lens, key areas of concern emerge. Cognitive validity prompts us to ask whether SDS tasks engage the same processes and interactional strategies as human-mediated exchanges. Context validity raises questions about authenticity: Does a conversation with a machine constitute a plausible communicative situation for test purposes, and how might learner variables, such as proficiency level, L1 background, or familiarity with digital agents, affect performance? Scoring validity requires us to examine whether rating scales designed for human interaction can be meaningfully applied, or whether new descriptors of “AI-mediated interactional competence” are needed. Finally, consequential validity highlights issues of fairness, washback, and test-taker perceptions of authenticity and anxiety.

By foregrounding the socio-cognitive perspective, we argue that uses of SDS-based assessments need to be grounded in validity frameworks that capture a range of considerations about fitness for purpose and score interpretability. Two examples, one from classroom-based assessment and one from a high-stakes context, will illustrate operational use and validity implications to guide principled decisions about when and how SDSs are most appropriately used. Ultimately, this talk underscores the imperative for AI-driven approaches to be evaluated not only on technical grounds but also in terms of validity considerations and their fitness for purpose in assessing human oral communication.

Symposium: Multilingual Assessment - Toward a Principle of Localization

Chairs: Saskia Van Viegen & Sunny Lau

Discussant: Constant Leung

Day: Friday, June 5th (1:30 p.m. – 3:00 p.m.)

Location: Symphonie 2

Transforming education practice requires concomitant change in current monolingual approaches to assessment toward localized, multilingual approaches that contribute to equity and inclusion across K-12, post-secondary and adult education contexts. Recognizing the colonial imprint of assessment, research across language testing, applied linguistics, and TESOL has begun to account for the fundamental legitimacy of bi/multilingual students' fluid, complex, and creative language practices and cross-linguistic awareness. This research suggests that traditional linguistic-cognitive models of language assessment do not reflect the dynamic, contextualized of students' language in use, potentially reinforcing English dominance and linguistic hierarchies. As an umbrella term, multilingual (ML) assessment integrates multilingualism in test design and learners' responses (e.g., Heugh et al. 2017; Schissel et al., 2019), recognizing students' fluid and dynamic language practices and multilingual language awareness. However, implementation of such efforts in educational contexts remains few. Despite great motivation to engage in and experiment with multilingual and cross-linguistic instruction and assessment practice, educators and researchers need clarity on the theoretical constructs of ML assessment as well as classroom-based research evidence and related professional development in theory and practice to continue to inform and support teaching and learning in their local context. Moreover, localization is critical to the design of ML assessment to ensure approaches are adapted to specific contexts, to ensure assessment tasks and tools are culturally appropriate, relevant, and functional. Engaging with this research and teaching issue, our proposed symposium synthesizes insights from four interrelated strands of inquiry, including a meta-analysis synthesizing research efforts and policy issues in localized language assessment in the Tanzanian context, an empirical study incorporating multilingualism and multimodalities for learning-oriented, classroom-based assessment in two Canadian provinces, another study focusing on the development of a context-specific scale for measuring cross-linguistic awareness for French-English biliteracy, and concluding with a systematic review of how translanguaging in assessment functions to support equity and linguistic justice. Taken together, these studies illustrate the possibilities and potentials in moving beyond a fixed conceptualization of language in assessment, challenging the traditional focus on achievement of standardized language norms and isolated skills. They demonstrate how localized assessment design, grounded in community language ecologies, can address equity and social justice for bi/multilingual students and the teaching and learning needs of Indigenous, racialized and other historically marginalized students. The symposium will conclude with a discussion and Q&A led by Prof. Constant Leung, to shed light on how to reconfigure epistemological traditions and methodologies of assessment in equitable ways to better align with the language and knowledge making evident in local classrooms and communities. By elevating localization (Chalhoub-Deville & O'Sullivan, 2020) as a central principle for multilingual assessment, the symposium expands the landscape of language assessment development and contributes to decentering traditional, global models, making assessment more relevant, valid, and equitable for diverse students and educational contexts.

Paper 1: Localizing Assessment in Tanzania's Language-in-Education Policy Context: A Systematic Review

Jamie Schissel (University of North Carolina Greensboro), Donather Magabe (University of North Carolina Greensboro)

Paper 2: Exploring Multilingual Primary Students' Cross-Linguistic Awareness in French-English Education

Susan Ballinger (McGill University), Beverly Baker (University of Ottawa), Claude Quevillon Lacasse (University of Ottawa)

Paper 3: Creating multilingual assessment ecologies in Canadian classrooms

Saskia Van Viegen (Education University of Hong Kong), Sunny Lau (Bishop's University)

Paper 4: Translanguaging for Equity and Justice in Assessment: A Systematic Review

Zhongfeng Tian (Rutgers University), Chia-Hsin Yin (The Ohio State University), Jessica McConnell (The University of Texas at San Antonio)

Paper 1: Localizing Assessment in Tanzania's Language-in-Education Policy Context: A Systematic Review

Jamie Schissel (University of North Carolina Greensboro), Donather Magabe (University of North Carolina Greensboro)

Our paper explores assessment approaches where English medium of instruction (EMI) occurs within communities that are supportive of local languages in education. We present the case of Tanzania, where the Ministry of Education has made attempts to include local languages in LEPs. Tanzania has more than 120 local languages. Kiswahili is the most common. In 2015, the government debated the use of Kiswahili as the dominant language in education. Though ultimately this did not become an official policy, discourses about the inclusion of local languages remain (Yogi, 2017) in part because EMI in secondary and higher education poses ongoing challenges for assessment (Qorro, 2013; Hiza & Paschal, 2023).

We conceptualize assessments in this context as defacto LEP (Shohamy, 2006; Van Viegen & Jang, 2021). In many parts of the world, assessments function as defacto policies that exclude local languages (e.g., Sub-Saharan Africa, Reilly et al. 2023; Northern/Eastern Europe, Jánk, 2025; USA, Menken, 2008; Park, 2024; Morocco, Chakrani et al., 2025). Our research question asks: how do studies document assessment practices around local languages in Tanzania?

We are using PRISMA methodology to review empirical literature from 2015 to present across the databases EBSCO, ERIC, and JSTOR. Our preliminary identification phase includes 191 articles. Screening to determine final inclusion will be completed before the conference.

Initial analyses indicate three literature trends: increased use of local languages in summative assessments, ongoing use of English-only in summative assessments, initial use of translanguaging approaches in formative assessments.

These trends indicate the Tanzanian context is cultivating a climate where support for local languages in assessments continues despite policies that maintain EMI. The implications for other contexts are potentially quite significant, given that many studies attribute policies and assessments as major barriers for inclusion of local languages and/or multilingual approaches.

Paper 2: Exploring Multilingual Primary Students' Cross-Linguistic Awareness in French-English Education

Susan Ballinger (McGill University), Beverly Baker (University of Ottawa), Claude Quevillon Lacasse (University of Ottawa)

In recent decades, researchers have highlighted the need to make multilingual education truly multilingual rather than treating learners' languages as siloed entities. Some researchers have also aimed to develop and assess what has been labeled as cross-linguistic awareness (XLA)—or learners' awareness of similarities and differences across languages in their repertoire (Jessner, 2006)—as a central ability to enact a multilingual paradigm in the language classroom. However, to date, researchers have operationalized this ability through comparative assessments of learners' metalinguistic awareness in individual languages (e.g., Jessner et al., 2016). We argue that this one-language-at-a-time assessment practice misses the dynamic and contextualized nature of XLA. The current study therefore employs Qualitative Retrodictive Modeling (QRM; Dörnyei, 2014), a method grounded in Complex Dynamic Systems Theory (Larsen-Freeman & Cameron, 2008), to empirically develop a rating scale capable of reflecting learners' XLA as a dynamic, localized, repertoire-based ability, rather than as a generalized ability housed in separate languages.

Our presentation will outline the second year of a multi-year project in Grades 3 and 4 of a private, French-English bilingual primary school in Montreal, Canada. In Year 1, three teachers of (N = 33) students collaborated to develop and deliver instruction targeting development of XLA. We used oral interview tasks to elicit learners' cross-linguistic comparison of English and French and administered them on two occasions. We then used QRM to first categorize students' responses into XLA profiles to develop a holistic rubric for researchers and teachers within this bilingual education context. This rubric was then piloted in a public French-English primary school and compared with learners' performance on biliteracy tasks. We will share insights from this pilot administration and the potential future use of this rubric as both a research tool and a teacher support for crosslinguistic pedagogy.

Paper 3: Creating multilingual assessment ecologies in Canadian classrooms

Saskia Van Viegen (Education University of Hong Kong), Sunny Lau (Bishop's University)

Despite trans/plural approaches to language and content instruction that invite students' multilingualism into learning official languages of instruction (LOI) in Canadian classrooms, languages are taught in compartmentalized ways with emphasis on achieving native-like proficiencies in single languages. Assessment continues to adhere to English- and individual-centric view of language with fixed norms. This traditional linguistic-cognitive model of language assessment does not reflect students' multilingualism, flexible language use, and agentive assemblage of digital and multimodal resources for communication and learning (Ou & Malmström, 2023). Within language assessment and education literature, researchers have called for attention to fluid and dynamic language practices, such as the integration of multilingualism and multimodality in test design and in learners' responses (e.g., Heugh et al. 2017; Schissel et al., 2019), but understandings of how to develop and implement such approaches remains few.

Grounded in classroom inquiry and dialogic engagement with teachers and students in two Canadian provinces, this paper shares insights from situated research with students and teachers into language practice, learning, and assessment. Elaborating research-practice partnership in two contexts, one in Ontario with students from immigrant and refugee backgrounds learning in English and the other in Quebec with students learning French as second language, we explore how educators are incorporating multilingualism, multimodality and multi-sensory resources for learning-oriented assessment (Shohamy & Pennycook, 2019). Centering learning in integrated instruction and assessment activities, teachers engaged students' linguistic, cultural and conceptual abilities, particularly pragmatic, sociolinguistic capabilities for interactive and academic purposes (Leung, 2022). Illustrated with classroom artifacts and student and teacher interviews, we show how students engaged with multilingual and multimodal resources for meaning co-construction and demonstration of understanding. The multilingual assessment approach invited teachers to apprehend the complexity of students' languaging practices and behold their learning potential, possibilities that are often veiled and rendered invisible by the target-language only policy.

Paper 4: Translanguaging for Equity and Justice in Assessment: A Systematic Review

Zhongfeng Tian (Rutgers University), Chia-Hsin Yin (The Ohio State University), Jessica McConnell (The University of Texas at San Antonio)

While translanguaging has gained global attention as a pedagogical approach that affirms multilingual learners' identities and promotes equitable participation (e.g., Hamman-Ortiz et al., 2025), assessment remains one of the most monoglossic and globally standardized domains of education. This systematic review examines how translanguaging has been conceptualized and enacted in educational assessment research across diverse local contexts, illuminating tensions between global assessment norms and localized multilingual realities.

Drawing on 33 empirical studies (2012–2023) identified through major databases (ERIC and Google Scholar), we employed an inductive–deductive coding approach informed by a heteroglossic view of language to analyze research settings, assessment types, methodological trends, and justice-oriented engagements. The studies span K–12, higher education, and community contexts across both the Global North and Global South.

Findings reveal that translanguaging assessments, primarily implemented in classroom-based, formative settings, enable multilingual learners to demonstrate knowledge more authentically, affirm linguistic identities, and reduce linguistic anxiety. Teachers report increased agency, stronger relationships with students and families, and more nuanced insights into student learning. However, systemic constraints, including monolingual policies, global standardization pressures, and limited institutional support continue to restrict translanguaging practices, particularly in high-stakes testing contexts.

To theorize these findings, we propose a “spectrum of justice” framework illustrating how translanguaging assessment functions as (1) a descriptive practice enhancing inclusion, (2) a pedagogical tool fostering engagement and agency, (3) a political act resisting linguistic hierarchies, and (4) a transformative model for systemic change.

By bridging global discourses on translanguaging with locally grounded assessment practices, this review contributes to a more context-sensitive understanding of equity and validity in multilingual assessment. We call for assessment designs that recognize translanguaging as both a linguistic resource and a justice-oriented stance, reimagining fairness and validity across educational systems worldwide.

Symposium: Contextualizing Standard Setting and Alignment in Canada: Local Practices, Global Frameworks, and Policy Relevance

Chairs: Angel Arias & Eunice Eunhee Jang

Discussant: Stephen Sireci

Day: Friday, June 5th (1:30 p.m. – 3:00 p.m.)

Location: Symphonie 3

Language assessment operates within interconnected local and global contexts that shape how validity evidence is gathered to ensure that test scores are interpreted appropriately and their uses are justified. In multilingual, policy-driven settings like Canada, language testing functions at the intersection of educational, professional, and immigration systems. This symposium examines how standard-setting and alignment research in Canada addresses the tension between language frameworks (e.g., CEFR, CLB/NCLC) and locally grounded communicative demands in diverse contexts. Across four coordinated studies, the symposium examines how contextualized assessment practices can yield empirically defensible yet socially responsive performance standards. The first presentation, *Landscape of Canadian Standard Setting and Alignment Research and Practice*, provides a macro-level overview of the Canadian standardsetting ecosystem. It situates emerging work within broader global trends, examining how regulatory bodies, test developers, and policymakers adapt traditional standard-setting approaches—such as the item descriptor matching method and psychometric analysis—to digital-first and AI-enhanced tests. By examining recent studies aligning the CLB and CEFR, it illustrates how methodological adaptations enhance fairness and interpretability across contexts, providing an empirical foundation for linking proficiency levels between frameworks and facilitating score comparisons across tests and jurisdictions. The second presentation, *A Triangulated Approach to Standard Setting for Language-Specific Purposes (LSP)*, focuses on standard setting for internationally educated nurses (IENs). Using a domain-expert model (Pill, 2016; Manias & McNamara, 2016), this study demonstrates how involving healthcare professionals as panelists ensures that cut scores reflect the linguistic realities of clinical communication. The triangulated design—integrating expert judgment and psychometric analysis—illustrates a context-sensitive approach that prioritizes safety and fairness in licensure decisions. The third presentation, *Renewing Language Proficiency Standards for Teacher Certification in Canada: A Multi-Phase Standard-Setting Initiative*, addresses the educational domain. This project, commissioned by a major Canadian teacher regulatory body, developed the target performance level descriptors (Egan et al., 2012) of the Minimally Competent Teacher-Candidate and conducted nine standard-setting studies across English and French tests. This work highlights the challenge of balancing consistency with bilingual equity and demonstrates how procedural validity evidence and expert consensus contribute to the development of defensible cut scores in accreditation contexts. The fourth presentation, *Aligning the Canadian Language Benchmarks with the AI-Enhanced, Adaptive Duolingo English Test Using Item-Descriptor Matching*, investigates innovation in large-scale test alignment. Using the Item-Descriptor Matching method (Ferrara et al., 2008) and the recommended guidelines for relating tests to language frameworks (Council of Europe, 2009). This study explores how standard setting can remain transparent and empirically grounded amid rapid technological change. Collectively, these papers illustrate how Canadian research bridges global frameworks and local needs through evidence-based standard-setting practices. Conceptually, the symposium advances understanding of contextual evidence of validity (DeLuca et al., 2019) and the interplay between psychometric defensibility and sociocultural appropriateness. Methodologically, it highlights innovations that adapt traditional standard-setting paradigms to AI-mediated, multilingual, and professional assessment contexts. Together, the four studies contribute to ongoing debates about how language testing can remain both globally comparable and locally meaningful—an issue of particular resonance for LTRC 2026 in Quebec’s bilingual, policy-rich environment.

Paper 1: Landscape of Canadian Standard Setting and Alignment Research and Practice

Eunice Eunhee Jang (University of Toronto)

Paper 2: A Triangulated Approach to Standard Setting for Language-Specific Purposes (LSP)

Maryam Wagner (McGill University)

Paper 3: Renewing Language Proficiency Standards for Teacher Certification in Canada: A Multi-Phase Standard-Setting Initiative

Angel Arias (Carleton University)

Paper 4: Aligning the Canadian Language Benchmarks with the AI-enhanced, adaptive Duolingo English Test using Item-Descriptor Matching

Jeanne Sinclair (Memorial University of Newfoundland)

Paper 1: Landscape of Canadian Standard Setting and Alignment Research and Practice

Eunice Eunhee Jang (University of Toronto)

Standard setting is central to the validity of language assessment, providing the rationale for translating test scores into performance levels that inform high-stakes decisions for educational placement or immigration policy (Cizek & Bunch, 2007). Establishing or aligning cut scores with frameworks like the CLB or CEFR provides evidence for the utilization inference that justifies score-based decisions (Kane, 2013).

Recent standard-setting and alignment research has expanded beyond academic tests to include digital-first assessments used in Canadian immigration and credentialing. Tests such as the Duolingo English Test, TOEFL Essentials, and PTE Core have aligned scores to the Canadian Language Benchmarks through mixed-methods studies involving expert panels, descriptor matching, and psychometric analyses (Duolingo, 2022; ETS, 2023; Pearson, 2024). Parallel research by North and Piccardo (2016) linking the CEFR and CLB supports score comparability across tests and jurisdictions, enhancing portability and fairness. Rising global migration has intensified the need to renew language standards for immigration and professional licensure. IRCC and regulatory bodies in healthcare and education (e.g., CNRC, OCT) now use standard-setting and alignment practices to ensure assessments reflect real communicative demands. With the advent of computer-adaptive testing, integrative task design, and automated scoring in tests such as the Duolingo English Test, TOEFL Essentials, and PTE Core, standard setting has become increasingly complex—requiring adapted or hybrid methods to maintain validity and enable alignment with frameworks such as the CLB and CEFR (Papageorgiou et al., 2019). North and Piccardo's (2016) CEFR–CLB alignment study provides a conceptual and empirical foundation for linking proficiency levels and facilitating score comparability across tests and jurisdictions.

This presentation provides an overview of the landscape of standard-setting practices in Canada, highlighting how diverse stakeholders, such as test developers, regulatory bodies, and policymakers, engage in standard-setting and alignment research to establish defensible proficiency standards aligned with national and international frameworks.

Paper 2: A Triangulated Approach to Standard Setting for Language-Specific Purposes (LSP)

Maryam Wagner (McGill University)

Nurses comprise the largest population of healthcare workers worldwide providing up to 80% of patient care (World Health Organization, 2020; 2025). In parallel, nursing is increasingly becoming a mobile profession with nurses drawn to work beyond the countries in which they were trained. In Canada, the credentialing requirements for these internationally educated nurses (IENs) requires demonstration of language proficiency through testing. Setting appropriate standards on these proficiency exams is paramount because if standards are set too low, the health and safety of the public is endangered, while standards that are set too high prevent competent nurses from entering practice (Davidson, 2022). Accordingly, it is vital to engage domain experts in standard setting activities so that they can apply “indigenous assessment criteria” (Pill, 2016) using their knowledge of both the clinical environment and the social uses of language in the target language use domain (Manias & MacNamara, 2016).

In this presentation, we report on a standard setting project guiding Canadian language proficiency requirements for IENs. We engaged domain experts from across Canada, representing three nursing designations, to serve as panelists (N=68). We facilitated five standard setting workshops for academic language testing (e.g., IELTS Academic), specific language testing (e.g., Canadian English Language Benchmark Assessment for Nurses), and testing French language proficiency (e.g., TEF-Canada). Each workshop involved establishing performance standards by defining the language characteristics of a minimally competent candidate using language skill descriptors from entry-to-practice competency frameworks. Next, we used both exam- and examinee-centred standard setting approaches to establish cut scores involving panelist training, discussing errors, using authentic test materials, psychometric analyses, and in-depth discussions. Finally, we employed rigorous quality control and evaluation processes to determine the quality of the standard setting procedures and the panel's confidence in the cut scores. We developed equivalence cut scores across the tests using the CEFR.

Paper 3: Renewing Language Proficiency Standards for Teacher Certification in Canada: A Multi-Phase Standard-Setting Initiative

Angel Arias (Carleton University)

Internationally educated teachers play a vital role in meeting Canada's growing demand for qualified educators; however, ensuring their readiness to teach in linguistically and culturally diverse classrooms requires valid and empirically grounded performance standards on measures of language proficiency. This imperative places renewed emphasis on the appropriate interpretation and use of test scores, a central concern in contemporary validity theory (AERA et al., 2014; Sireci, 2016). In this study, we report on a two-phase initiative commissioned by a major teacher regulatory body in Canada to renew its language proficiency requirements for teacher certification.

The first phase defined the Minimally Competent Teacher-Candidate (MCTC) through expert consultations and focus group discussions that examined the communicative competencies required for effective and ethical teaching. The resulting MCTC profile aligned approximately with CLB/NCLC 9 (CEFR B2+/C1, North & Piccardo, 2023), representing the necessary proficiency for effective communication in K-12 classrooms. This work laid the foundation to refine the target performance level descriptors (Egan et al., 2012) for subsequent standard setting studies.

The second phase comprised nine standard-setting studies to recommend empirically supported cut scores for five English and four French language tests accepted for certification. The English studies included IELTS Academic, IELTS General, TOEFL iBT, CELPIP-General, and PTE Core; the French studies included TEF, TCF, DELF B2, and DALF C1. Panelists applied modified Angoff methods, borderline descriptive, and regression methods, drawing on established frameworks for standard setting studies (Cizek, 2012). These approaches emphasized procedural validity, expert judgment, and the link between data and performance-level descriptors (Kane, 2001).

Findings across both language streams highlighted shared commitments to fairness and relevance. Grounded in research-based methods and expert consensus, this initiative enhanced the defensibility of certification outcomes and ensured that proficiency benchmarks align with the evolving communicative, cultural, and pedagogical demands of teaching in Canada.

Paper 4: Aligning the Canadian Language Benchmarks with the AI-enhanced, adaptive Duolingo English Test using Item-Descriptor Matching

Jeanne Sinclair (Memorial University of Newfoundland)

The increasing use of computerized adaptive and AI-enhanced language assessments requires renewed attention to standard-setting and language framework alignment methodologies. This study aligned the Duolingo English Test (DET), a fully computerized adaptive test employing multiple natural language processing (NLP)-scored tasks and machine-learning-based scoring, with the Canadian Language Benchmarks (CLB). Our goal was to establish empirically grounded cut scores for interpreting DET results in terms of CLB proficiency levels, enabling their potential use for immigration purposes in Canada. This study offered an opportunity to utilize an efficient, reliable, and valid approach to aligning multiple, integrated task types to multiple proficiency levels when total (sum) scores are not available.

We adhered to the phases proposed by the Council of Europe (2009): familiarization, construct-content congruence analysis, standard-setting, and multi-source evaluation. Methodologically, we adapted the Item-Descriptor Matching method (Cizek & Bunch, 2007; Ferrara et al., 2008; Ferrara & Lewis, 2012), organizing materials by difficulty, using calculations based on IRT parameters. The standard-setting activities were optimized by providing panelists subject-matter expert-derived descriptions of knowledge, skills, and abilities for each DET task across a range of score levels. A panel of 15 CLB experts first achieved consensus on performance level descriptors for each CLB level and then engaged in two iterative rounds of judgment supported by adaptive data visualization to establish multiple cut points. The final cut scores demonstrated strong internal consistency (Krippendorff's $\alpha > .84$) and close external alignment with existing DET–CEFR concordances.

Empirically, the study confirmed that DET content covers CLB levels 2–12 for reading and speaking, CLB 2-11 for writing, and CLB 3-11 for listening. This study demonstrates how standard-setting frameworks can be applied to AI-scored, adaptive assessments, and it advances dialogue on maintaining transparency, fairness, and interpretability in AI-enabled testing where multiple cut points are required for policy-relevant decisions.

Symposium: Reframing Language Assessment Literacy in Latin America: The Multiple Cases of Brazil, Colombia and Mexico

Chair: Elsa Fernanda Gonzalez

Discussant: Gladys Quevedo-Camargo

Day: Friday, June 5th (1:30 p.m. – 3:00 p.m.)

Location: Symphonie 4

This symposium brings together multiple research-based cases on the development of Language Assessment Literacy (LAL), the knowledge, methods, techniques and principles various language assessment stakeholders may need to decide, design and conduct assessment (Vogt, Bøhn & Tsagari, 2024), across diverse teacher populations in Latin America. The presentations span Brazil, Colombia, and Mexico, offering critical insights into how pre-service and in-service language teachers navigate, perceive, and enact assessment practices. Together, these studies highlight contextual needs, emerging practices, and possibilities for shaping context-sensitive, teacher appropriate, and pedagogically grounded LAL frameworks in the region.

The first presentation discusses a university-based LAL course for pre-service teachers (PSTs) in Brazil, detailing the course design, task structure, and short-term impacts on PSTs' assessment understanding. Drawing on end-of-course questionnaires and analysis of participant-designed tools, it explores the extent to which such localized initiatives foster principled, reflective LAL development. This presentation also situates the course within global discourses, encouraging comparative reflection and the development of responsive LAL curricula elsewhere.

The second study explores LAL among teachers of (very) young English language learners (ages 0–11) across Latin America, a demographic increasingly targeted for early language instruction yet underrepresented in research. Findings from an online questionnaire (n=65) reveal a fragmented landscape: while formative assessment is valued, practices are highly inconsistent and often driven by institutional demands rather than developmental or theoretical frameworks. The study highlights major gaps in formal preparation and calls for professional development opportunities that help teachers align their assessment choices with young learners' cognitive and linguistic needs.

The third presentation reports on a workshop implemented in five Colombian regions (rural and urban), designed to identify and address the LAL needs of in-service teachers. While initially aimed at English teachers, the initiative attracted educators from various disciplines, reflecting widespread interest in assessment capacity-building. Findings point to both the structural challenges and the transformative potential of participatory, needs-based professional development, with teachers designing tailored assessment instruments and reflecting critically on their classroom practices.

Finally, the fourth study examines an online language assessment training course offered to educators in Latin America, analyzing perceptions from both instructors and participants. Semi-structured interviews show that while the course encouraged self-reflection and tool improvement, issues such as lack of contextualization and tensions between teachers and institutional structures limit its long-term impact. The study calls for more adaptive, dialogic, and systemically supported training initiatives.

Collectively, the presentations foreground how LAL in Latin America cannot be disentangled from its socio-educational realities marked by policy variation, limited training infrastructures, and tensions between top-down mandates and grassroots needs. Additionally, it highlights how localising LAL may lead to meaningful support for teachers by cross referencing common ground practices among the contexts. The symposium invites participants to consider how local experiences can inform broader theoretical and practical understandings of LAL, and how teacher education can be reimagined to support meaningful, and ethical, assessment practices across diverse regional contexts.

Paper 1: Presenting a Brazilian Language Assessment Literacy Course: Syllabus, Pre-Service Teachers' Perceptions and Short-Term Impact

Isadora Teixeira Moraes (Federal University of Minas Gerais)

Paper 2: Developing Language Assessment Literacy in Early English Education: Insights from Latin America

Juliana Reichert Assunção Tonelli (State University of Londrina)

Paper 3: Links between teaching and assessment in context: a workshop to develop in-service teachers' (L)AL

Sonia Patricia Hernández-Ocampo (Pontificia Universidad Javeriana)

Paper 4: Mexican young learner EFL teachers and training instructors' perspectives: How useful is online LA training?

Elsa Fernanda Gonzalez (Autonomous University of Tamaulipas)

Paper 1: Presenting a Brazilian Language Assessment Literacy Course: Syllabus, Pre-Service Teachers' Perceptions and Short-Term Impact

Isadora Teixeira Moraes (Federal University of Minas Gerais)

In this presentation, I detail a language assessment literacy (LAL) course for pre-service teachers (PSTs) offered at a university in Brazil. I begin by briefly outlining how the syllabus was designed and structured. Next, I discuss PSTs' perceptions of the proposed tasks, drawing on data from an end-of-course questionnaire. I also report findings from an analysis of assessment tools designed by PSTs, aiming to examine the course's short-term impact on their LAL development. Some implications and directions for future research are outlined. In view of contrasting local and global realities, I consider how this course compares to other initiatives for promoting PSTs' LAL. I hope this presentation will not only highlight a local LAL initiative but also encourage teacher educators elsewhere to reflect on their own courses.

Paper 2: Developing Language Assessment Literacy in Early English Education: Insights from Latin America

Juliana Reichert Assunção Tonelli (State University of Londrina)

This study investigates the state of Language Assessment Literacy (LAL) among teachers of (very) young English language learners (ages 0–11) in Latin America—a region where early English instruction is expanding, yet systematic research on assessment practices remains scarce. While LAL is commonly understood as the combination of knowledge, skills, and principles related to language assessment (Inbar-Lourie, 2013; Fulcher, 2012; Taylor, 2013), little is known about how these components are applied in early childhood education, especially in under-researched contexts such as Latin America.

To address this gap, an online questionnaire was distributed via social media over three months, yielding 65 valid responses from teachers working in a range of settings, including public and private schools, bilingual programs, and language institutes. Findings reveal a fragmented and heterogeneous landscape of assessment practices. Teachers reported varying understandings of assessment's purpose and scope, with some using regular formal assessments and others relying solely on formative observation—or not assessing at all. Vocabulary, listening comprehension, and oral communication were the most frequently assessed domains, while writing and grammar were less emphasized. Teachers used diverse tools, such as quizzes, checklists, portfolios, and games, often without a clear rationale, indicating gaps in principled knowledge.

Although most teachers valued formative assessment for guiding instruction, many lacked confidence in aligning practices with developmental and language acquisition principles. Institutional mandates often shaped their approaches more than pedagogical frameworks. A consistent theme was the absence of formal training, with teachers expressing a strong need for professional development, particularly in formative assessment and error analysis.

Overall, the study highlights the need for context-sensitive, research-informed LAL frameworks that move beyond tool use toward a deeper, principled understanding of assessment in early language education across Latin America

Paper 3: Links between teaching and assessment in context: a workshop to develop in-service teachers' (L)AL

Sonia Patricia Hernández-Ocampo (Pontificia Universidad Javeriana)

This presentation gives account of a workshop offered to in-service teachers in a study that aimed at identifying their language assessment needs in their different contexts (rural and urban schools in five regions in Colombia). This workshop was intended to help participants overcome their difficulties in assessment, both as a retribution to their participation in the study and as an attempt to promote LAL among in-service language teachers. Although it was initially offered to teachers of English, it resulted in including all those in-service teachers in the participant institutions who were interested in assessment disregarding the subject they taught. Thus, I will first describe the workshop structure, which responded to the needs that were identified in the study and the challenges we had to face for the implementation. Next, some findings on the interactions that occurred will be presented, as well as an overview of the participants' resulting projects that responded to the design of sound assessment instruments for the different subjects they taught. Finally, the participants' perceptions of the workshop and its usefulness will be addressed.

Paper 4: Mexican young learner EFL teachers and training instructors' perspectives: How useful is online LA training?

Elsa Fernanda Gonzalez (Autonomous University of Tamaulipas)

The study focuses on the perspectives of two instructors and two participants of the course, aiming to understand their views on the course's structure, relevance, and impact on classroom assessment practices. Using data collected through semi-structured interviews, the thematic analysis revealed that the training fostered increased self-reflection among participants and contributed to improvements in how they designed or adapted assessment tools for their teaching contexts.

Despite these positive outcomes, both instructors and participants identified key areas for improvement. Specifically, they highlighted the need to better adapt course content to local realities, manage textbook-based instruction more effectively, and promote more collaborative, trusting relationships between teachers and institutional decision-makers.

The paper concludes by discussing these findings and their implications for future training initiatives in the region. It suggests that while online courses hold promise for enhancing Language Assessment Literacy (LAL), they must be carefully contextualized to align with teachers' realities and foster sustained, meaningful professional growth. It explores how an online language assessment training course was perceived by those directly involved in its delivery and participation. By focusing on participants' lived experiences, this case study emphasizes the value of listening to educators' voices in shaping language assessment-related professional development across Latin America.

Symposium: Connecting Global Assessment Systems and Local Practices: Trust, Affect, and AI in Language Education

Chair: Canadian Association of Language Assessment

Discussant: Samira ElAtia

Day: Friday, June 5th (1:30 p.m. – 3:00 p.m.)

Location: Ovation

Language testing research has traditionally emphasized psychometric concerns such as validity, reliability, and comparability standards, concepts largely developed within global frameworks and applied across diverse contexts. Within these paradigms, affective and emotional dimensions have often been treated as construct-irrelevant variance rather than integral to assessment meaning. Yet assessment is always situated: it unfolds within social, cultural, and institutional contexts that shape how tests are designed, interpreted, and experienced. This symposium explores how emotional, relational, and ethical dimensions of assessment move across scales from global infrastructures and shared models to locally situated practices, revealing how affect and relational trust give meaning to assessment in context. Across four interrelated papers, the symposium examines the interplay between large-scale assessment systems and local educational realities, showing how human relationships, cultural perspectives, and technological mediation reshape traditional understandings of validity and fairness. Together, the studies highlight how relational dynamics and affective processes complicate universal assessment models and offer new ways of understanding locally grounded approaches to testing and evaluation.

The first paper analyzes Japan's internationalization policy (*guroubaru-jinzai*) to show how global mobility programs often overlook socio-emotional and relational competencies that are essential for long-term development. It explores how observable language skills in real-life, local environments challenge globally standardized notions of proficiency and success. The second paper investigates TESL educators' professional identities in AI-enhanced classrooms, tracing how teachers reinterpret their roles and relationships within globally circulating technological infrastructures. Their reflections underscore how relational trust sustains pedagogical integrity amid rapid technological change. The third paper centers on newcomer youth in Canada and illustrates how Learning-Oriented Assessment (LOA) takes on emotional significance when feedback, courage, and positive emotion support learning and belonging within diverse educational contexts. The final paper examines how human-in-the-loop approaches can supplement automated feedback and contribute to a more integrated understanding of how mediated interaction supports the development of higher-order language skills across tasks.

Together, the papers present four vantage points that trace a progression from global policy and assessment design to local enactments involving teachers, learners, and technological systems. This structure underscores how affect and relationality connect institutional testing logics with personal learning experiences while illustrating the multiple scales at which assessment meaning is produced. A recurring theme across all contributions is trust and relational integrity within global and local contexts: for international learners navigating institutional expectations, teachers balancing authority within technological systems, students seeking affirmation through feedback, and practitioners striving to combine fairness with empathy. Affect thus emerges not as interference but as a meaningful part of evidence, showing how fairness, inclusion, and validity are inseparable from the emotional and social dimensions of assessment. By foregrounding the ties between emotional experience, relational trust, and contextual interpretation, this symposium invites reflection on how language assessment can remain both globally rigorous and locally responsive in an era of technological and educational transformation.

Paper 1: Beyond Language Proficiency Scores: Exploring the Socio-Emotional Learning Dimension of Study Abroad

Suzanne Springer (Université du Québec à Montréal), Martyna Kozłowska (Université du Québec à Montréal), Jaime Demperio (Université du Québec à Montréal)

Paper 2: Reframing Teacher Identity and Relationality in the AI Era: A Mixed-Methods Study of TESL Educators' Role Perceptions

Johanathan Woodworth (Mount Saint Vincent University)

Paper 3: Taking a positive turn for the affective dimension of the Learning-Oriented Assessment Framework: "These things give us courage"

Christine Doe (Mount Saint Vincent University), Krista Ritchie (Mount Saint Vincent University)

Paper 4: Responsive Human Mediation and Higher-Order Language Skills Development: A Bayesian Time-Series Case Study

Melissa R. Hunte (University of British Columbia), Nanda Warren (University of British Columbia)

Paper 1: Beyond Language Proficiency Scores: Exploring the Socio-Emotional Learning Dimension of Study Abroad

Suzanne Springer (Université du Québec à Montréal), Martyna Kozłowska (Université du Québec à Montréal), Jaime Demperio (Université du Québec à Montréal)

Recent figures indicate that annually, around 7,000 students from Japan participate in study abroad (SA) programs in Canada, and this number has been steadily increasing. An important motivating factor that encourages universities in Japan to seek SA programs, primarily in English-speaking countries, is the nation-wide effort on the part of the Japanese government to promote *guroubaru-jinzai*, or “global talent”. This initiative aims to cultivate globally competent human resources. It centers on developing a broad worldview and a desire to contribute to society while improving the socio-emotional and communicative abilities required to build cooperative and multicultural relationships. Yet of all these objectives, only language skills are formally measured, and performance on internationally recognized language tests such as IELTS and TOEFL is of particular value for students, universities, and employers. This emphasis on the importance of test scores is evidenced in the selection criteria for SA opportunities and becomes one means through which students can demonstrate their improved ‘global talent’ post SA experience for future study and employment opportunities. However, large-scale international tests have clear limitations in capturing aspects of students’ personal growth that SA experiences foster, such as empathy, adaptability, and collaborative communication - dimensions linked to later life outcomes and civic engagement. As attaining the objectives of global talent programs are not a purely linguistic feat, we explore ways in which alternative measurements might play a role in accounting for the multiple dimensions of intercultural competency and socio-emotional growth that study abroad affords.

Paper 2: Reframing Teacher Identity and Relationality in the AI Era: A Mixed-Methods Study of TESL Educators' Role Perceptions

Johanathan Woodworth (Mount Saint Vincent University)

This study examines how pre-service and in-service TESL educators interpreted the implications of generative artificial intelligence for teacher identity, student-teacher relationality, and pedagogical mediation in language education. Although research on artificial intelligence in education has often emphasized curriculum, efficiency, assessment, and technological adoption, less empirical attention has been given to how teachers understand their professional roles and relational responsibilities in AI-mediated contexts. TESL teacher education programs such as this represent upstream sites of pedagogical formation, where emerging and practising educators negotiate the professional identities, relational commitments, and evaluative criteria that will later shape whether GenAI is adopted, resisted, or critically mediated in classroom and assessment practice. To address this gap, a convergent mixed-methods design was implemented in a graduate-level teacher education course, including a three-week professional development module focused on the ethical and critical use of generative artificial intelligence. Quantitative data were collected through matched pre- and post-intervention surveys with 15 participants, using 30 Likert-type items organized into three domains: teacher identity and role, relationality and trust, and pedagogical mediation and classroom enactment. Analyses using the Wilcoxon signed-rank test indicated significant pretest to posttest shifts across all three domains, with the greatest change evident in participants' confidence in the enduring value of teacher expertise, adaptability, and human judgment. More modest positive movement was also observed in relationality, trust, and pedagogical mediation, although item-level patterns indicated continued concern about depersonalization, reduced face-to-face interaction, and student overdependence on AI. Qualitative survey responses and structured reflections, analyzed thematically, revealed four interrelated shifts: reframing generative artificial intelligence as supportive rather than displacing, reaffirming emotional labour and relational trust as irreducibly human, consolidating teacher identity around judgement and adaptability, and articulating more explicit strategies of boundary-setting, transparency, and critical oversight in pedagogical and assessment contexts. Taken together, the findings suggest that critically oriented professional development can support TESL educators in repositioning themselves not as displaced actors, but as ethical, relational, and interpretive mediators of technological change. The study contributes empirical evidence to current debates on teacher education and generative artificial intelligence by showing that meaningful preparation for AI integration requires more than technical familiarity; it also requires deliberate attention to human connection, pedagogical responsibility, and the conditions under which AI use can be educationally justified.

Paper 3: Taking a positive turn for the affective dimension of the Learning-Oriented Assessment Framework: "These things give us courage"

Christine Doe (Mount Saint Vincent University), Krista Ritchie (Mount Saint Vincent University)

This presentation considers the affective dimension of the Learning-Oriented Assessment (LOA) framework (Turner & Purpura, 2017) from a positive psychology lens (Cabrera & Donaldson, 2023). From a traditional testing perspective, the affective dimension, the emotions and feelings, are the noise contributing to the context-irrelevant variance. This is in contrast to a learning-oriented perspective, where the emotions and basic psychological needs are understood to shape and inform language learning and assessment experiences. In other words, the affective experiences and continuously developing social and emotional skills while learning a language are assets for teachers and students to enhance learning opportunities. The positive psychology framing of the affective domain of the LOA framework will be explained in relation to 'global' or broadly generalizable evidence about human emotion, perception, regulation, and learning from neuroscience in parallel with a complementary explanation of how the subjective experience, language, and stigma components of emotions are shaped by sociocultural contextual complexity. For this study, we interviewed 21 newcomer youth (immigrants, refugees, and international students) through individual and focus groups about their perspectives on their classroom learning and assessment experiences and future goals. The youth's development trajectory emerged as a key finding in how it shaped their articulation of their current and future goals, and the extent they drew on previous educational experiences to inform their perspectives. The youth recognized timely and encouraging feedback as integral when being vulnerable in practicing their language skills, and often expressed positive emotions related to the situations when they saw growth or achievement of their goals. The youth are navigating two educational systems in Canada, adult language learning and/or the K-12 system, and in the presentation, we will consider how the findings relate to these two contexts and how the affective dimension can be expanded from a positive psychology framing.

Paper 4: Responsive Human Mediation and Higher-Order Language Skills Development: A Bayesian Time-Series Case Study

Melissa R. Hunte (University of British Columbia). Nanda Warren (University of British Columbia)

Research suggests that integrating responsive human mediation with automated feedback can support students' reading comprehension (Hunte et al., 2025), particularly language learners who experience difficulty with inferential and discourse-level understanding. These dimensions of comprehension require integrating higher-order cognitive processes, including reasoning, imagining, interpreting, and constructing coherent mental representations (Cain, 1999; Graesser et al., 1994; Kamhi, 2007). Such processes depend on students' ability to draw on prior knowledge, establish connections within and across texts, and extend meaning beyond literal interpretations. From a functional perspective, comprehension also relies on pragmatic knowledge, enabling learners to interpret intended meaning within social and linguistic contexts rather than relying solely on surface-level decoding (Bachman & Palmer, 2010; Hoover & Tunmer, 2020). Responsive human mediation paired with automated feedback have been shown to support the development of these complex language processes by scaffolding learners' meaning-making during interaction (Hunte et al., 2025; Poehner and Infante 2016; Vygotsky 1986).

However, given the dynamic and interactive nature of language development, the effects of such supports are unlikely to be confined to reading comprehension alone. Rather, both the challenges associated with higher-order language processes and the benefits of mediation are expected to extend across broader language use contexts. This study employed a mixed-methods, single-case design to examine these dynamics. Bayesian time-series analysis was used to model the effects of responsive human mediation on students' higher-order language comprehension over time, and qualitative vignettes of student-tutor interactions provided contextualized insights into the observed changes. Findings contribute to a more integrated understanding of how mediated interaction supports the development of higher-order language skills across tasks.

Symposium: Situating Language Proficiencies: Assessment From An Ecological Language Perspective

Chairs: Constant Leung, Eunice Eunhee Jang & Angel M. Y. Lin

Day: Saturday, June 6th (1:40 p.m. – 3:10 p.m.)

Location: Opera AB

The notion of language proficiency in English Language Teaching (ELT) has tended to be understood with reference to the concept of communicative competence in recent decades. Conceptually communicative competence has been adapted from the work of anthropologists such as Hymes (1964, 1974) and Gumperz (1964, 1982). In practice communicative competence has been operationalized in terms of a selection of lexicogrammar associated with normative patterns of discursal features in social use (e.g. Bachman and Palmer, 1996, 2010; GreenBaker, 2021). The selection of lexicogrammatical and discursal features has been primarily premised on conventionalized usages, comprising samples of language use in transactional, professional and educational activities in Anglophone communities. This purview of communicative competence has been the bedrock of ELT teaching and assessment (particularly in large-scale standardized testing) for the best part of fifty years (Leung, 2023; Xi, 2025). Recent research in the fields such as academic literacies, English as a Lingua Franca, flexible multilingualism, and translanguaging has progressively eroded the apparent universality and durability of this approach to modelling language proficiency. On-going research and scholarship in related fields such as community interpreting, health/medical communication, intercultural communication, professional knowledge (in action) and use of AI point to the situated fluidity of language proficiency. The established code-based stability is ill-equipped to deal with the 'leaky' and mutable language boundaries, and the emergent and contingent communicative enactments in situated activities. There is now wide recognition of the fact that communicative activities involve (a) participant speakers' own linguistic and cognitive resources and agentive dispositions, and (b) situated factors such as context, function, task, and relevant ecological affordances (e.g. material environment, AI/technology). The introduction of plurilingual mediation, in its myriad manifestations, as part of language proficiency in the 2020 Companion Volume of the CEFR has crystalized the need for conceptual development that would feed productively into language assessment and language education more generally. At this time, a focussed effort to expand the notion of language proficiency through the lens of assessment and testing is also likely to make a timely and important contribution to both theory and research in Applied Linguistics and language education more generally. In this proposed symposium we will report work-in-progress in four areas: the onto-epistemological challenges to the prevailing approach to conceptualizing communicative competence and their direct impacts on the assessment of language proficiency and language education more generally; the arguments for moving away from a synthetically packaged language code-based view of language use to a concept of Ecological Languageing Competencies (ELC) in which communicative co-action is regarded as primary datum for construing language proficiency; the empirical issues of identifying, describing and analyzing ELC episodes; this discussion will report the conceptual and operational questions arising from working with non-contrived classroom-based speaking data for assessment; the methodological and operational issues associated with adopting an AI-based process-oriented diagnostic writing assessment for assessing and tracking ELC in elementary classrooms; machine learning models and Bayesian progression-tracking systems will be considered in relation to charting developmental trajectories and capturing the uncertainties inherent in student enactments of ELC longitudinally.

Paper 1: Language Proficiency – From the Global to the Situated

Constant Leung (King's College London)

Paper 2: Ecological Languageing Competencies (ELC) – Localizing Assessment

Angel M.Y. Lin (The Education University of Hong Kong), Paul J. Thibault (The Education University of Hong Kong)

Paper 3: Glocalizing Language Assessment: Redesigning EAP Speaking Rubrics Through The Ecological Languageing Competencies (ELC) Lens

Peichang He (The Education University of Hong Kong), Paul J. Thibault (The Education University of Hong Kong), Wen-Li Tsou (National Cheng Kung University), Fay Chen (National Cheng Kung University), Angel M. Y. Lin (The Education University of Hong Kong)

Paper 4: Assessing and Tracking ELC through AI-Based Process-Oriented Writing Assessment

Eunice Eunhee Jang (University of Toronto)

Paper 1: Language Proficiency – From the Global to the Situated

Constant Leung (King's College London)

The focus of this presentation is on analyzing and expanding the conceptualization of English Language proficiency in ELT and language assessment. For the best part of fifty years, the central core of English-as-an-additional/second language proficiency has been underpinned by the concept of communicative competence. Initially developed from ethnographically oriented empirical observations of situated language in use, the concept of communicative competence has been morphed into an orthodoxy of curricularized language specifications, modelled on conventionally valued (and normed) varieties of language use in English-speaking educational, institutional and professional settings, for assessment, teaching and learning generally. The discussion will first trace the recontextualization and reification processes that have transformed a research orientation to curricularizing selections of valued language samples through a historical lens. After that the discussion will turn to the more recent research findings and theorizations in fields such as academic literacies, English as a Lingua Franca, flexible multilingualism, translanguaging, and use of technology to show the need for language assessment to adopt a more situated and dynamic approach to agentive and fluid language communication and language proficiency; the inclusion of volitional and contingent communicative acts such as plurilingual mediation in the Companion Volume of the CEFR (Council of Europe, 2020) unambiguously calls for a thorough re-consideration of the prevailing psychometrically informed paradigm in language assessment, with particular reference to concepts such as construct, validity and fairness in the designing and validation of standardized language testing (Leung, 2023; Leung and Jenkins, 2020). This part of the symposium is intended as a reflexive response to the looming challenges to language assessment, both theoretically and operationally. This presentation will conclude with some remarks on the directions for reframing our approaches to language assessment and language education more broadly.

Paper 2: Ecological Linguaging Competencies (ELC) – Localizing Assessment

Angel M.Y. Lin (The Education University of Hong Kong), Paul J. Thibault (The Education University of Hong Kong)

This presentation explores the relevance, applicability and potential affordances of the concept of Ecological Linguaging Competencies (ELC) for language assessment. Linguaging is something that we do with others rather than just a performance of static pre-existing set of forms that we use. Linguaging is a social mode of (inter)action, or communicative co-action, which has specific situated characteristics (Leung et al, under review). The concept of Ecological Linguaging Competencies (ELC) comprises two dimensions: (1) the “offline” (i.e., learned, habituated) ability of persons to create utterances that serve their own and others’ purposes in languaging activity (part of real world activities); and (2) the “online” (i.e., real-time, moment-to-moment) participation in and management of the flow of languaging. ELC requires the achievement of functional fits between the learned capacities and skills of languaging selves and the affordances (artefacts, tools, technologies, objects, etc.) of the environments (taskscape, situations, activities and practices) in which individual participants are required to contribute productively to the flow of languaging. Current formulations of communicative competence focus on the individual and seek to assess the competence of the individual in terms of language code-based outputs of typified communication activities in (implicitly) monolingual English-speaking settings. However, languaging is always co-production in different ways that cannot be entirely pre-scripted. If we assume that the individual is the sole locus in the ways in which we conceptualize language and language use in real world settings, it would be virtually impossible to take account of the nature of agentive, contingent and sentient languaging in sociality. ELC emphasizes languaging as ecologically embedded and embodied co-participation and co-production whereby persons coordinate with other persons and with relevant aspects of their worlds. This discussion will provide examples of how ELC works as units of analysis for pedagogy and assessment in classroom activities (Thibault et al, forthcoming).

Paper 3: Glocalizing Language Assessment: Redesigning EAP Speaking Rubrics Through The Ecological Linguaging Competencies (ELC) Lens

Peichang He (The Education University of Hong Kong), Paul J. Thibault (The Education University of Hong Kong), Wen-Li Tsou (National Cheng Kung University), Fay Chen (National Cheng Kung University), Angel M. Y. Lin (The Education University of Hong Kong)

For decades, English-as-an-additional-language proficiency has been abstractly framed by a putatively globally applicable concept of communicative competence and operationalized as code-based, monolingual outputs, often reinforced through high stakes testing—practices that can constrain learner agency and hamper performance. This presentation reports a study that redesigns English for Academic Purposes speaking assessment through the Ecological Linguaging Competencies (ELC) lens (The New Territories Group, forthcoming), which foregrounds situated co action with people and artefacts, whole body sense making, and the alignment of teaching, learning, and assessment.

Using a longitudinal mixed methods design, we developed and operationalized an ELC informed rubric for academic presentations. An initial rubric, drafted from ELC principles and stakeholder consultations, was trialed by linguists/researchers on 60 student presentation videos and iteratively refined until descriptor saturation. Subsequent expert review and teacher workshops produced further adjustments before classroom implementation. Surveys and interviews with assessors and students examined usability, validity in practice, and perceived impact on teaching and learning.

Findings highlight three contributions to plurilingual, ecological assessment: (1) Glocalization—criteria aligning with local classroom practices and learners’ lived, plurilingual repertoires, shifting emphasis from statically normed, delimited contextual language and topic coverage to action-oriented, ecological, embodied participation, and mediation (see CEFR Expert Group, Council of Europe, 2023); (2) Theoretical significance—producing evidence that ELC offers a viable alternative to code-based conceptualization of communicative competence for EAP speaking assessment in higher education; (3) Professional learning—co constructing the rubric across linguist–researcher–practitioner communities fostered shared understanding, calibration, and sustainable assessment literacy.

The study shows how ELC reframes EAP speaking rubrics from measuring a static “construct” to supporting learners’ agentive development in situated communication. By integrating ecological conditions of tasks, tools, and interlocutors into observable descriptors, the rubric makes assessment a driver of learning rather than a gatekeeper, advancing the symposium’s aim to reconceptualize proficiency as situated languaging.

Paper 4: Assessing and Tracking ELC through AI-Based Process-Oriented Writing Assessment

Eunice Eunhee Jang (University of Toronto)

This presentation reports a case study of using AI-based process-oriented diagnostic writing assessment for tracking ELC in elementary classrooms. Drawing on the writing assessment module of BalanceAI, a digital assessment platform, I illustrate how students engage in an integrated sequence of tasks—watching a stimulus video, planning ideas, generating self-questions, drafting multiple versions, receiving instant AI-powered feedback, and reflecting through self-assessment (Ignacio et al., 2024). These activities elicit not only written output but also the cognitive and metacognitive processes that underlie ELC's languaging in action, such as monitoring, self-regulation, and agentic decision-making in communicative acts. Empowered by AI, data of students' planning, reflective monitoring, and goal settings across writing process are automatically collected and analysed.

Unlike traditional product-focused writing assessment, this approach positions writing as situated languaging—an embodied, agentic, and socially embedded process. By automating assessment of fine-grained features across multiple drafts, AI tools provide dynamic and ecologically valid evidence of students' developing competencies. Machine learning models and Bayesian progression-tracking systems further allow us to chart developmental trajectories and capture the uncertainties inherent in student enactments of ELC over time (Hannah et al., 2023). The discussion highlights how AI-based diagnostic assessment can enhance both learner and teacher agency. Learners gain opportunities to reflect on their writing decisions, thereby strengthening metacognitive growth and ownership of learning. Teachers, in turn, receive actionable, real-time insights to differentiate instruction in ways that are sensitive to the situational and ecological contexts of writing. I will conclude by considering theoretical, practical, and ethical issues, including authenticity of tasks, developing defensible performance descriptors, and fairness in AI scoring. Together, these discussions show how AI-infused diagnostic writing assessment can be integrated into the ELC framework by embedding assessment in the dynamic flow of classroom learning.

Symposium: Navigating Inclusivity: How Can We Balance Global and Local Challenges in Assessment Practices?

Chairs: Carla Pastorino-Campos, Amy Devine & Evelina Galaczi

Discussant: Lynda Taylor

Day: Saturday, June 6th (1:40 p.m. – 3:10 p.m.)

Location: Symphonie 1

Representation has become an increasingly important topic in second/foreign (L2) language assessment, taking centre stage in discussions of inclusivity and ethical test design. A quintessential aspect of inclusive test design is (or should be) adequate representation of the diversity of people in the target language use domain. Inclusive representation has been promoted in policy decisions (International Language Testing Association, 2024). Despite general agreement at policy level that representation and the broader concept of inclusivity should play a key role in assessment (and learning), the practice of inclusivity can be a complex endeavour underpinned by highly contextual ideologies. Even when policies and theoretical principles promote the ideal of inclusive representation in assessment materials, historical, political, cultural and socio-economic factors may covertly or explicitly present barriers to implementation and shape decisions about assessment design and materials. This tension between global policy imperatives and local contextual implementation barriers means that test designers and constructors have to manage the finelytuned balancing act of translating global principles into the complex realities of local contexts. Systematic guidance and published research problematising this challenge is rare in L2 assessment and case studies tackling the theory-into-practice complexities are virtually non-existent. In this symposium we aim to address this important area by focusing on considerations and practices of inclusive representation in L2 assessment. We draw on recent research to explore the following questions: What does good practice in inclusive representation in L2 assessment look like? What are the considerations and barriers to implementing inclusive representation in local contexts? How can we lay the groundwork for principled implementation of representation in L2 assessment while taking into account local contextual constraints? These questions constitute the organising principle behind the symposium and are addressed in papers that critically engage with key issues. Common themes are briefly synthesised in an introduction which sets the scene by outlining key debates arising from global/local tensions in representation and inclusivity in L2 assessment. This is followed by a paper delineating best practice regarding inclusive representation in language assessment and providing an overview of potential barriers to implementation. Specific contexts are then explored in three case studies which provide direct insights into the tensions between policy and implementation. These case studies focus on (i) the introduction of different types of representative materials in tests intended for a global audience; (ii) the considerations which emerge when incorporating Indigenous topics into the testing materials of higher-education institutions; and (iii) deliberations in creating EDI-compliant content for young learners. The symposium ends with a reflection from a discussant and time for audience discussion. Through this symposium we aim to bring together a diversity of opinions and experiences offered by scholars and practitioners from five different contexts who bring complementary theoretical and practical expertise in understanding and problematising representation in L2 assessment. That spectrum of views will, we hope, provide an opportunity for both audience and presenters to engage with new insights to benefit our field.

Paper 1: Whose voices count? Representation, agency, and language assessment

Nahal Khabbazzbashi (Centre for Research in English Language Learning and Assessment, University of Bedfordshire)

Paper 2: (Trying to) Do the right thing: Constraints to representation in international English language tests

Carla Pastorino-Campos (Cambridge University Press & Assessment), Amy Devine (Cambridge University Press & Assessment)

Paper 3: Language test content on Indigenous topics: Guidelines for item writers

Janine Metallic (McGill University), Beverly Baker (University of Ottawa)

Paper 4: Creating EDI-compliant content for young learners: Hurdles and discoveries

Gemma Bellhouse (IDP Education)

Paper 1: Whose voices count? Representation, agency, and language assessment

Nahal Khabbazzbashi (Centre for Research in English Language Learning and Assessment, University of Bedfordshire)

Hall (1997) makes a fundamental distinction between reflection and representation, the latter involving ‘the active work of selecting and presenting, of structuring and shaping; not merely the transmitting of an already-existing meaning, but the more active labour of making things mean’ (Hall, 1997: 15). Building on this perspective, this paper frames language assessments as cultural sites where agency, ideology, economy, and politics intersect and offers an overview of key research strands on representation in language learning and assessment.

Drawing on critical applied linguistics, I synthesise work exploring representational meaning through multiple lenses: gender and sexuality (e.g. Gray, 2013, Keleş et al. 2023), race (Bowen & Hopper, 2023), class (Gray & Block, 2014), cultures and communities (Jenkins & Leung, 2019), and ideology (Noori & Mirhosseini, 2021). Across these contexts, three recurring themes emerge. Firstly, representations consistently reproduce exclusions: e.g., gender stereotypes, absence of LGBTQ+ figures, racial imbalance privileging whiteness, and the dominance of male, white, middle-class identities. Secondly, assessment designers may adopt ‘safe zone’ strategies that create an ‘artificial sense of inclusion’ while avoiding issues salient to local realities and experiences (Selvi et al., 2021). Thirdly, much scholarship, while incisive, remains at the level of critique, leaving many questions unresolved.

In international exams, whose ideologies and identities are foregrounded? Should assessments reflect test takers’ demographics or the cultural norms of the target country? If diversity is acknowledged, which accents, voices, and communities are privileged or deemed legitimate? To what extent are inclusion and representation genuine concerns – or are they ultimately at the service of the million-dollar TESOL industry?

This paper sets the scene, inviting reflection and dialogue on how we might move beyond critique to consider whose agency is at work in assessments and what inclusive and equitable assessments might look like.

Paper 2: (Trying to) Do the right thing: Constraints to representation in international English language tests

Carla Pastorino-Campos (Cambridge University Press & Assessment), Amy Devine (Cambridge University Press & Assessment)

English language tests intended for global distribution are often designed and produced in contexts where policies and legislation typically include some basic aspects of inclusive education. This includes countries in the Global North, such as the UK, USA or Canada, which tend to adhere to global EDI policies (e.g. from UNESCO) or international professional codes (e.g. from ILTA).

However, many of the language tests produced in these contexts are intended to be taken by test takers all over the world and, for this reason, need to acknowledge the preferences of a very diverse audience that may have equally diverse views regarding the need for inclusive education. The tension between the desire to produce inclusive testing materials and the need to commercialise them in markets with varying understandings of inclusion (and sometimes explicit legislation against it) is the focus of this presentation.

We will explore two cases where the introduction of inclusive materials has faced barriers of a very different nature: 1) the inclusion of Global Englishes accents for listening tasks, which has typically been resisted based on potential impacts on the psychometric properties of the test (e.g. the shared-L1 effect) but is now more generally accepted; and 2) the inclusion of non-heteronormative content in reading tasks, which has been almost summarily opposed due to risk of a negative, morality-based reaction from more conservative social contexts.

Constraints can include logistical challenges and operational costs associated with developing representative content, concerns about exam uptake and the difficulty of introducing innovation while still maintaining continuity in language assessment. These will be discussed from the perspective of the test provider conundrum, which seeks to balance the need to deliver tests that have broad appeal for a variety of consumers with the imperative to produce high quality, ethically designed tests.

Paper 3: Language test content on Indigenous topics: Guidelines for item writers

Janine Metallic (McGill University), Beverly Baker (University of Ottawa)

In this presentation, we describe a collaboration between a non-Indigenous test developer and an Indigenous scholar to improve inclusivity in a university language testing office. Specifically, we focus on the development of a checklist for the responsible inclusion of Indigenous content in our testing materials.

Guided by principles of relational accountability (Wilson, 2001, 2008) and critical language testing (Shohamy, 2001; Schissel, 2019), we worked to decolonise not only the content development process but also our partnership. Drawing on literature that emphasizes the importance of relationships as the foundation of research in Indigenous contexts (Gordon & Around Him, 2025), we adopted a partnership model to shape how we built trust, addressed power dynamics, and translated theory into practice. We also drew on calls for “culturally responsible use of language assessment and evaluation” (Schissel, p. 7) to ground our efforts.

The outcome of our collaboration was a practical, easy-to-use checklist for developing reading and listening test passages. One item on the checklist, for example, advises developers to avoid overgeneralisation by not misapplying traditions of one group to all Indigenous people (also called “pan-indigenisation”). We subsequently trialled the checklist with a listening comprehension task, which revealed “blind spots” in the office’s prior coverage of Indigenous topics, despite good intentions and commitment to existing international guidelines and practices. These findings underscored not only the value of the checklist but also the ongoing need to interrogate colonial legacies in language testing. The new checklist has since been incorporated into content development guidelines.

This process illustrates how language testers can overcome barriers in implementing Indigenous content to faithfully reflect their local context in assessment materials. By sharing both our collaborative approach and the resulting checklist, we aim to offer a model that can inform more culturally responsible assessment practices in other local settings.

Paper 4: Creating EDI-compliant content for young learners: Hurdles and discoveries

Gemma Bellhouse (IDP Education)

Assessments for young learners must prioritise creating a positive experience, offering children opportunities to demonstrate their capabilities while fostering a sense of progression and achievement (Papp & Walczak, 2016; Rea-Dickins, 2000). The role of technology is continually increasing in early formal and informal learning and can enhance engagement and enjoyment, but also introduces challenges to young learners, such as increased cognitive demands and test anxiety. Digital platform functionality and other Equality, Diversity and Inclusion (EDI) related compliant criteria specific to young learners (e.g. authority representation, storyline safety and role responsibilities) must be addressed in language test policies, and the test content and test delivery must be monitored for policy regulation.

To ensure test content for young learners is compliant with EDI values, one global organisation and international test provider developed a new section for an existing, internal EDI test content creation policy. The section provides clear and practical guidance for item writing and is relevant for language test production and quality assurance. The young learners’ section, implemented in early 2025, includes three parts. The first is ‘Platform functionality: motivation & attention’ and outlines inclusive guidelines for test piloting and delivery. The second part ‘Appropriate topics and vocabulary’ introduces the new acronym ‘TURNIPS’ to guide content creators to carefully treat or completely avoid the following topics: Threat, Unfairness, Responsibility, Negligence, Inequality, Profanity, and Self-consciousness. The third part ‘Representation and role modelling’ speaks to character representation and the avoidance of negative stereotypes.

This case study presentation will describe the project stages and the organisation’s EDI Mainstreaming Guide tool used to structure the panel discussions. The project outcomes and lessons learned during the development of the policy will also be shared, such as the importance of policy implementation by design and ensuring diversity and expertise in the panel.

Symposium: Refining Principles and Practice in Score Concordance Research

Chair: Sarah R. Hughes

Discussant: Sara T. Cushing

Day: Saturday, June 6th (1:40p.m. – 3:10 p.m.)

Location: Symphonie 2

Score concordance studies play a vital role in shaping public understanding of score meaning and policy surrounding score requirements. The primary purpose of score concordance studies is to provide information to score users, such as institutions and governments, seeking to set fair and consistent score requirements in an environment where multiple tests are used for the same purpose across varied national and local contexts. In practice, these studies also serve a broader audience, encompassing test takers, teachers, and the general public, who may use concordance studies to understand the relationships between tests. In the past year, multiple large-scale score concordance studies have been published exploring the relationship between scores on high-stakes English language proficiency tests. All studies adhered to good practice principles established by Knoch & Fan (2024), resulting in several high-quality studies that represented considerable investment from testing organisations. This symposium will feature presentations from the lead authors of three concordance studies (Hughes & Clesham, 2025; Ikeda et al., 2025; Lampropoulou et al., 2024), who will discuss the insights they gained through the research process. First, an empirical investigation of variables impacting concordance precision builds on the Knoch & Fan guidelines, offering insights into the elements that can be flexed to accommodate different contexts and uses of concordance information. The second paper similarly aims to propose revisions to or expansion of these guidelines, highlighting considerations for the use of concordance information in decision-making. The final paper emphasises the importance of contextualising concordance research through discussions of construct comparability and guidance on the interpretative limits of concordance data to support fair and valid uses of test score comparisons. A panel discussion with the study authors will consider future directions for concordance research and how language testing organisations can collaborate to better support both score users and test takers in understanding and using concordance research findings. The “year of the concordance study” presents a unique opportunity to gather a wealth of expertise to reflect on the principles and practice of score concordance research, as this important topic in language assessment continues to impact real-world policy decisions and opportunities for test takers.

Paper 1: An Empirical Evaluation of Good Practice Principles for Concordance Studies

Sarah R. Hughes (Pearson), Anastasia Ulicheva (Pearson)

Paper 2: Establishing Fair Score Requirements across Language Proficiency Tests: Reflections on Score Concordance Research

Spiros Papageorgiou (ETS), Emma Bruce (British Council), Tony Clark (Cambridge University Press and Assessment)

Paper 3: From Construct to Communication: Rethinking Good Practice in Concordance Research

Leda Lampropoulou (LANGUAGECERT), Johnathan Jones (CRELLA, University of Bedfordshire)

Paper 1: An Empirical Evaluation of Good Practice Principles for Concordance Studies

Sarah R. Hughes (Pearson), Anastasia Ulicheva (Pearson)

The good practice principles for conducting concordance studies established by Knoch & Fan (2024) provide the first set of comprehensive criteria researchers should seek to meet when conducting concordance studies. The principles, which have been influential in recent concordance studies, are theoretically grounded in prior research and professional standards for educational evaluation and language assessment. They are also practically difficult to achieve, requiring significant investment to meet in full. This research seeks to build on the good practice principles by empirically investigating the impact of specific principles on the stability and accuracy of concordance. The findings of this research distinguish between design choices that are essential for fairness and validity, and those that may be optional or adjustable depending on resources and proposed uses of concordance.

Expanding on the recent large-scale concordance study between PTE Academic and IELTS Academic (Hughes & Clesham, 2025), this study investigates the stability of concordance under a variety of conditions. The PTE-IELTS concordance data set (N = 1,522 participants) was analysed to systematically test how concordance stability changes with different sample sizes, counterbalancing designs, and time-between-tests intervals. Controlled simulations were used to establish the conditions needed to detect practically meaningful differences in concordance outcomes.

As concordance becomes increasingly important in the interpretation and use of test scores as entry requirements, it is essential that concordance methodology continues to evolve. Ultimately, this project contributes to the methodological foundations of concordance research, offering new evidence to refine and improve good practice in the field and support the responsible use of concordance research as one component of setting fair and consistent score requirements across distinct tests.

Paper 2: Establishing Fair Score Requirements across Language Proficiency Tests: Reflections on Score Concordance Research

Spiros Papageorgiou (ETS), Emma Bruce (British Council), Tony Clark (Cambridge University Press and Assessment)

Drawing on our experience conducting a collaborative score concordance study between the IELTS Academic and TOEFL iBT tests (Ikeda et al., 2025a; 2025b), we review the evaluative framework proposed by Knoch and Fan (2024), explore methodological best practices and challenges, and offer future directions for score concordance research. We emphasize the importance of methodological rigor in collecting test taker score data, transparency in analysing such data to build score concordance tables, and construct comparability as a prerequisite for conducting a score concordance study, while also highlighting the limitations of concordance tables as standalone tools for admissions decisions. We note that some aspects of Knoch and Fan's (2024) guidelines are more straightforward to implement in practice than others, and that some features might be updated or adjusted after real-world application, further contributing to best practice for concordance researchers going forward. We conclude with recommendations for decision-making that are based on fair score requirements irrespective of which test is selected by test takers, and suggestions for rigorous test provider collaborative research in the area of score and test comparison.

Paper 3: From Construct to Communication: Rethinking Good Practice in Concordance Research

Leda Lampropoulou (LANGUAGECER), Johnathan Jones (CRELLA, University of Bedfordshire)

Concordance research increasingly shapes how test users, institutions, and policy makers interpret and apply test scores. The value of such research, however, depends not only on robust statistical relationships but also on clear evidence of construct comparability and on transparent communication of findings. The LANGUAGECER–IELTS concordance study illustrates how these dimensions interact to support interpretive validity, showing that meaningful concordance emerges from the interaction of empirical rigour and transparent stakeholder communication.

Evidence of content comparability and observed correlations jointly underpin the technical defensibility of concordance relationship. Against this backdrop, we discuss how these layers of evidence define the interpretive boundaries within which concordance tables can be used meaningfully. Attention then turns to how concordance findings are communicated to external stakeholders to ensure understanding aligns with the empirical limits of the data. Particular focus is given to challenges of preventing over-interpretation or misrepresentation of concordance tables in public discourse.

By linking methodological integrity with communicative responsibility, the contribution calls for a conception of “good practice” in concordance research which recognises that statistical precision and transparent interpretation are equally essential for supporting fair and valid score use across assessments.

Symposium: Mid-level Actors in Global–Local Negotiations

Chair: Laura Schildt

Discussant: Bart Deygers

Day: Saturday, June 6th (1:40p.m. – 3:10 p.m.)

Location: Symphonie 3

Language assessment is shaped by macro-level policies and global frameworks such as the Canadian Language Benchmarks (CLB) and the Common European Framework of References for Languages (CEFR). However, in both highly centralized and decentralized systems mid-level actors must interpret and apply these policies and frameworks taking into account contextual demands. Mid-level actors are individuals straddling both policy design and delivery such as middle-management policy officials, program coordinators, language assessment experts and specialists and test administrators. Researchers are just beginning to explore the central role of these individuals. While some attention has been paid to classroom teachers (Menken & García, 2010), more work is needed to understand the unique role played by these individuals as they mediate between macro policy priorities and local realities.

The presentations in this symposium combine theory and practice. They report on the real-world challenges and experiences of people bridging the policymaking and implementation spheres. The presentations apply conceptual frameworks as a basis for understanding these experiences. In this way they advance both empirical insights and theoretical innovation by demonstrating how local groups and individuals do not merely transmit macro assessment policies; they actively negotiate between universal scales and locally situated needs. Their success as assessment practitioners depends not only on their beliefs about language, but also their agency, the policy environment and institutional support.

The contributions to this symposium span four continents to provide case studies from settings where language testers respond to policy mandates. The two-hour program features four presentations (fifteen minutes each with five minutes for questions) which report on empirical studies in different language assessment contexts. The first presentation highlights administrators, teachers, and policy officers in the implementation of the National Matriculation Test (NMT) language policy in China. The second presentation brings into focus language test developers responding to the Costa Rican Ministry of Education's Alliance for Bilingualism language policy. The third presentation shines a light on the policy advocacy work of Australian language specialists and macro-level language policy. The fourth presentation addresses Austrian school headmasters' policy implementation with regards to the language test for German Language Support Classes. Taken as a whole, these contributions explore new directions in theorising language assessment by foregrounding actors often overlooked. Collectively, they form a compelling and coherent narrative of the soft power and important voice of mid-level actors in in shaping language testing practices.

Paper 1: Policy Intentions and Reality: How Do Meso-Level Stakeholders Shape Language Assessment in Practice?

Chenyang Zhang (University of Melbourne), Jason Fan (University of Melbourne)

Paper 2: The Power of Context: Rethinking Language Proficiency Testing in Costa Rica

Ana Carolina González Ramírez (University of Costa Rica)

Paper 3: Context Sensitive Policy Literacy: Insights from Australian Language and Assessment Specialists

Laura Schildt (Ghent University)

Paper 4: Subordinate Strawman or Arbiter with Agency? School Headmasters as Funnel of Language Assessment Policies

Benjamin Kremmel (University of Innsbruck), Eva Konrad (University of Innsbruck), Viktoria Ebner (University of Innsbruck)

Paper 1: Policy Intentions and Reality: How Do Meso-Level Stakeholders Shape Language Assessment in Practice?
Chenyang Zhang (University of Melbourne), Jason Fan (University of Melbourne)

Under China's Belt and Road Initiative, multilingual competence has been positioned as critical in fostering international collaboration and has gained increasing prominence in China's foreign language education. Against this backdrop, this study examines how meso-level policy actors in senior secondary schools interpret and appropriate national language policies and how their agency shapes language education and assessment. The focus of this study is the Languages Other Than English (LOTE) subject in the National Matriculation Test (NMT), one of China's most high-stakes examinations. Drawing on Johnson and Johnson's (2015) funnel model, this case study investigates two senior secondary schools in Shanghai with contrasting curricular orientations: one emphasizing LOTE and the other English-focused. The study was conducted in two stages. First, 25 national policy documents on LOTE education and assessment were analysed to trace policy intentions. Second, semi-structured interviews were conducted with two administrators, four LOTE teachers, and one officer from the local educational examinations authority responsible for developing the LOTE subject in the NMT. The data were thematically analysed using grounded theory. Findings reveal that while both schools prioritize student success in the NMT, their strategies differed according to institutional context. Administrators in both schools sought to enhance LOTE education for test preparation, drawing on external proficiency tests as benchmarks and collaborating with universities to address resource gaps. Teachers exercised agency by adapting higher education textbooks and designing classroom assessments despite limited resources. Yet important differences also emerged: the LOTE-oriented school supported teacher collaboration in curriculum and pedagogy through institutional structures, whereas the English-focused school relied more on external experts, with teachers independently seeking resources and focusing primarily on test preparation. This study highlights the importance of a meso-level perspective on policy appropriation, demonstrating how actors' agency, institutional resources, and policy intentions collectively shape the implementation of LOTE education and assessment.

Paper 2: The Power of Context: Rethinking Language Proficiency Testing in Costa Rica
Ana Carolina González Ramírez (University of Costa Rica)

This paper argues that localized, context-sensitive assessment design can reconcile international benchmarking demands with national educational realities. Within an increasingly globalized educational landscape, Costa Rica's Ministry of Public Education launched the Alliance for Bilingualism, an initiative aimed at ensuring that secondary school graduates achieve a B1 English proficiency level according to the Common European Framework of Reference for Languages (CEFR). To support this goal, the Ministry restructured the national English curriculum in 2016 and identified the need for a large-scale proficiency assessment to measure learning outcomes. However, available international exams were limited in number, financially inaccessible to most students, geographically restricted, and insufficiently aligned with the communicative and contextual emphases of the renewed 2018 curriculum. In response, the School of Modern Languages at the University of Costa Rica assumed the responsibility of designing and administering a localized, cost-effective proficiency test that would uphold the CEFR's communicative principles while addressing the country's specific sociolinguistic and socioeconomic conditions. The resulting Language Proficiency Test (Prueba de Dominio Lingüístico; PDL) integrates internationally recognized assessment practices within a performance-based, digital format adaptable to varying connectivity levels across the country. The PDL embodies collaborative agency among policymakers, university experts, and educators, balancing global standards with local curricular content. Initially limited to receptive skills, the test has undergone continuous refinement and now includes an AI-assisted oral component, widely accepted by national employers as evidence of speaking proficiency. Findings from the 2022 administration revealed that students from urban public schools consistently outperformed those from rural areas, highlighting deep-seated disparities in resource allocation, infrastructure, and teacher training. The PDL experience illustrates how localized assessment initiatives can operationalize global benchmarks in equitable and contextually meaningful ways, promoting both national accountability and social inclusion in language education.

Paper 3: Context Sensitive Policy Literacy: Insights from Australian Language and Assessment Specialists

Laura Schildt (Ghent University)

There is a pressing need within the field of language testing and assessment to engage in substantive, contemporary debate on the practice of policy literacy. It is now well-established that test developers, researchers and other stakeholders involved in the process of designing and using tests require policy literacy (Deygers & Malone, 2019; Fulcher & Harding, 2022; Lo Bianco, 2019). This study examined what policy literacy means in the professional lives of language experts, and how their perspectives can inform a more context-sensitive conceptualization of the term. These questions were explored in qualitative interviews with 21 Australian language experts engaged in policymaking. The thematic analysis of this data supports a contextualized view of policy literacy. The findings indicate that meso-level actors mediate between policy aims and policy impacts on the ground. Ricento and Hornberger's (1996) onion model of policymaking, and Johnson and Johnson's (2015) expansion, highlight the mediating role of meso-level actors in shaping how language policies are interpreted and implemented locally. Traditionally, this mediation has been examined through the lens of language beliefs. This presentation argues that a view of language ideologies alone is insufficient in understanding how language policies are developed and enacted. Instead, the policy literacy of meso-level actors—their capacity to understand policy discourses, navigate institutional processes, and exercise agency and soft power—also plays a decisive role in how language policies are made and interpreted. Furthermore, the results reveal that the scope of policymaking is changing (Howlett et al. 2025) and the type of knowledge and skills that were considered relevant in the past are being pushed out to the periphery of policymaking, replaced by a greater diversity of voices and policy complexity.

Paper 4: Subordinate Strawman or Arbiter with Agency? School Headmasters as Funnels of Language Assessment Policies

Benjamin Kremmel (University of Innsbruck), Eva Konrad (University of Innsbruck), Viktoria Ebner (University of Innsbruck)

The Austrian curricula for primary and secondary schools include regulations on German Language Support Classes (GLSC) to foster the acquisition of the country's official language amongst children with first languages other than German (Konrad et al., 2024). The main stated objective is to support pupils whose language skills do not enable them to follow regular classes (BGBl. II 230/2018, BGBl. II 235/2019). The policy, however, has been subject to controversy from the start, with critics stating that such a segregated approach which singles out individual students for support classes and removes them from their regular classroom environment is detrimental for both emotional/motivational as well as academic development and does not yield the anticipated language learning benefits (Erling et al., 2021; Flubacher, 2023; Schwab et al., 2024). The allocation of students to GLSC is made on the basis of a language test. The administration of this test, as well as the decision as to who needs to take it in the first place, (i.e. whether they fall into the category of not being a L1 German speaker), is the responsibility of the school headmasters, who receive a brief training on the matter, but are not necessarily second or foreign language (learning) specialists. This pilot study therefore aims to explore how headmasters implement the policy and use or expand the agency provided to them regarding this policy. Interviews with six school headmasters from primary and secondary schools in two provinces were conducted, transcribed and thematically coded to shed light on how these language policy arbiters (Johnson & Johnson, 2015) navigate potential tensions between the policy directive and their own attitudes towards the policy. Findings show that these individuals, although relatively low in the overall policy hierarchy, still play a critical role with some agency in the implementation of the educational language policy.

Symposium: Global Advances and Local Applications: Expanding L2 Classroom Assessment in the Age of GenAI

Chairs: Vera Vesala & Dmitri Leontjev

Discussant: Erik Voss

Day: Saturday, June 6th (1:40p.m. – 3:10 p.m.)

Location: Ovation

The rapid increase in the use of Generative Artificial Intelligence (GenAI) has produced profound changes in various aspects of society, including education. Research has documented its impact on language learning, teaching, and assessment, as illustrated by a large-scale survey of English teachers (Edmett et al. 2024). Learners and particularly teachers' attitudes towards AI range from seeing its undeniable benefits to noting the serious challenges it has brought to the classroom, such as threats to academic integrity (Sütçü & Sütçü, 2023). What regards classroom assessment, the novel affordances and tensions introduced by GenAI interact with two notable developments in educational assessment at the turn of the century: a growing focus on assessment supporting learning, building on the work of such scholars as Black and Wiliam (e.g., 1998; see Leontjev et al., 2025) and criticality introduced to educational assessment (e.g., Shohamy, 2001). These developments drew the scholars' attention to the consequences of assessment, educational equity, and recognition of teachers and learners as actors to be closely involved in exploring ways assessment can support teaching and learning. Diagnostic (DiagA) and dynamic assessment (DA) frameworks, informed by these developments, are particularly relevant to this symposium, as the symposium papers are to various degrees informed by one or both frameworks. Building on SLA research, L2 DiagA seeks to identify learners' strengths and weaknesses and provide actionable feedback to learners and teachers. Unlike DiagA, assessing learners' independent performance, L2 DA, rooted in Vygotskian sociocultural theory (Vygotsky, 2012), aims at identifying learner abilities in the process of developing by including a teaching component as a part of the assessment process.

The symposium serves as a forum to discuss the tensions and affordances that GenAI proliferation as global societal change has created in the local contexts of L2 classrooms, and ways in which AI in language assessment can support and promote language learning. The symposium includes a ten-minute introduction by the organisers and four papers drawn from China, Finland, and the United States, each lasting fifteen minutes, followed by a short discussion. Informed by the L2 DiagA, the discussion starts by building an argument for expanding learner feedback literacy in light of AI-generated actionable feedback on learner writing. It then moves to the L2 DA framework, exploring the potential of ChatGPT in diagnosing and promoting the development of learners' argumentative writing ability. Continuing from this, derived from the project that combines DiagA and DA, the following paper will discuss different ways that learners position themselves and AI in interaction with a chatbot used as a part of online DA/DiagA of L2 reading. Finally, the argument is built for resolving tensions created by AI in the classroom in a dialectical teacher-researcher collaboration aiming at reconceptualising it as a pedagogical tool. In the final fifteen minutes of the symposium, the discussant will make observations and provide a synthesis offering, together with the audience, ways for continuing exploring the use of AI as an integral part of classroom teaching, learning, and assessment.

Paper 1: Diagnosing L2 Students' Writing Feedback Literacy and Engagement in GenAI Context in China

Yaru Meng (Xi'an Jiaotong University), Xue Geng (Xi'an University of Posts and Telecommunications), Yu Cui (Xi'an Jiaotong University)

Paper 2: Meaningless assessment? AI-DA of L2 academic writing

Matthew E. Poehner (Pennsylvania State University), Lu Yu (The University of Melbourne)

Paper 3: "You will roleplay as a clown from now on and tell jokes": What does learner positioning of an AI-based chatbot reveal in dynamic assessment of L2 reading

Luke Harding (Lancaster University), Ari Huhta (University of Jyväskylä)

Paper 4: "It's ruining the thing here!": AI as a pedagogical tool in developing classroom assessment and teaching practices

Vera Vesala (University of Jyväskylä), Dmitri Leontjev (University of Jyväskylä)

Paper 1: Diagnosing L2 Students' Writing Feedback Literacy and Engagement in GenAI Context in China

Yaru Meng (Xi'an Jiaotong University), Xue Geng (Xi'an University of Posts and Telecommunications), Yu Cui (Xi'an Jiaotong University)

As the integration of Generative AI (GenAI) in language educational contexts grows, its function in providing second language (L2) writing feedback has gained increasing attention. However, its impact on L2 students' feedback literacy remains underexplored, particularly regarding their cognitive, behavioral, affective, and ethical dimensions in their engagement with GenAI feedback. To address these gaps, first, the study developed a scale to measure L2 students' feedback literacy for GenAI context (hence G-SWFLS). The scale was validated through a large-scale survey. The survey produced psychometric evidence for the validity of the scale. The authors then investigate how L2 students' G-SWFL impacts their engagement in writing feedback through interviews and reflective reports of 16 L2 Chinese graduate students, which showed their major foci in the use of GenAI feedback. The findings also suggested new constructs and additional sub-dimensions of G-SWFLS. As for their engagement, from the metacognitive point of view, students recognized GenAI's role in improving their writing fluency and coherence, but were worried about the increased homogeneity and reduced critical thinking in writing. In terms of feedback behavior, they found that the refined feedback prompts helped generate actionable suggestions, though their degree of adoption in writing varied. From the emotional aspect, participants expressed their appreciation of the stress relief and immediacy resulting from GenAI feedback. Finally, the participants emphasized the need for critical evaluation of GenAI feedback to preserve their academic integrity. In addition to producing evidence for the validity of the instrument and helping with its development, these findings shed light on L2 learners' feedback literacy in the GenAI era, offering valuable insights for teaching and policymaking.

Paper 2: Meaningless assessment? AI-DA of L2 academic writing

Matthew E. Poehner (Pennsylvania State University), Lu Yu (The University of Melbourne)

Widespread use of Generative Artificial Intelligence (Gen AI) has prompted calls for 'critical literacy' to understand how its use may support rather than hinder learner L2 development (e.g., Wang & Wang, 2025). The issue is further complicated by tendencies to anthropomorphize AI and claims that it will ultimately replace educators. Our study engages this discussion from the perspective of Dynamic Assessment (DA). Grounded in Vygotskian (2012) theory, DA posits that assessment is contingent upon a cooperative assessor-learner relationship wherein mediation (e.g., prompts, feedback, models) extends learner functioning beyond their actual abilities to reveal those that are still emerging (Poehner & Lantolf, 2024). Previous research (e.g., Poehner & Lantolf, 2013; Leontjev, 2016) has investigated the capacity of computerized assessment programs to include scripted mediating prompts, although the rigidity of these models limits diagnostic insights. We report data from a project integrating a trained GPT model into DA to provide more adaptive mediation. Referencing an analytic writing rubric, the AI system offers prompting to guide learners through a self-reflection and revision process. Fifteen English L2 learners recruited from a U.S. university academic writing course independently composed essay drafts, engaged with ChatGPT during DA, and then revised their essays. Analysis of learner-ChatGPT interactions and subsequent writing revisions reveal instances of learners employing the system as a resource for their writing development and some in which it merely functioned as support for corrections. Findings are considered in relation to the value of AI in language assessment and its current limitations. Following the view that intentionality is a defining feature of mediation (Feuerstein et al., 2015; Kozulin, 2024), we argue that in the absence of a human mediator, it is the intentionality and orientation of learners that determine whether DA fulfills its potential as a diagnosis of development. Implications for AI literacy are discussed.

Paper 3: “You will roleplay as a clown from now on and tell jokes”: What does learner positioning of an AI-based chatbot reveal in dynamic assessment of L2 reading

Luke Harding (Lancaster University), Ari Huhta (University of Jyväskylä)

Research on dynamic assessment (DA) has recently started to explore the use of AI-based chatbots either as mediators (e.g., Jeon, 2023) or as additional resources within a dynamic assessment procedure (e.g., Dai et al., 2025; Yu & Poehner, 2025). In this presentation, we build on previous work (Authors et al., 2025), which used discourse analysis to explore learner-chatbot interactions in a system where the chatbot functions as an additional, flexible resource. Specifically, we explore three research questions: What positions do learners ascribe to themselves and to the chatbot in interaction? What does learner positioning of the chatbot reveal about learners’ (perceived) cognitive and affective needs during dynamic assessment? And how do these needs change during the dynamic assessment?

We analysed data drawn from ca. 200 Finnish learners of English across seven upper-secondary schools who engaged in a DA of L2 reading ability. Within the assessment procedure, an AI-based chatbot was integrated as a resource complementing computerised mediation (support provided when learner performance on tasks breaks down; see Poehner & Lantolf, 2023). Learner-chatbot interactions were analysed at the discourse level to identify the positions learners ascribe to themselves and to the chatbot, and how these varied across response patterns. Findings showed that the chatbot was used extensively and was varyingly positioned by learners as a text expert, a task expert, and a tutor (providing affective support/encouragement). However, learners also often “played” with the chatbot, particularly after items where no mediation was required, leading to complex negotiations about the chatbot's role and exploration of its limitations.

On a practical level, our initial findings suggest ways to further refine chatbot interactions and enhance the chatbot’s diagnostic potential. Theoretically, our findings illustrate the sophisticated ways in which learners engage with AI, suggesting dynamic assessments as rich contexts through which to explore facets of human-computer interaction and their impact on learning.

Paper 4: “It’s ruining the thing here!”: AI as a pedagogical tool in developing classroom assessment and teaching practices

Vera Vesala (University of Jyväskylä), Dmitri Leontjev (University of Jyväskylä)

The research of teachers’ conceptions about AI and its use in the classroom has shown that teachers are concerned about AI creating obstacles to learning, including learners cheating on tasks they set for them (Sütçü & Sütçü, 2023). The same is true for Finland, where discussions with regard to using AI have ranged from embracing it as a part of the teaching/learning process to largely forbidding its use in the classroom (FNAE, 2025; Sinkko-Westerholm, 2025). The aim of this presentation is to illustrate how the wide societal changes brought by the proliferation of AI can be translated to the local context of the Finnish upper-secondary school foreign language classroom. Our principled approach for that is Vygotskian praxis (Lantolf & Poehner, 2014), which positions the relationship between theory and practice, and teachers and researchers, as dialectical.

This presentation focuses on our collaboration with two English teachers informed by theoretical concepts coming from the dynamic assessment (DA) framework, a dialectical unity of teaching and assessment (Poehner, 2008). The data comes from transcripts of audio and video recorded meetings as well as lesson plans developed collaboratively with the teachers in which we built our joint understanding of the role of AI in classroom assessment, teaching, and learning. Leading the audience through these data, we illustrate how our praxis allowed for shifting teacher conceptions from AI as an efficiency tool to a tool for learning and showcase one lesson developed collaboratively with one of the teachers in which this conceptual understanding was put to test. The discussion will focus on the value of Vygotskian praxis as a principled approach in the development of teacher assessment literacy and classroom practices with the goal of transforming the role of AI from creating obstacles in the classroom practices to a tool that supports assessment and teaching.

Special sessions

Graduate Student Assembly (AGM)

Session facilitators: Che Han Chen, Karen Haeng-A Kim, and Shanshan He

Day: Thursday, June 4th (1:20 p.m. – 2:20 p.m.)

Location: Opera AB

The International Language Testing Association (ILTA) Graduate Student Assembly (GSA) will hold its first-ever in-person annual general meeting (AGM) at LTRC! Whether you are a current ILTA Graduate Student Assembly member or a (post)graduate student looking to learn more about what the ILTA GSA is, what it does, who's a part of it, and how to get involved, you're warmly invited to attend this interactive session!

During this 50-minute session, you will:

- meet members from the 2026 GSA Steering Committee;
- recognize how the GSA fits within ILTA;
- learn about the assembly's initiatives like the ILTA Bibliography, Webinar Committee, Peer Review Group, Language Tester Mentor Program, and Statistics Club;
- informally engage with ILTA Executive Board members; and
- socialize with (post)graduate students from all over the globe!

Using the Revised ILTA Code of Ethics and Guidelines for Practice in the Wild: An Overview and Strategies

Session facilitators: Margaret Malone and Dan Douglas

Day: Thursday, June 4th (1:20 p.m. – 2:20 p.m.)

Location: Symphonie 1

The language testing field has traditionally wrestled with the ethical issues inherent in assessing language learners and promoting the fair treatment of test takers—and all others in the assessment process—as humans with inherent dignity. Moreover, language test developers are an integral part of the process of supporting ethics in the field. With the ubiquitous nature of AI in daily life, it has become even more important to consider an ethics code, as well as guidelines for daily practice of language test development, administration and rating. This professional session will provide a short overview of the revisions to the ILTA COE and GFP. The session leaders will provide opportunities for participants to brainstorm ways to use these resources in test development, policy discussions and language testing courses as well as to discuss ways to maintain ethical processes in our work.

Equity, Diversity and Inclusion Issues in Language Assessment: Challenges, Practical Strategies and the Role of ILTA

Session facilitators: Monique Yoder (WIDA), Gladys Quevedo (Universidade de Brasília), and Richard Spiby (British Council). Contributor: Niles Zhao (National Institute of Education, Nanyang Technological University, Singapore)

Day: Thursday, June 4th (1:20 p.m. – 2:20 p.m.)

Location: Symphonie 2

Drawing on the recent work of the ILTA EDI Working Group, this session centres on a guided discussion of equity, diversity and inclusion (EDI) in language assessment, with attention to both field-wide challenges and organisational responsibilities. The panel addresses three key areas: (1) macro-level EDI challenges in the field, including bias in test design, accessibility barriers and socioeconomic impacts on outcomes; (2) practical strategies for creating inclusive language assessment frameworks that recognise linguistic diversity while maintaining fairness; and (3) ILTA's organisational efforts to advance EDI, including reflection on past and current initiatives and consideration of possible future directions. Participants will have opportunities to voice their experiences, raise concerns, and contribute suggestions for advancing EDI in research, practice and professionalisation.

CALA Annual General Meeting (AGM)



Session facilitators: Canadian Association for Language Assessment

Day: Thursday, June 4th (1:20 p.m. – 2:20 p.m.)

Location: Symphonie 3

Special Interest Groups (SIGs)

Language Assessment Literacy (LAL SIG)

Title: Looking back, moving ahead

Facilitators: Elsa Fernanda González and Isadora Teixeira Moraes

Day: Friday, June 5th (12:20 p.m. – 1:25 p.m.)

Location: Symphonie 1

During this meeting, we will briefly present some actions undertaken by the SIG since 2025, such as our webinars and reading groups. We will also engage with the audience with reflective questions to discuss possible suggestions for future actions to further develop the LAL field. Through this discussion, we hope to better cater to members' needs and strengthen their engagement with the ILTA and LAL communities. By the end of the session, we are hoping to find out more about our group members' interests in regards to LAL and whether we can improve our SIG actions in the remainder of 2026.

Language Assessment for Young Learners (YL Learners' SIG)

Title: Language assessment research in local to global contexts: A Young Learner's SIG Colloquium

Facilitators: Jeanne Beck and Hannah Kim

Day: Friday, June 5th (12:20 p.m. – 1:25 p.m.)

Location: Symphonie 2

This YL SIG colloquium brings together four studies examining language assessment for young learners (YLS) across varied contexts. Hyunah Kim and Eunice Eunhee Jang (2025) from the University of Toronto examine how cultural familiarity may (dis)advantage some students on a large-scale reading test in Canada, with a focus on length of residence. Veronika Timpe-Laughlin and colleagues (under review) analyze young learners' oral performances in the Netherlands, Finland, and Denmark to trace the development of speaking proficiency across levels measured by TOEFL Young Student assessments. Mark Chapman and colleagues (Chapman et al., 2025) from WIDA examine how task design and complexity influence fluency and other features of spoken language in U.S. multilingual school settings. Gary Ockey and Iowa State University colleagues (Ockey et al., 2026) review the globally used, digitally delivered Pearson Versant assessment through the lens of its YL test takers. The session concludes with a panel discussion on the common themes and issues.

Integrated Assessment (IA SIG)

Title: Integrated assessment: Ideas, questions, and on-going work

Facilitators: Ping-Lin Chuang, GoMee Park, Renka Ohta

Day: Friday, June 5th (12:20 p.m. – 1:25 p.m.)

Location: Symphonie 3

This informal session offers an open and welcoming space for conversation about integrated assessment. Rather than having formal presentations, the session will be organized around several groups, each dedicated to a specific topic such as task design and implementation, scoring and rating, or validation. Each group will be guided by a facilitator, and participants are encouraged to move between groups, engaging in dynamic discussions and interacting with others. Whether you are actively working on integrated assessment or simply curious about how it is defined and used across contexts, you are encouraged to join. Works in progress, tentative ideas, and practical challenges are all welcome. The goal is to create a relaxed environment where SIG members can exchange ideas, learn from one another, and reflect on current practices and future directions.

Automated Language Assessment (ALA SIG)

Title: A debate on the roles of humans in reading assessment development in the AI era

Facilitators: Jing Xu, Xiaoming Xi, and Hye-won Lee

Day: Friday, June 5th (12:20 p.m. – 1:25 p.m.)

Location: Symphonie 4

GenAI can generate reading passages and various types of test questions at different complexity levels. It has shown great promise in improving efficiency and reducing cost in reading assessment development. However, researchers have warned that AI-generated test materials may include social biases, poorly functioning questions, and non-factual information (Chuang & Xun, 2025; Xi, 2023). In practice, reading passages and questions generated by AI are usually reviewed by human experts for quality assurance. The ALASIG will organize a debate on the roles of humans in reading assessment development amid AI-driven changes in language assessment. Four panellists from academia and testing organisations will argue for and against the following statement: The use of generative AI in reading passage and test item generation eliminates the need for entry-level test developers, reserving human roles for only higher-level review, revision, and approval. The audience will cast their votes both before and after the debate.

Paper Sessions Summaries

Paper Session 1.1 (Day 1: 9:40 a.m. - 10:10 a.m.)

Assessing with AI: How Raters Negotiate LLM Suggestions in Writing Scoring

Authors: Hongyi Yang (University of Cambridge), Gabrielle Gaudeau (University of Cambridge), Andrew Caines (University of Cambridge), Andreas Säuberli (Ludwig-Maximilians-Universität München)

Day: Thursday, June 4 (9:40 a.m. - 10:10 a.m.)

Location: Opera AB

As large language models (LLMs) increasingly support human scoring in writing assessment, key questions emerge: How do raters engage with AI-generated input? Do they defer to these suggestions, critically evaluate them, or disregard them altogether? Which analytic traits are more susceptible to AI influence, and how might this reshape scoring cognition?

This study investigates how experienced raters respond to LLM-suggested analytic scores during essay evaluation. Using a mixed-methods design, we combine eye-tracking with interviews to explore rater cognition in real time. Thirty experienced raters participated in two 2h-long sessions, evaluating a total of 30 essays drawn from the ELLIPSE corpus. A split-screen interface displayed the analytic rubric (left) and the essay plus score entry (right). In the AI-assisted phase, rubric bands aligned with AI-suggested scores were visually highlighted.

This study contributes to ongoing conversations on scoring consistency and fairness in human–AI collaboration. It offers timely insights for designing rater training and developing responsible assessment policy as co-scoring becomes more prevalent.

Global Frameworks, Local Contexts: Evaluating the Cultural Appropriateness of English Language Tests for Vietnamese Candidates

Authors: Brigita Seguis (Cambridge University Press and Assessment), Quynh Nguyen (Vietnam National University), Nadir Zanini (Cambridge University Press and Assessment)

Day: Thursday, June 4 (9:40 a.m. - 10:10 a.m.)

Location: Symphonie 1

This paper examines the cultural appropriateness of the Cambridge English Qualifications (CEQs) within the Vietnamese educational context, addressing the broader challenge of reconciling global standardisation with local relevance in language assessment.

The research draws on the CEFR's concept of pluricultural competence (Council of Europe, 2020) and Messick's (1989) unified theory of construct validity. It employed a mixed-methods design, comprising findings from curriculum mapping, comparisons with major national English standardised tests, questionnaires for candidates ($n = 394$), as well as item-level Differential Item Functioning (DIF) and candidate performance analyses. Comparative statistics from other countries and candidate groups were also analysed to contextualise the results for Vietnam.

The curriculum mapping revealed strong alignment between the CEQ constructs, Vietnam's national English curriculum, and its high-stakes standardised exit and entry examinations. Most candidates reported familiarity with the test content and expressed strong intercultural curiosity; however, they were less confident in articulating aspects of Vietnamese culture in English. The DIF analysis indicated no systematic bias, while performance statistics showed that Vietnamese candidates performed comparably to the global cohort, with some nuanced differences across individual skills.

Overall, the findings demonstrate strong congruence between the CEQs and the requirements of the Vietnamese educational context. However, candidates' lack of confidence in articulating their cultural identity in English raises questions about how large-scale proficiency tests and local instruction can foster pluricultural competence, as envisioned in the CEFR. The paper concludes with practical recommendations, contributing to ongoing conversations about reconciling global standardisation with local authenticity in language testing.

The Impact of Rater Cognition on the Assessment of Interactive Listening in Vietnamese Higher Education Context

Author: Huong Tram Anh Nguyen (University of Queensland)

Day: Thursday, June 4 (9:40 a.m. - 10:10 a.m.)

Location: Symphonie 2

Interactive listening (IL), one of the five domains of Interactional competence (Galaczi & Taylor, 2018), refers to listeners' active engagement in co-constructing meaning and sustaining interactional flow. IL is realised through moment-by-moment displays of listeningship, topical contiguity, and collaborative completions (Lam, 2021), which are central to interactional achievement. In assessment contexts, however, conventional listening and speaking tests overlook IL, overemphasising production and monologic output, highlighting the need to foreground IL as a distinct construct. In Vietnam, interaction-based assessment is promoted under the National Foreign Language Project and CEFR-VN framework (MOET, 2014). Yet, descriptors for IC and IL remain broad and underspecified, while speaking tests often prioritise linguistic accuracy, creating a mismatch between policy intentions and assessment practices. Given this lack of operationalisation, IL assessment hinges on raters' cognition, and variability in these orientations may lead to differences in interpretation (Barkaoui, 2011), highlighting the need of examining rater cognition for validity and fairness. To address this gap, this study investigated how rater cognition influences IL assessment in English interactive speaking tasks in Vietnamese higher education. Four experienced raters assessed 74 performances across speaking tasks. Using a sequential exploratory mixed-methods design, verbal reports and semi-structured interviews were thematically analysed, complemented by MFRM of IL performances. Findings show that raters' construct understandings and scale interpretations shaped what they attended to, how they weighted features, and how severe they scored. The study contributes insights into how cognition shapes rating, informing IL descriptors, rubric development, and rater training for more valid, context-sensitive assessment in Vietnam.

Integrating Criticality in Language Assessment Education

Authors: Lia Plakans (The University of Iowa), Kwangmin Lee (Western Michigan University), Marina Durinova (University of Northern Iowa)

Day: Thursday, June 4 (9:40 a.m. - 10:10 a.m.)

Location: Symphonie 3

While global frameworks in Applied Linguistics and Language Assessment emphasize equity, inclusion, fairness, and justice, less attention has been paid to how these values are enacted locally—particularly in preparing language teachers, researchers, and test developers to understand and apply them in context. This presentation explores a curricular revision of a graduate-level language assessment course aimed at fostering critical engagement with these principles.

Designed by practitioner-researchers, the revised curriculum integrated foundational and contemporary scholarship in Critical Language Testing (Shohamy, 1998, 2001; Lynch, 2001), fairness and justice (McNamara & Ryan, 2011; Kunnan, 2004, 2014; Xi, 2010), and broader critical perspectives (Kubota & Miller, 2017; De Costa, 2018; Randall, 2021, 2023; Schissel & Khan, 2021). Students engaged with readings, reflective writing, discussions, and assignments that encouraged interrogation and application of key concepts.

The study asked: (1) What pedagogical approaches effectively integrate criticality into language assessment learning? (2) What challenges do students face in applying a critical lens? Data from a 16-week course with five graduate student participants were analyzed using thematic analysis, axial coding (Saldaña, 2013), and cross-case analysis (Denzin & Lincoln, 1994), with rigorous qualitative checks.

Findings show students' learning was shaped by prior experiences, conceptual development, and tensions between theory and practice. Pedagogical choices significantly influenced their ability to engage with global discourses through a localized lens. The study offers implications for language assessment literacy and professional development, advocating for socially responsive and transformative assessment education.

Reverse-Engineered Validity: Retrofitting a Legacy Multiple-Choice Test Through Assessment Engineering Principles

Authors: Petia Alexevia (Defense Language Institute English Language Center), Troy L. Cox (Brigham Young University), Matthew Wilcox (Brigham Young University)

Day: Thursday, June 4 (9:40 a.m. - 10:10 a.m.)

Location: Symphonie 4

This study describes how a long-standing multiple-choice English proficiency test was retrofitted using the principles of Assessment Engineering (Luecht, 2025) to align it with national proficiency frameworks. Originally created by non-specialists, the test consisted largely of grammar and vocabulary items and lacked an explicit construct definition. Yet it had become globally entrenched as an institutional standard, making full replacement impractical.

To improve responsiveness, stability, and cohesiveness, the development team reconstructed a construct map to clarify what the test was actually measuring and how that related to national descriptors. Using task-modeling procedures, the existing item bank was systematically analyzed for linguistic focus, cognitive complexity, and evidentiary contribution. Poorly written or redundant items were removed, and new reading and listening passages aligned with proficiency descriptors were added to better reflect the interpretive construct being measured.

The resulting re-engineered test preserved institutional stability while improving interpretive validity and policy alignment. This project demonstrates how legacy, non-specialist assessments can evolve through principled engineering—linking evidence, construct, and framework alignment—to meet both global operational needs and local policy expectations without disruptive redevelopment.

Paper Session 1.2 (Day 1: 10:15 a.m. - 10:45 a.m.)

A GenAI Writing Examiner for ESL writing: Scoring quality and revision engagement in test-preparation and classroom contexts

Authors: Daniel Yu-Sheng Chang (University of Bristol), Chun-Chia Chang (National Taiwan Normal University)

Day: Thursday, June 4 (Day 1: 10:15 a.m. - 10:45 a.m.)

Location: Opera AB

This mixed-methods study investigated the feasibility of Writing Examiner (WE), a self-prompted GenAI chatbot designed for writing practice in both test-preparation and classroom contexts. WE, built using ChatGPT-4 with ongoing integration of ChatGPT-5, was developed and piloted by the researchers, who provided free access to all participants. WE follows a structured four-step feedback process. Learners were encouraged to ask questions, request clarifications, and even challenge the given feedback during their revisions.

Thirty ESL learners completed eight writing tasks, working at their own pace. During the study, participants were preparing for standardised English tests and their English proficiency ranged from B2 to C1+. Data sources included original and revised essays, chatbot conversation logs, reflective journals, and analytic scores assigned by WE. Two expert raters evaluated the same essays to provide a benchmark, compared with ratings given by WE.

WE's ratings correlated moderately to strongly with expert ratings across the four subscales. No significant differences across task types were observed between WE and expert ratings. Severity levels were comparable. However, WE displayed an overly fitting (scoring pattern, suggesting a central tendency as it avoided extreme scores. Learners expressed overwhelmingly positive attitudes toward revising with WE, valuing its timely feedback and judgement-free environment and endorsing linguistic-focused comments. However, mixed perceptions emerged regarding a) the reliability of scores, b) the provision of global feedback, and c) whether AI-generated responses should be fully trusted. Engagement patterns varied, with some learners showing increased motivation while others displayed more superficial use.

Global Measures, Local Outcomes: ELP Test Predictive Validity in Graduate Admissions

Authors: Allison Michalowski¹, Ramsey Lee Cardwell¹, Jill Burststein¹, Steven Nydick¹, Anthony Verardi¹

Organization: 1: Duolingo

Day: Thursday, June 4 (Day 1: 10:15 a.m. - 10:45 a.m.)

Location: Symphonie 1

We report a multi-institution analysis of the comparative predictive validity at the graduate level of three English-language proficiency (ELP) admissions tests: Duolingo English Test (DET), IELTS Academic, and TOEFL iBT. Using admissions records for 1,841 international students who matriculated between 2022 and 2024 at four U.S. universities that accept all three tests, we examined group differences in mean GPA and score–GPA correlations. Analyses combined descriptive statistics, expectancy plots mapping score bands to GPA ranges, Pearson and Spearman correlations with Thorndike Case-2 adjustment for range restriction, outlier analysis, and psychometric meta-analysis of institution-level estimates.

Mean first-year GPAs did not differ by test within institutions, and expectancy plots showed monotonic increases in the likelihood of 4.0 GPAs with higher ELP bands across all tests. Pooled correlations were positive but modest (DET $r = .127$; IELTS $r = .095$; TOEFL $r = .233$) and generally increased after adjustments without changing substantive conclusions. For some institution–test combinations, removing a small number of outliers materially shifted point estimates. Across institutions, the three tests yielded effectively comparable admitted-student outcomes, and no single test systematically produced stronger score–GPA relationships. Theoretically, these results support a view that, at decision-relevant score bands, the three ELP tests afford a similar extrapolation warrant to early academic performance. Variability in correlations (negligible to modest) suggests contextual moderators—grading practices, disciplinary task demands, and instructional supports—and highlights that predictive-validity coefficients are sensitive to statistical artifacts. ELP test results should therefore be interpreted within the context of a holistic admissions process.

Integrated Speaking Tasks: Investigating Content Across CEFR Levels

Authors: Nazlinur Gokturk (Republic of Turkey Ministry of National Education), Sermin Vardal Ocakli (Ankara University), Gokce Fidan (Center for Measurement, Selection, and Placement)

Day: Thursday, June 4 (Day 1: 10:15 a.m. - 10:45 a.m.)

Location: Symphonie 2

Integrated speaking tasks can be useful for measuring L2 speaking ability authentically and comprehensively (e.g., Barkaoui et al., 2013); however, their complex nature poses challenges for evaluating the content-related aspects of task performance (e.g., Frost et al., 2021). While most studies focus on the tasks with written and audio input, the increasing use of video technology in educational contexts calls for an exploration of audio-visual input in integrated speaking tasks (Khabbazzbashi et al., 2023). Towards this goal, the current study investigates the operationalization of “content” in a prototype Watch and Speak Integrated Task developed for the Electronic Test of English (e-TEP), a large scale four skill-based English test designed for adult L2 learners in Türkiye. A sample of 60 task responses at different CEFR levels (A1, A2, B1, and B2) was randomly selected from the data collected and rated in pilot test administrations. Adopting a discourse-analytic approach (Brown et al., 2005), the transcripts of the task responses were first segmented into idea units (Inoue & Lam, 2021), which were then analyzed for their rhetorical functions. Frequency-based analysis was employed to describe patterns across CEFR levels, supplemented by illustrative excerpts from the task responses. Findings revealed that the prototype task elicited progressively richer content at higher proficiency levels, providing preliminary evidence for the appropriateness of the task for assessing L2 speaking ability. These findings contribute to a better understanding of the characteristics of performance on video-based integrated speaking tasks in terms of content, with implications for rating scale and task design.

Pre-service teachers' knowledge, practices and conceptions of writing assessment: A Comparative Study of Anglophone and Francophone Sectors in Quebec

Author: Amira Ben Hmida (University of Montréal)

Day: Thursday, June 4 (Day 1: 10:15 a.m. - 10:45 a.m.)

Location: Symphonie 3

Historically, Anglophone systems have been linked to holistic, learner-centred, and formative assessment approaches emphasizing communication, autonomy, and feedback, whereas Francophone systems have prioritized linguistic accuracy, structure, and adherence to normative standards. Québec offers a unique context where both traditions coexist within the same ministerial policy environment—including the Policy on the Evaluation of Learning (MEQ, 2003), the Progression of Learning, and the Reference Framework for Professional Competencies for Teachers (MEQ, 2020)—raising questions about whether these long-standing distinctions persist or have begun to converge in practice.

Drawing on the framework of Writing Assessment Literacy (WAL), this qualitative comparative study investigates the knowledge, conceptions, and practices of 18 pre-service teachers: nine in Teaching English as a Second Language (TESL) and nine in French as a Second Language (FLS). Semi-structured interviews examined three dimensions of WAL: understanding of assessment principles, conceptions of writing and its evaluation, and intended classroom practices related to feedback, rubrics, and professional judgment.

Preliminary findings indicate that, although the two groups conceptualize writing and assessment differently, their practices converge around similar focal points—particularly the assessment of linguistic features within students' written texts. These results illuminate the evolving assessment cultures of Québec's dual teacher-education systems and provide insight into the readiness and assessment literacy of future teachers, contributing to more coherent and equitable approaches to writing assessment across linguistic sectors.

The Role of Language Proficiency and Background Factors in Predicting International Students' Academic Success

Authors: Okim Kang (Northern Arizona University), Yuna Bae (Duolingo), Masha Kostromitina (Northern Arizona University), Harriet Dentaa (Northern Arizona University)

Day: Thursday, June 4 (Day 1: 10:15 a.m. - 10:45 a.m.)

Location: Symphonie 4

While proficiency tests are intended to assess students' ability to function successfully in academic contexts (Zareva, 2005; Weir, 2005), their predictive validity depends on how well test scores relate to real-world academic performance (Chapelle, 2021). Previous research has often directly correlated high-stakes test scores with academic outcomes but has overlooked the influence of institutional contexts and learner background factors. This study examined the extent to which background variables (proficiency scores, L1, major, gender, withdrawal, and high school GPA) contribute to international undergraduate and graduate students' cumulative GPA and first-year academic writing performance. It particularly focused on three high-stakes proficiency tests (DET, TOEFL, IELTS) widely accepted by U.S. institutions.

Institutional data from 29,168 international students (10,180 undergraduates and 18,988 graduates) at a public university in the southwestern United States were analyzed. Multiple regression results showed that all three proficiency tests significantly predicted cumulative GPA ($\beta = .10-.11$, $p < .001$), though high school GPA, withdrawal status, and major were stronger predictors. The models explained 4–12% of GPA variance. For first-year English course grades, DET ($\beta = .06$, $p = .0167$) and IELTS ($\beta = .10$, $p < .001$) were significant predictors, while background factors (high-school GPA, withdrawal, L1, major, and gender) were consistently significant ($p < .001$).

Overall, findings suggest that although English proficiency tests contribute to predicting academic success, background factors play a greater role. The three tests performed similarly within a local context, highlighting the need to consider contextual and individual factors in admissions and placement decisions.

Paper Session 2.1 (Day 1: 2:20 p.m. - 2:50 p.m.)

Investigating the implementation of automated writing scoring for adolescent multilingual learners in grades 6-12

Authors: Mark Derek Chapman¹, Katie Schultz¹, Ahyoung Alicia Kim¹, Kristin Erin Graw¹, Heather Elliot¹

Organization: 1: WIDA at UW-Madison

Day: Thursday, June 4 (Day 1: 2:20 p.m. - 2:50 p.m.)

Location: Opera AB

This multiphase study explores educator perceptions of automated writing scoring (AWS) for diverse groups of adolescent multilingual learners (MLLs) and reports on a trial of AWS for student responses in grades 6-12. In the first two phases, researchers collected qualitative data from educators about their perceptions of the use of AWS with MLLs via focus groups and a survey. Most educators believed that AWS would be able to reliably evaluate the writing performance of MLLs. Others raised concerns about AWS such as students' levels of digital literacy, learning disabilities, or physical limitations. In the third phase, WIDA and Pearson collaborated on a trial of AWS using Pearson's Intelligent Essay Assessor (IEA). The study utilized student responses to six distinct WIDA MODEL writing prompts across two grade bands (6-8 and 9-12), as well as the scores assigned to those responses by human raters. We applied data augmentation techniques to generate synthetic responses at underrepresented score points on the 0–18-point rubric used to score the WIDA MODEL Writing test. The performance of the scoring engine was evaluated using standard psychometric metrics, including agreement with human scores, reliability, and score point recall.

The presenters will share a synthesis of the findings from the three research phases, with a focus on the third phase. The findings provide evidence for the feasibility of implementing an AWS system for diverse groups of adolescent MLLs in a large-scale, standardized assessment context along with suggestions regarding how to assure educators of the meaningfulness of AWS for MLLs.

Global Frameworks, Local Classrooms: Aligning the AAPPL with the CEFR Scale

Authors: Jennifer Dobberfuhl Quinlan (Language Testing International), Young-A Son (Language Testing International)

Day: Thursday, June 4 (Day 1: 2:20 p.m. - 2:50 p.m.)

Location: Symphonie 1

The ACTFL Assessment of Performance toward Proficiency in Languages (AAPPL) is a large-scale test for learners in grades 3–12, aligned with the ACTFL Proficiency Guidelines (2024). As use of the AAPPL increases outside the United States, particularly in contexts where the Common European Framework of Reference (CEFR) is preferred, linking AAPPL scores to the CEFR has become essential for local educators to interpret scores and align curricula with global standards.

This presentation reports on early stages of aligning the English AAPPL with the CEFR, focusing on the familiarization and specification phases outlined in the CEFR Alignment Handbook. Prior research has largely centered on adult assessments and later alignment stages such as benchmarking and standard setting while few studies have examined earlier alignment stages or addressed young learner assessments.

A panel of experts completed asynchronous and synchronous familiarization activities to internalize CEFR scales and descriptors, followed by a specification phase analyzing how AAPPL content and score descriptors correspond to CEFR scales. Evidence sources included the ACTFL Proficiency Guidelines (2024), ACTFL Performance Descriptors (2012), AAPPL Score Descriptors (2020), and the Extended AAPPL Demo (n.d.).

The process yielded preliminary mappings between AAPPL test components and the CEFR scale, including Interpretive Listening, Interpretive Reading, Interpersonal Listening and Speaking, and Presentational Writing. Findings suggested CEFR levels most closely aligned with AAPPL scores. These results lay groundwork for subsequent benchmarking and validation, advancing efforts to connect U.S.-based proficiency frameworks with international standards and thereby support globally interpretable yet locally meaningful assessment outcomes for younger learners.

Self-reported strategic planning and avatarcharacter comprehension in a computerized-adaptive speaking assessment

Author: Sarah Goodwin (Duolingo)

Day: Thursday, June 4 (Day 1: 2:20 p.m. - 2:50 p.m.)

Location: Symphonie 2

Avatar-mediated speaking tasks represent a scalable solution for language assessment, offering greater efficiency and standardization than face-to-face interviews. However, their validity depends on understanding their impact on test takers (Saville & Khalifa, 2016). Test-taker perceptions are a key source of validity evidence, as learners at different proficiency levels bring varying cognitive resources to a task (Skehan, 1998), shaping their engagement and their view of the assessment's fairness, or face validity (Jin & Fan, 2023). Understanding these proficiency-based differences is crucial for informing equitable assessment design that is both psychometrically sound and perceived as meaningful by the test taker.

This study explored perceptions of an avatar-mediated speaking task. Test takers (N=5,066) took a practice proficiency test and were divided into beginner, intermediate, and advanced groups based on their results before completing the task and a post-test survey. While higher proficiency correlated with better comprehension of the avatar, ratings for how authentic and adaptive the conversation felt peaked with the intermediate group. These divergent perceptions provide critical evidence for the test's validity argument. For beginners, the high cognitive load of the task raises validity concerns. For advanced learners, a mismatch between their high-level abilities and the task's perceived simplicity has negative implications for face validity. These findings demonstrate that perception data is essential for validating the use of a standardized task across a wide proficiency spectrum and ensuring fair and meaningful score interpretations for all learners.

Exploring the link between linguistic features and psychometric difficulty in vocabulary assessment

Authors: Ping-Lin Chuang¹, Ben Naismith¹, Allison Michalowski¹, Masha Kostromitina¹, Geoff LaFlair¹

Organization: 1: Duolingo

Day: Thursday, June 4 (Day 1: 2:20 p.m. - 2:50 p.m.)

Location: Symphonie 3

The difficulty of a language test item can be conceptualized through the task's demands, the cognitive processes it elicits, and its linguistic characteristics. Linguistic difficulty, the extent to which linguistic forms affect comprehension, arises from multiple sources. In vocabulary assessment, it is influenced by word features relating to meaning, form, and use, alongside sentence-level factors including sentence length and syntactic complexity (e.g., Laufer, 1997; Toyama, 2021). Understanding how linguistic characteristics influence psychometric difficulty clarifies what aspects of language knowledge shape item performance.

This study analyzes the relationship between linguistic features and psychometric difficulty in a vocabulary-in-context task (13,212 items), where test takers complete a sentence containing a partial word. A set of linguistic features (initially, $n = 476$) was modeled to predict psychometric difficulty, employing regression as a proxy for expected linguistic difficulty. Features were refined using dimensionality reduction techniques, and items with large residuals between predicted and actual difficulty were flagged. Qualitative analysis of these mismatches identified aspects of linguistic difficulty not captured by the models, informing iterative model improvements. This process incorporated lexical features (e.g., frequency, density, predictability) and task factors like syllable-boundary truncation for the partial word.

Findings show that linguistic features effectively predicted item difficulty ($R^2 = .49-.74$), deepening understanding of the relationship between item difficulty and the language construct measured. The most predictive features included COCA-related frequency metrics, age of acquisition, and surprisal. These results inform assessment design, providing an evidence-based foundation for item evaluation, test development, and interpretation of vocabulary item difficulty.

Global AI-Scoring and Feedback, Local Fairness: TESOL Teachers' AI Literacy and Identity in Assessment Practices

Author: Johanathan Woodworth (MSVU)

Day: Thursday, June 4 (Day 1: 2:20 p.m. - 2:50 p.m.)

Location: Symphonie 4

AI-scoring and AI-mediated feedback are transforming language testing worldwide, yet their adoption unfolds locally as teachers adapt them to classroom assessment and feedback practices. This study investigates how K–12 TESOL teachers in Canada developed AI literacy and negotiated professional identity while participating in a graduate-level course on AI integration. Fifteen in-service teachers completed a 3-week intensive course. Data included pre- and post-surveys measuring nine AI literacy domains (technical knowledge, ethics, assessment applications) and three professional identity domains (role, relationships, classroom dynamics), as well as structured reflections at three time points. Analyses employed Wilcoxon signed-rank tests and reflexive thematic analysis. Findings showed significant gains across eight of nine AI literacy domains, with largest improvements in technical understanding, ethical reasoning, and social impact awareness. Teachers reported reduced anxiety about professional displacement and greater confidence in their essential role in assessment and feedback. Thematic analysis revealed five themes: progression to technical understanding, reconceptualization of AI-scoring as a supplement to teacher judgment, critical awareness of bias and equity concerns, context-specific integration planning, and attention to Universal Design for Learning principles.

Rethinking Item Calibration: Hybrid Approaches for Alternative Item Calibration

Authors: Ying Zheng (University of Southampton), David Booth (Pearson)

Day: Thursday, June 4 (Day 1: 2:20 p.m. - 2:50 p.m.)

Location: Ovation

Item calibration is central to ensuring fair and meaningful language assessment, but traditional pretesting within Item Response Theory (IRT) frameworks often clashes with local operational realities. Large-scale pretesting is expensive, time-consuming, and raises security concerns, especially in rapidly evolving assessment contexts. This study explores multiple alternative methods of item calibration that maintain psychometric quality while reducing the reliance on extensive pretesting.

Drawing on validity theory (Messick, 1989; Kane, 2013) and evidence-centred assessment design (Mislevy, 2018), a series of investigations were conducted using anchor items as benchmarks. A Standard Setting study (modified Angoff; 14 raters, 60 items) showed feasibility but was limited by rater variability. Two Comparative Judgement (CJ) studies demonstrated stronger potential: the first (14 raters, 60 items) produced difficulty estimates closely aligned with pretesting for multiple-choice and repeat-sentence tasks; the second (20 raters, 160 items) improved stability and comparability, showing scalability in real-world contexts. Text Complexity profiling was conducted to identify linguistic features that predicted reading item difficulty.

These complementary approaches offer credible alternatives to traditional calibration methods. CJ emerged as particularly effective, while text complexity added predictive power for reading tasks. The presentation concludes by proposing an Automated–Assisted Item Calibration Tool (A-AIT) integrating these methods, offering a pragmatic solution to the challenges of large-scale, secure, and sustainable item calibration.

Paper Session 2.2 (Day 1: 2:55 p.m. - 3:25 p.m.)

Rating Scale Adaptation for an AI-mediated L2 Speaking Test

Authors: Yasin Karatay¹, Jing Xu¹, Nicholas Glasson¹, Masaki Eguchi¹, Mao Saeki¹, Yoichi Matsuyama¹

Organization: 1: Cambridge University Press & Assessment

Day: Thursday, June 4 (Day 1: 2:55 p.m. - 3:25 p.m.)

Location: Opera AB

Recent advances in conversational (AI) have enabled the development of spoken dialogue systems (SDS) capable of simulating human examiners' verbal behaviors in direct speaking assessments (Author, 2025). These systems offer scalable, resource-efficient solutions for eliciting candidates' interactive behaviors that computer-delivered semi-direct speaking tests cannot capture (e.g., Ockey & Chukharev-Hudilainen, 2021). However, existing SDS studies (e.g., Gokturk & Chukharev, 2024; Timpe-Laughlin et al., 2025) have predominantly used rating scales developed for human-human interaction (HHI) to assess human-computer interaction (HCI). Given that contextual parameters influence test-taker and examiner behaviors (Inoue et al., 2024), the distinctive features of HCI, such as turn-taking dynamics, response timing, and constraints, necessitate modifications to existing scales.

This study develops and validates an HCI-specific rating scale for an SDS-mediated Oral Proficiency Interview (OPI) task using a mixed-method multiphase research design (Creswell & Plano Clark, 2012). It investigates: 1) issues arising from rating HCI using an HHI-based scale, 2) the effectiveness of a newly developed HCI rating scale, and 3) rater reliability and the optimal number of raters.

In Stage 1, two raters evaluated 30 non-native English speakers' SDS-mediated responses using the IELTS Speaking rating scale. Based on rater feedback and literature, "Fluency and Coherence" was replaced with "Interactiveness and Discourse Coherence," incorporating HCI-specific descriptors.

In Stage 2, four raters trained on the new scale rated 50 HCI responses. Reliability and validity were examined using Multi-faceted Rasch measurement and a fully crossed G-study and D-study. The study provides a construct-aligned framework for evaluating oral communication in HCI.

From Global Validity to Local Gatekeeping: Language Proficiency in South Korea's Marriage-Immigrant Naturalization Interviews

Author: Jiyeon Lee (University of Maryland-Baltimore County)

Day: Thursday, June 4 (Day 1: 2:55 p.m. - 3:25 p.m.)

Location: Symphonie 1

This study investigates language requirements in South Korea's naturalization exams for marriage immigrants, highlighting tensions between global validity theory and local gatekeeping. Since 2009, marriage immigrants, who are spouses of Korean citizens, have had to pass exams beyond the residence requirements. Using Kane's argument-based approach and Messick's unified validity framework, the study deductively codes naturalization amendments (2009–2024), National Assembly debates, media coverage, and interviews with applicants, NGO workers, and civil servants. Kane structures evaluation across inferences; Messick integrates evidential and consequential validity, including fairness. Emergent patterns are also noted. Preliminary findings show threats across the interpretive-use chain. Scoring lacks transparency and controls: the absence of public rubrics, limited rater training, and weak quality assurance allow for construct-irrelevant variance (memorized answers, demeanor). Generalization suffers from variable questions, interviewer styles, and site procedures; no public reliability studies exist. Extrapolation is weak: brief, scripted interviews offer little evidence of citizenship readiness (e.g., longitudinal integration). Decisions appear locally discretionary, with tenuous links from performance to outcomes. Interviewees report construct ambiguity and drift: applicants rehearse scripts instead of demonstrating civic reasoning, while civil servants describe inconsistent expectations. Triangulation shows policy rhetoric emphasizing "fit" and integration without technical rationales (e.g., standard setting). Messick-informed analysis finds insufficient documentation of content alignment, response processes, internal structure, external relations, and consequences. Recommendations: separate civics and language constructs, publish standardized materials and reliability evidence, and provide supports or bilingual options for civics components.

Eliciting Interactional Competence with Generative AI: A Mixed-Methods Study in a U.S. English Placement Test Context

Author: Inyoung Na (Iowa State University)

Day: Thursday, June 4 (Day 1: 2:55 p.m. - 3:25 p.m.)

Location: Symphonie 2

Assessing interactional competence (IC) in large-scale oral tests is challenging because paired and group tasks, while effective at eliciting IC features, introduce variability from interlocutor traits such as personality, proficiency, and speech variety. This variability makes it difficult to isolate individual IC ability. Attempts to reduce variability with spoken dialogue systems (SDS) have been critiqued for producing unnatural discourse. Recent advances in generative AI make it possible to design SDSs that interact more naturally and adapt dynamically to test takers. However, highly competent AI interlocutors may reduce the peer-to-peer dynamics that effectively elicit IC. If interactions are too smooth, key IC features such as comprehension checks, clarification requests, and repair sequences may not surface. To address this, a prototype large language model-driven SDS was developed to simulate paired discussions while creating deliberate communication challenges. The design strikes a balance between global fairness, achieved through standardized performance, and local authenticity by incorporating speech varieties and task conditions that reflect the communicative realities of international students. This study evaluates the system in the English Placement Test of a large Midwestern U.S. university. The data include forty international students completing paired discussion tasks, rated by four raters using 12 IC rating scales. Analyses include chi-square tests to examine feature ratability, many-facet Rasch measurement (MFRM) for scale functioning and rater consistency, qualitative discourse analysis to document IC feature presence, and thematic analysis of rater interviews. Preliminary findings suggest AI-mediated tasks enhance IC ratability and provide consistent opportunities for demonstration, with implications for balancing practicality, fairness, authenticity, and validity in high-stakes oral assessment.

Comparing signaling features in the stimuli of an academic listening test and authentic lectures

Author: Zhi Li (University of Saskatchewan)

Day: Thursday, June 4 (Day 1: 2:55 p.m. - 3:25 p.m.)

Location: Symphonie 3

Given its impact on test score interpretation, input authenticity in listening comprehension assessments has received considerable attention, especially from the perspective of corpus-based comparisons. This study investigated the use of signaling features, the linguistic devices for organizing structure and guiding the audience, in 88 mini-lectures from IELTS listening tests and 88 abridged authentic lectures from the British Academic English (BASE) corpus. Four signaling features, namely, questions, directives, connectives, and signaling nouns (also known as meta-linguistic nouns or relevance markers) were automatically extracted using a Python script and then manually checked for accuracy. Log likelihood ratio tests were applied to compare their frequencies across the corpora. The results indicate that while both questions and directives are visible in the IELTS mini-lectures, their usages are significantly more pronounced in the abridged lectures from the BASE corpus, highlighting a need for improvement in the interactional aspect of mini-lectures used in this academic listening test. On the other hand, the mini-lectures have significantly more instances of both connectives and signaling nouns than the abridged BASE lectures, suggesting clear signposting and higher information density in the mini-lectures. In addition, it is found that in the mini-lectures, the utterances that contain necessary information (NI) or content cues for answering questions share similar patterns, except for the use of signaling nouns. Qualitative analysis of the examples of signaling features provided additional information about the differences in engagement and signposting functions between the two corpora. Implications for developing listening materials will be discussed.

Rubrics for classroombased language assessment: Teachers' use of GenAI

Author: Meg Montee (Georgetown University)

Day: Thursday, June 4 (Day 1: 2:55 p.m. - 3:25 p.m.)

Location: Symphonie 4

Rubrics are particularly important in classroom-based language assessment because they can be used to consistently evaluate student work, communicate expectations to students, and provide feedback for learning (see Jonsson and Svingby, 2007; Crusan, 2015). Using data collected during a six-week online CPD program, this study addresses the following research questions:

- (1)How do instructors understand the purposes and uses of rubrics in their own contexts?
- (2)Through what processes do instructors develop LAL competencies in creating and using rubrics?
- (3)How do teachers use Generative AI to develop rubrics and how does the use of the tools affect their LAL development?

Participants include N = 20 instructors of languages other than English working in both K-12 and university US contexts. Data sources are a participant needs analysis, pre- and post-program participant interviews, course participation data, evaluation surveys, and rubrics created by instructors. Within the convergent parallel mixed methods design (Creswell & Plano Clark, 2011), the analysis synthesizes the multiple data sources to reveal developmental patterns of LAL for participants and the processes through which learning occurs.

Results show that there is often a threshold of expertise for utilizing Generative AI effectively and that at times the use of AI tools can obscure language instructors' learning processes and needs as they create products that may seem effective but do not reflect a deepening understanding of classroom-based assessment. The discussion focuses on strengthening partnerships between language testers and instructors through CPD as a site of mutual learning and collaboration.

Paper Session 3.1 (Day 1: 4:00 p.m. - 4:30 p.m.)

University staff's language assessment literacy development: cross-contextual comparisons

Authors: Angela Gayton¹, Daniel MK Lam¹, Beverly Baker², Coral Yiwei Qin², Emma Bruce³, Mojtaba Heydari²

Organizations: 1: University of Glasgow, 2: University of Ottawa, 3: British council

Day: Thursday, June 4 (Day 1: 4:00 p.m. - 4:30 p.m.)

Location: Opera AB

In this presentation, two groups of scholars, one Canada-based and one UK-based, compare our recent language assessment literacy (LAL) initiatives with post-secondary admissions decision-makers, reflecting on what we have learned from each other and from our participants.

In university admissions contexts, LAL incorporates knowledge, awareness and understanding of how standardised language test scores are used in admissions decision-making processes (Fulcher, 2012; Kremmel and Harding, 2020). Research to date indicates that relevant university staff are interested and motivated to develop their LAL specifically related to their work functions (Baker, 2016; Baker et al., 2024; O'Loughlin, 2013), around key issues such as which tests to accept, what minimum scores to set, and crucially the rationale behind such decisions.

In HE contexts currently facing myriad language assessment challenges, there is value for the language assessment community to take initiative on enhancing relevant stakeholders' LAL capacity and more robustly equipping them for such high-stakes processes and decisions. We report on separate initiatives in the two contexts, comparing our workshop objectives and target audiences; workshop designs; and respective discoveries on the basis of our workshop evaluations (both quantitative and qualitative).

We conclude by considering the foundation that this comparative analysis between the two teams' projects provides for future cross-cultural collaborative initiatives in the LAL field, theoretically in terms of its conceptualisation for these stakeholder groups, and also practically in terms of further stakeholder engagement. We explore opportunities for joint learning in LAL development, considering what principles are globally applicable, and which are locally specific.

Evaluating Consequences in Dynamic/Diagnostic Language Assessment: An Integrated Framework for Macro-Micro and Global-Local Impact

Author: Yong-Won Lee (Seoul National University)

Day: Thursday, June 4 (Day 1: 4:00 p.m. - 4:30 p.m.)

Location: Symphonie 1

Dynamic (DA) and diagnostic language assessment (DiagA) have recently drawn much attention from language testers. These two approaches pursue a common goal of facilitating further learning and development through assessment and assessment-based actions. Social mediation and intervention are often used as the major means of promoting further learning and growth in DA (Lantolf & Poehner, 2004) while diagnostic feedback and advice/guidance for remedial learning are provided to support subsequent learning activities in DiagA (Alderson, Brunfaut, & Harding, 2014; Kunnan & Jang 2011; Lee, 2015).

One important implication of such growth-oriented nature of these two approaches is that consequences should be treated as central pieces of assessment validity evidence. Thus, it seems critically important to have a comprehensive validity framework that can guide us through the process of collecting, evaluating, and synthesizing multiple lines of evidence coming from various stages/levels of assessment interpretation and use in DA and DiagA.

The main goals of this presentation are to: (a) review a brief history of validity theorizing and validity frameworks, particularly in connection with the evaluation of consequences; (b) present an integrated, argument-based validity framework containing two sub-frameworks that address quantitative and qualitative evidence; and (c) illustrate how it can be used to create, and implement, a validation plan for systematically collecting, evaluating, and synthesizing multiple streams of evidence, including individual-level, learning- and development-inducing effects, micro-level washback, macro-level impact, and all other consequences in the contexts of DA and DiagA. Some additional issues are discussed in relation to validating DA and DiagA.

Introducing Behavioral Differential Item Functioning in Language Testing

Authors: Sébastien Béland (Université de Montréal), Sandip Sinharay (Educational Testing Service), Yikai Lu (University of Minnesota)

Day: Thursday, June 4 (Day 1: 4:00 p.m. - 4:30 p.m.)

Location: Symphonie 2

We introduce a new concept, termed “behavioral DIF”, along with a hierarchical location–scale framework for investigating differential item functioning (DIF) in language testing. Instead of modeling item responses directly, we focus on the log-likelihood of each respondent’s full response pattern, derived from an item response theory (IRT) model. Both the conditional mean and variance of these log-likelihoods are modeled as functions of latent ability, group membership, and their interaction.

This approach allows us to examine two complementary dimensions of test fairness, which is essential in validity. First, differences in the conditional means (of log-likelihoods) reveal whether groups with the same ability level produce systematically more or less coherent response patterns, pointing to potential structural DIF. Second, differences in the conditional variance capture behavioral DIF, showing whether groups differ in the variability of their responses when ability is held constant.

By focusing on response behavior and strategy rather than accuracy alone, the method extends traditional DIF analyses. It highlights cases where linguistic or cultural factors influence how test-takers engage with tasks, which may not be visible through item-level accuracy models. The inclusion of random intercepts and group slopes per item further accounts for item-specific variability, ensuring that both global and local sources of misfit are identified. Additionally, our method integrates person-fit analysis (PFA) and DIF because several statistics used in PFA is based on log-likelihoods.

An application to TOEFL data illustrates how our approach can enrich validity arguments in language testing by uncovering nuanced group differences in response behavior.

Can Paired Tests Assess the Interactional Competence of Beginner Learners in EFL Classroom Settings?

Author: Kazuyo Kawamura (Kogakkan University)

Day: Thursday, June 4 (Day 1: 4:00 p.m. - 4:30 p.m.)

Location: Symphonie 3

Paired speaking tests, in which learners interact with one another, are an effective format for assessing interactional competence (IC). However, the focus of most studies has been limited to learners with relatively high English proficiency, typically those at the CEFR B1 level or above, such as university students studying in English-speaking countries or learners preparing for higher education in such contexts. To address this gap, this study investigated whether paired speaking tests could effectively assess the IC of Japanese junior high school students at the CEFR A2 level or below within EFL classrooms. Forty students participated in the paired speaking test, and their interactions were audio- and video-recorded for analysis. Macro-level analysis using the dyadic interaction model (Storch, 2020) revealed that 7 out of 20 dyads engaged in collaborative dialogue, characterized by high equality and mutuality. In contrast, 10 dyads exhibited parallel interaction patterns characterized by low mutuality, clearly distinguishing those that demonstrated IC from those that did not. These findings suggest that paired speaking tests can serve as a valid tool for assessing IC, even among learners with lower proficiency levels. Micro-level analysis showed that the frequent use of turn taking, responding, questioning, and breakdown repair correlated with higher IC levels, suggesting that these behaviors are useful indicators for assessing IC. Furthermore, the most common topic development pattern was a learner asking a question related to their partner's previous utterance, followed by a response with both a minimal answer and additional information.

Contextualizing Automated Reading Assessment with Dynamic Human Mediation

Authors: Melissa R Hunte (University of British Columbia), Eunice Eunhee Jang (University of Toronto), Angelie Ignacio (University of Toronto)

Day: Thursday, June 4 (Day 1: 4:00 p.m. - 4:30 p.m.)

Location: Symphonie 4

Automated assessments and digital learning platforms are increasingly utilized in K–12 education. However, concerns persist about whether AI adequately considers children's language and contextual realities, particularly multilingual learners (MLLs), students with learning differences (LDs), and minority groups. When educators tap into students' cultural and linguistic strengths, build trust, and scaffold responses to feedback, learners are more likely to revise, self-assess, and achieve mastery. Therefore, this Ontario-based study examined whether integrating dynamic assessment (DA), through human mediation, can complement automated reading comprehension assessment. Using a complementarity mixed-methods design, we asked: (1) What reading-skill profiles emerge from Cognitive Diagnostic Modelling (CDM) of automated scores? (2) How do students perceive and engage with human versus automated mediation? (3) How do profiles change after DA?

Students in Grades 3–8 (N = 445) completed five increasingly challenging digital reading passages. The assessment targeted three skills: explicit, inferential, and discourse-level comprehension, and the CDM confirmed eight latent classes. After each passage, the digital platform provided diagnostic feedback, and 32 students received additional human DA mediation. Tutors utilized platform diagnostics to support each learner's zone of proximal development and facilitate metacognitive reflections and self-regulated learning (SRL). Students responded with guided reflections while tutors observed and documented students' behaviours, affect, and strategies. These qualitative observations were analyzed alongside CDM profiles to track within-student changes.

Combining automated diagnostics with human mediation proved especially beneficial for MLLs and students with LDs. Effective mediation involved rapport-building and culturally responsive teaching practices. Furthermore, DA revealed qualitative changes in engagement, motivation, persistence, and SRL that the automated system or CDM alone could not detect.

Investigating the impact of calm music on L2 listening test anxiety through electrodermal activity (EDA) sensors and self-reports

Authors: Ga Eul You (University of Western Ontario), Ruslan Suvorov (University of Western Ontario)

Day: Thursday, June 4 (Day 1: 4:00 p.m. - 4:30 p.m.)

Location: Ovation

Listening tests are often among the most anxiety-inducing tasks for second language (L2) learners, with test anxiety shown to impair both performance and emotional engagement (Elkhafaifi, 2005; Zhang, 2019). While various classroom-based strategies have been explored (Horwitz et al., 1986; MacIntyre & Gardner, 1994), most studies rely on self-reported data that are vulnerable to bias and fail to capture real-time changes in anxiety. To address this gap, we investigated whether calm music—known to reduce anxiety in mental health contexts (Harney et al., 2023)—could alleviate L2 listening test anxiety and improve performance. Using a mixed-methods design grounded in theories of foreign language anxiety (Horwitz et al., 1986) and the arousal–mood–performance link (Thompson et al., 2001), we collected IELTS Listening scores, electrodermal activity (EDA) measures, and retrospective interview data from 38 L2 English speakers at a Canadian university. Participants completed the IELTS Listening test under two conditions: before and after a brief music intervention, while wearing EDA sensors. Findings revealed a complex relationship: Test scores declined following the intervention ($t(37) = 5.28, p < .001$), but physiological measures indicated reduced arousal ($t(37) = 3.99, p < .001$), and participants reported feeling less anxious. No significant correlations emerged between arousal, performance, and perceptions. These results underscore the importance of integrating physiological data with self-reports to better capture test anxiety. They also highlight the complex interplay between emotion and cognition in assessment, with implications for language testing research, classroom practice, and test design.

Paper Session 3.2 (Day 1: 4:35 p.m. - 5:05 p.m.)

Putting measurement to the test. Assessing the construct validity of implicit and explicit measurement instruments for vulnerable learners

Authors: Emma Maes (Ghent University), Bart Deygers (Ghent University), Marieke Vanbuel (KU Leuven)

Day: Thursday, June 4 (Day 1: 4:35 p.m. - 5:05 p.m.)

Location: Opera AB

Implicit and explicit measurement instruments are omnipresent in SLA research, where they operationalize language proficiency constructs. However, these instruments have been often validated with highly educated learners, raising questions about their ability to capture learning gains of those with interrupted schooling or limited literacy (LESLLA). Evidence suggests that LESLLA learners process and acquire language differently from their literate peers, highlighting the risk that low scores reflect biased measurement rather than limited learning. The current study investigates the construct validity of two implicit and explicit oral proficiency measures - an Elicited Imitation test (EI) and an Untimed Grammaticality Judgment test (GJT) - in a sample of LESLLA learners and highly educated learners ($n > 300$) enrolled in Dutch as a second language programs in Flanders. Many-facet Rasch analysis and regression analyses were applied to detect bias at item level for both measures and assess the suitability of these instruments for the target population. Results indicate that implicit tasks - such as the Elicited Imitation test - better capture progress among LESLLA learners, suggesting that widely used explicit measures may misrepresent learning gains. The findings emphasize the need for context-sensitive measurement tools by revealing how widely used language measures may misrepresent learning gains in vulnerable learner populations, with important implications for policy, test design, and equitable practices.

An ethnographic approach to domain analysis in LSP assessment design: Exploring local and global dimensions of English in a Turkish Tech Start-up

Author: Jennifer Kay Morris (Lancaster University)

Day: Thursday, June 4 (Day 1: 4:35 p.m. - 5:05 p.m.)

Location: Symphonie 1

This study explores the value of a context-rich, ethnographic approach to domain analysis in Language for Specific Purposes (LSP) assessment, using a case study of “TechX,” a tech start-up in Türkiye. While typical domain analyses and needs analyses often rely on surveys, interviews with domain participants on the periphery, this research demonstrates how immersive ethnographic methods with domain inhabitants can capture the complexity of workplace language use in tech-driven, globalized environments. Anchored in Mislevy et al.’s (2017) Evidence-Centered Design (ECD) framework and Dursun’s (2023) multi-dimensional DA framework, the study introduces an “embedded-actualised” technique (Adams, 2024), grounded in standpoint theory, to examine both online and physical workplace contexts.

The methodological approach focuses on Documents (texts and artifacts) and Contexts (in-situ language use and linguistic landscapes), addressing underexplored dimensions of DA. Data collection followed three stages: (1) embedding: remote interviews, observations, and daily surveys; (2) actualizing: on-site observations and informal interviews; and (3) domain compositing: synthesizing findings into a holistic workplace profile.

Findings reveal role-specific English (L2) proficiency requirements among Turkish (L1) professionals, with advanced skills needed for global engagement. However, a disconnect exists between stated language priorities and hiring practices that favor technical skills. Additionally, the increasing use of generative AI tools demands communicative flexibility and new literacies beyond traditional language competencies. By centering lived experiences and contextual realities, this study advocates for ethnographically grounded LSP assessment design that links global frameworks with local practices. It calls for more ethically responsive, situated approaches to capturing language use in evolving workplace domains.

Exploring the Impact of Learning Journals as a Formative Assessment Tool on Chinese English Major Students’ Self-Regulated Listening Strategy Use

Authors: Qiaozhen Yan (Chongqing University), Shiyao He (Chongqing University)

Day: Thursday, June 4 (Day 1: 4:35 p.m. - 5:05 p.m.)

Location: Symphonie 2

This study addresses how learning journals, as a formative assessment tool, can foster self-regulated learning (SRL) in L2 listening. Although formative assessment has been widely recognized for its role in enhancing L2 students’ SRL, existing research has predominantly focused on methods like self-and peer assessment, leaving the potential of learning journals as a formative assessment tool under-explored. Moreover, there remains a notable lack of research on formative assessment implementation in L2 listening contexts.

To address these gaps, this study investigated the impact of learning journals as a formative assessment tool on Chinese English majors’ self-regulated listening strategy use and the factors influencing the effectiveness of journal implementation. Guided by Zimmerman’s (2002) SRL model, a qualitative study was conducted in a first-year English major students’ listening course at a Chinese university. Over one semester, students kept weekly learning journals to reflect on their experiences with BBC news dictation tasks. Data were collected from two rounds of semi-structured interviews with 13 students selected through purposive sampling strategy and from their journal entries, which were analyzed using a thematic analysis approach.

Findings revealed that learning journals had positively influenced students’ SRL capabilities across the three phases of Zimmerman’s (2002) model: forethought, performance, and self-reflection. However, the effectiveness of learning journals was moderated by four categories of factors related to the student, the task, the teacher and the institution. This study contributes to our understanding of the facilitative role of learning journals as a formative assessment tool in supporting self-regulated listening strategy use.

AI Feedback for Speaking Assessment

Authors: Sherry Hsin-Ying Li¹, Jessica Row-Whei Wu¹, Rachel Yi-Fen Wu¹, Berlin Chen², Fu-An Chao²

Organizations: 1: Language Training and Testing Center, 2: National Taiwan Normal University

Day: Thursday, June 4 (Day 1: 4:35 p.m. - 5:05 p.m.)

Location: Symphonie 3

This presentation reports on the development and validation of an AI-driven scoring and feedback system for speaking assessment, designed to enhance both large-scale English proficiency testing and pedagogical feedback in Taiwan. Unlike most learner corpus and automated scoring research, which focus on writing (e.g., Attali & Burstein, 2006), this project draws on spoken test performance data from 1,200 GEPT intermediate-level (CEFR B1) test-takers. Using open-ended Picture/Chart Description tasks, the system captures extended discourse that integrates visual and textual input, offering richer evidence of communicative ability than constrained oral tasks.

The model integrates diverse inputs—Chinese prompts, visual stimuli, exemplar responses, learner audio, and human scores—within a multimodal large language model (LLM) framework. It generates both holistic speaking scores and diagnostic feedback on semantic relevance, language use, and delivery, moving beyond traditional automated scoring toward a feedback-rich and pedagogically responsive process.

Challenges included ASR imperfect transcription, imbalanced score distributions, and limited analytic scoring data (Lo et al., 2024). A further challenge lay in interdisciplinary collaboration, requiring sustained dialogue between assessment specialists and AI engineers to balance construct validity with technical feasibility.

By shifting the focus from scoring accuracy to feedback quality, this project illustrates how locally grounded AI systems, built on authentic GEPT speaking data, can foster learning and teaching.

Harnessing Generative AI to Develop L2 Reading Comprehension Tests: A Human–AI Comparison

Authors: Ray J. T. Liao (National Taiwan Ocean University), GoMee Park (University of Texas Rio Grande Valley), I-Chun Hsiao (University of Iowa)

Day: Thursday, June 4 (Day 1: 4:35 p.m. - 5:05 p.m.)

Location: Symphonie 4

With the emergence of Generative Artificial Intelligence (GenAI), researchers have increasingly explored its applications in second language (L2) assessment. While prior studies focused mainly on question generation, this study examined GenAI's capacity to develop both reading texts and comprehension questions, comparing its output to human expert–developed tests.

Three English reading comprehension tests were created: two by ChatGPT-4.0—one fine-tuned using high-quality English tests and one generated without fine-tuning—and one by human experts. Each test included science and humanities passages aligned to CEFR B2 difficulty, with multiple-choice comprehension questions.

Thirty-two experienced English teachers and test developers completed Likert-scale questionnaires evaluating text and question quality, analyzed via repeated measures MANOVA. Follow-up interviews with 12 participants provided qualitative insights, analyzed inductively.

Quantitative results showed that AI-generated materials—particularly untrained AI—were generally rated higher than human-developed tests, especially for science passages and humanities comprehension questions. Qualitative findings revealed that raters valued coherence and structure but noted formulaic tendencies and a lack of “human voice” in AI texts, contrasted with stylistic richness yet weaker coherence in human-written passages. Science texts—regardless of source—were generally preferred over humanities texts, particularly AI-generated ones. Question-level insights indicated that both AI- and human-developed items exhibited misaligned skill labels, incoherent sequencing, unclear stems, and weak distractors, affecting validity.

Collectively, findings suggest that GenAI can produce L2 reading assessments comparable to human experts, with implications for enhancing classroom assessment practices and guiding broader language testing initiatives, though careful item design and nuanced text development remain essential.

AI or No AI: Contextualizing Listening Assessment Across Languages and Proficiency Levels

Authors: Ruth Maria Baptista (Brigham Young University), Jennifer Hansen (Brigham Young University), Matthew Wilcox (Brigham Young University)

Day: Thursday, June 4 (Day 1: 4:35 p.m. - 5:05 p.m.)

Location: Ovation

The rapid spread of artificial intelligence has created new possibilities for language testing, including the use of AI-generated voices to deliver listening passages. For test developers, these voices reduce costs and allow for large-scale scalability. However, for learners, questions remain about authenticity, fairness, and potential effects on item difficulty. This study examines the authenticity and perception of AI-generated versus human-generated listening items in language proficiency assessments at intermediate and advanced levels in Portuguese and English. In alignment with the ACTFL proficiency scale, it develops listening items that reflect real-world communicative contexts. Two learner populations—approximately 100 English L2 learners in a university-affiliated intensive program and 100 university students learning Portuguese (L2)—will complete parallel forms with identical passages and items; only the audio source varies (human narrators versus high-quality AI-generated voices) under standardized conditions. Data will be analyzed to determine whether item difficulty differs between conditions, and interviews and post-test surveys will gather perceptions of authenticity, fairness, and preference. This combination of performance and perception data will determine whether synthetic voices affect test outcomes and how they shape learners' experiences and evaluations of the test across languages and settings. The study provides timely evidence on AI's role in language testing and directly addresses whether AI-generated voices are acceptable substitutes for human narration in listening proficiency assessment, both in terms of psychometric performance and test-taker perceptions, highlighting how AI audio can be used without sacrificing authenticity, validity, or fairness. It connects to the LTRC 2026 theme of balancing global trends with local contexts.

Paper Session 3.3 (Day 1: 5:10 p.m. - 5:40 p.m.)

How Raters with Different Backgrounds Judge West-African Test Takers' Video Interviews

Authors: Kadidja Kone (Ecole Normale d'Enseignement Technique et Professionnel), Paula Winke (Michigan State University)

Day: Thursday, June 4 (Day 1: 5:10 p.m. - 5:40 p.m.)

Location: Opera AB

Over 2,000 colleges and universities world-wide use Slate® as their online submission software. The software allows institutions and/or admissions-committee members to see an international applicant's Duolingo English Test (DET) "Video Interview" within their test-score-report portals if they took the DET. Thus, the video evaluators can use them for admissions-decision-making purposes alongside the students' DET scores. However, the video evaluators may not have training in fair and equitable practices in rating, which allow for judgement biases. To test this hypothesis, we had seven novice raters evaluate the Video Interviews of 60 Francophone West-African test takers for comprehensibility, fluency, lexicogrammar, and response quality. The raters, who came from three different backgrounds (i.e., Africa, Asia, Europe, and USA) additionally provided written comments while rating. Using Rasch analysis and Differential Item Functioning (DIF) analyses of the test scores, and a grounded-theory, thematic analysis of the written comments, we elucidate potential biases and suggest a draft of a global tip sheet with "dos and don'ts" in evaluating Video Interviews to promote equity and fairness for stakeholders.

Bridging Global Assessments and Local Contexts: How Learner Background Shapes Young Learners' English Proficiency

Authors: Mohammadreza Dalman (Winona State University), Okim Kang (Northern Arizona University), Maria Nelly Gutierrez Arvizu (Universidad de Sonora Mexico), Francesca Grixoni (Northern Arizona University)

Day: Thursday, June 4 (Day 1: 5:10 p.m. - 5:40 p.m.)

Location: Symphonie 1

The introduction of English in primary schools has increased the need for age-appropriate proficiency assessments for young learners (YLS). While family socioeconomic status (SES), learner background, and educational context are known to influence language development, little research has examined these factors in formal assessment settings. This study explored (1) how learner background variables—language contact, technology access and use, and assessment practices—and family SES indicators—parents' education, English proficiency, income, occupation, and residential location—predict YLS' proficiency scores, and (2) how performance varies across instructional contexts. Participants were 145 fifth- and sixth-grade students from three school types in Mexico: Bilingual schools (English as medium of instruction), Private schools (English as a foreign language), and Public schools (limited English exposure). Learners completed online questionnaires and took the British Council Primary English Test (PET). Multiple regression analyses revealed that language contact outside school ($R^2 = .18$), frequency of digital technology use ($R^2 = .27$), and number of quizzes completed ($R^2 = .26$) significantly predicted proficiency scores. Learners with only smartphone access scored lower than those with tablets and computers ($p < .001$). Among SES variables, parents' English proficiency, income, and residential location were significant predictors ($R^2 = .33$). ANOVA results showed that learners in bilingual schools outperformed peers in other contexts ($F(2,142) = 214.70$, $p < .001$). Findings highlight the combined influence of family and learner-related factors on YLS' English proficiency and underscore the need for equitable access to resources and fair assessment design.

Supplementing calibration with linguistic and cognitive item features: lessons from large-scale English language tests

Authors: Nadir Zanini¹, Mark Elliott¹, Yan Huang¹, Lina Loxley¹, Siyuan Tang¹

Organization: 1: Cambridge University Press & Assessment

Day: Thursday, June 4 (Day 1: 5:10 p.m. - 5:40 p.m.)

Location: Symphonie 2

Calibration plays a crucial role in high-stakes testing and test validation, traditionally relying on Item Response Theory (IRT) models applied to representative samples of test-takers. This can be a lengthy and expensive process that does not always yield the information needed. To improve the efficiency of calibration, recent research has explored the use of Natural Language Processing (NLP) in predicting item parameters. These studies, however, often pertain to the calibration of a limited number of items, specific item types, or focuses exclusively on difficulty.

This paper investigates whether combining linguistic and cognitive features can enhance calibration efficiency in large-scale Reading assessments. It expands the scope beyond difficulty to include discrimination and model fit statistics, which are vital for test construction. Additionally, it examines the feasibility of reducing calibration sample sizes and excluding certain language groups.

Using over 1,200 items from Cambridge English's B1 Preliminary and B2 First exams, the study extracts linguistic features via NLP, alongside cognitive features derived from expert analysis based on a hierarchical processing model. Regression modelling is employed to evaluate feature predictive power and reduce dimensionality. Alternative calibration approaches are trialled across varying sample sizes and representativeness. Their performance is compared against live-calibrated item parameters and fit statistics in both pre-testing and live settings.

Findings will inform a discussion on how technology can support more efficient, inclusive, and valid calibration practices, highlighting the potential of integrating machine learning with psychometrics to improve language testing and provide a suitable even with less globally representative samples.

Tailoring the Learning Journey: Unveiling Personalized Learning Paths in L2 Speaking through Cognitive Diagnosis

Author: Shuting Zhang (Zhejiang University)

Day: Thursday, June 4 (Day 1: 5:10 p.m. - 5:40 p.m.)

Location: Symphonie 3

The paradigm shift from “assessment of learning” to “assessment for learning” has led language assessment to prioritize providing diagnostic information to facilitate subsequent learning. Cognitive Diagnostic Assessment (CDA) offers fine-grained insights into learners’ cognitive structures, identifying strengths and weaknesses to inform personalized learning. However, CDA remains underexplored in speaking assessment, particularly in the Chinese EFL context, where the need for oral proficiency diagnosis is substantial but unmet.

This study develops a CDA framework to assess Chinese EFL learners’ speaking proficiency and construct personalized learning progressions. Spoken responses from 508 examinees in an in-house English Proficiency Test were analyzed using a 16-item diagnostic rating scale aligned with the Chinese Standards of English Language Ability (CSE). Four attributes—pronunciation, fluency, lexico-grammatical ability, and content development—were specified in a Q-matrix. After comparing multiple CDMs, the Mixed-CDM was selected as the best-fitting model. Ability estimates (θ) were then derived from a 3PL IRT model to establish hierarchical patterns and infer learning paths.

The results show sixteen distinct mastery profiles across the four attributes. Learners with identical total scores displayed different underlying knowledge states. The most prevalent trajectory—0000 → 1000 → 1010 → 1011 → 1111—reveals a progression from pronunciation and lexico-grammatical ability to content development and finally fluency. By integrating CDA and IRT, this study provides an effective data-driven approach to modeling hierarchical relationships among speaking subskills and mapping personalized learning paths, thereby highlighting CDA’s potential as a powerful tool for advancing learning-oriented assessment in L2 speaking.

Design and Validation of a Language Placement Test: Insights from Class Performances, CEFR-Based Self-Assessment, and Test Perceptions

Author: Punchalee Wasanasomsithi (Chulalongkorn University)

Day: Thursday, June 4 (Day 1: 5:10 p.m. - 5:40 p.m.)

Location: Symphonie 4

Every year, Chulalongkorn University Language Institute offers English language instruction to approximately 6,000 first-year students with different proficiency levels. A placement test is therefore essential to create homogeneous groups of learners (Purpura, 2004). This presentation focuses on the design and validation of the Chulalongkorn University English Language Placement Test with the aim of enhancing the effectiveness of instruction and teacher support. The CU-ELPT comprises 80 multiple-choice items assessing listening, academic vocabulary, language use, and reading, along with a 250-word essay, which, collectively, assess the skills required for undergraduate studies. The newly developed test was piloted with 303 first-year students at the Faculty of Medicine. Placement test scores were compared with the scores from the English for Medical Profession course, a required course for first-year medical students. Additionally, students completed a CEFR-based self-assessment, and in-depth interviews were conducted with 12 students, evenly distributed among high, middle, and low proficiency, as well as nine instructors, to elicit insights into students’ perceptions of the test and their own performance and instructors’ perceived appropriateness of student assignment. This presentation highlights how the comprehensive analysis stemming from this initial trial has paved the way for necessary revisions of the placement test to be subsequently administered on a larger scale and led to the development and validation of a more effective placement system. The presentation ends with a discussion of future directions for the Language Institute regarding the use of the placement test and illustrates how the design and validation process of the test can be adopted by interested parties.

Investigating the cognitive validity of an integrated listen-to-summarize test: A mixed-methods analysis of test-taker notes

Authors: Rebecca Yeager (University of Illinois Urbana-Champaign), GoMee Park (University of Texas at Rio-Grande Valley), Junhee Park (University of Iowa)

Day: Thursday, June 4 (Day 1: 5:10 p.m. - 5:40 p.m.)

Location: Ovation

This study examines the cognitive validity of a listen-to-summarize task on a university placement test. Cognitive validity is underinvestigated in listening assessment, especially for higher-order listening processes (Holzknecht et al., 2017). However, student notes from listening tests can record evidence of these processes (Cushing, 1991; Rukthong, 2021). This study employed an explanatory sequential mixed-methods design to investigate the higher-order listening processes elicited by an integrated listen-to-summarize task. First, notes from two administrations of the same test form – one with brief scaffolding instructions on notetaking strategies ($n = 25$), and one without ($n = 25$) – were analyzed qualitatively, drawing on Field's (2013) model of higher-order listening processes: Selecting, Integrating, Monitoring, and Structure-Building. Second, qualitative findings informed the development of a rubric to quantify higher-order listening processes across administrations and in relation to summary and multiple-choice task scores on a third test administration ($n = 118$). Qualitative codes revealed substantial evidence of higher-order processes in student notes across test administrations. Linear regression models comparing notes from the first two test forms confirmed that Selecting, Integrating, and Structure-Building strategies were significantly more frequent when notetaking scaffolding was provided. Finally, linear regression modeling scores on the third test form indicated that Selecting processes significantly predicted test scores for both the integrated task and the multiple-choice task, while Structure-Building processes were significant only for the integrated task. Results support the cognitive validity of the listen-to-summarize task and the potential for mixed-methods notetaking analysis to investigate the cognitive validity of listening assessment in diverse contexts.

Paper Session 4.1 (Day 2: 9:40 a.m. - 10:10 a.m.)

LLM surprisal as a meaningful indicator of writing proficiency in L2 French

Authors: Guillaume Loignon (Université du Québec à Montréal), Angel Arias (Carleton University), Émilie Courteau (Dalhousie University)

Day: Friday, June 5 (9:40 a.m. - 10:10 a.m.)

Location: Opera AB

This study investigates linguistic surprisal as an interpretable and theoretically grounded indicator of L2 writing proficiency. Surprisal quantifies how unexpected a word is given its context and can now be efficiently computed using large language models (LLMs). Psycholinguistic research has shown that surprisal reflects cognitive effort during comprehension, yet few studies have applied it to written production. We analyzed 234 French L2 essays produced under examination conditions and rated for proficiency by human evaluators. Surprisal values were computed using a French unilingual LLM ModernCamemBERT. Spelling, grammar, and syntax errors were identified using Antidote 12. Results show that low-level writers had significantly higher surprisal than high-level writers, and that error rates increased with surprisal. These results suggest that surprisal correlates with writing proficiency but also with classroom-relevant features. Overall, surprisal shows promise as an interpretable and psycholinguistically grounded feature for automated essay scoring and second-language writing research.

Predictive roles of nonverbal behavioral patterns for analytic scores in L2 speaking assessment

Authors: Jiayi Wu (Zhejiang University), Shangchao Min (Zhejiang University), Zhuohan Hou (Zhejiang University)

Day: Friday, June 5 (9:40 a.m. - 10:10 a.m.)

Location: Symphonie 1

This study examines whether nonverbal behavior can be grouped into distinct behavioral patterns and how these patterns explain L2 speaking scores. Thirty Chinese university students with varying English proficiency completed two 8-minute paired conversation tasks with the same trained interlocutor, drawn from five campus-related topics to ensure comparability. Participants' nonverbal behavior was annotated using ELAN with high interrater reliability ($ICC > .80$) and converted into normalized frequency measures. Audio-based speaking scores in interactional competence, language use, fluency, and pronunciation on a 0-5 continuous scale were obtained from two trained raters. The resulting measures and test scores were analyzed using linear mixed-effects models and k-means clustering. Significant effects in the mixed-effects models were further explored with post hoc pairwise comparisons using Tukey adjustment. K-means analysis identified three distinct behavioral modes: an information-enhancement mode (rich head nods, eyebrow movements, beat and non-beat gestures), an interaction-management mode (sustained gaze and smiles), and a low-nonverbal mode (uniformly low frequency). Linear mixed-effects models revealed that both interaction-management and information-enhancement behavioral patterns corresponded to significantly higher scores in IC, language use, pronunciation, and fluency, compared with the low-nonverbal pattern. Post hoc pairwise comparisons with Tukey adjustment further indicated that the information-enhancement pattern was associated with significantly greater scores than the interaction-management one. These findings advance our understanding of interactional competence by uncovering distinct nonverbal behavioral patterns and their role in shaping L2 interaction, and emphasize the need to incorporate multimodal evidence into rater training and performance descriptors for more valid and nuanced interpretations of communicative ability.

Investigating the Role of Rater Profiles in Applying Rating Rubrics for Second Language Comprehensibility

Authors: Aki Tsunemoto (Tohoku University), Rie Koizumi (University of Tsukuba), Ryo Maie (Tohoku University), Yo In'nami (Chuo University), Makoto Fukazawa (University of the Ryukyus), Mariko Abe (Okayama University)

Day: Friday, June 5 (9:40 a.m. - 10:10 a.m.)

Location: Symphonie 2

Developing oral proficiency is a major goal for many second language (L2) speakers, yet learners in English as a Foreign Language (EFL) contexts often face challenges due to limited opportunities for L2 exposure and experience. Comprehensibility, defined as the ease or difficulty with which speech can be understood, is a key dimension of L2 oral proficiency that seems amenable to L2 instruction (Saito & Plonsky, 2019). Previously, Isaacs et al. (2018) developed comprehensibility rating rubrics for formative assessment, with the aim of helping teachers and learners identify the linguistic factors affecting comprehensibility. Our precursor study adapted these rubrics to an EFL context and found that language teachers in Japan could reliably evaluate learners' comprehensibility. However, because our raters and speakers shared the same first language (L1), the rubrics' generalizability to more diverse rater groups remained yet to be tested.

Using the holistic (comprehensibility) and analytic (pronunciation, fluency, vocabulary, grammar) rubrics, 55 L1 Japanese teachers and 18 L1 non-Japanese teachers evaluated 50 speech samples (45 seconds each) from a corpus of extemporaneous speaking tasks performed by Japanese secondary school students (e.g., describing a favorite meal). Using an online rating interface, raters completed brief training before independent rating. A many-facet Rasch analysis revealed considerable variation in rater severity: L1 Japanese teachers were generally stricter than non-Japanese teachers who showed greater within-group variation. A generalizability study indicated that two raters and one criterion were sufficient to maintain adequate reliability. Implications will be discussed for classroom assessments involving teachers from different linguistic backgrounds.

Does word class matter? Investigating the relationship between word frequency and item difficulty across parts of speech in L2 French vocabulary testing

Authors: Ji-young Shin¹, Jeffrey Steele¹, Magda Tigchelaar¹

Organization: 1: University of Toronto Mississauga

Day: Friday, June 5 (9:40 a.m. - 10:10 a.m.)

Location: Symphonie 3

Second language (L2) vocabulary tests are often structured by corpus-based lexical frequency, reflecting the usage-based assumption that more frequent words are easier to learn and thus yield easier items (Milton, 2009; Schmitt, 2010). Validity arguments for vocabulary size tests (VSTs) built on the Frequency Model rest on evidence that word frequency aligns with item difficulty. Previous studies have found positive, though moderate, correlations between the two (Richards & Malvern, 2007; Milton & Daller, 2007; Hashimoto & Egbert, 2019), suggesting that other lexical features such as part of speech (POS), complexity, and polysemy also contribute (Hashimoto, 2021). Despite its importance in lexical processing (Faroqi-Shah et al., 2021; Zareva, 2011), POS remains underexplored. This study examined whether frequency effects on item difficulty vary by POS in L2 French vocabulary testing, given different form–function mappings across categories.

Data came from 487 students completing the vocabulary section of an L2 French placement test, comprising fifteen definition-matching items across 1K–5K frequency bands (Lonsdale & Le Bras, 2009) and three POS (nouns, adjectives, verbs). After excluding the overly easy 1K band, a rating scale IRT model yielded reliable item difficulties, analyzed using a two-way ANOVA. Frequency and its interaction with POS showed significant and large effects ($\eta^2_p = .66, .42$), while POS had a marginal but moderate effect ($\eta^2_p = .21$). Item difficulty increased as frequency decreased, most strongly for nouns. Findings support frequency-informed test design while suggesting word class–specific frequency patterns for improved construct validity, reflecting varying lexical learning demands across POS, but further examination on different L2 tests is needed.

Applicability of Dynamic Assessment in EFL Classrooms: An Investigation of Teachers' Perceptions

Author: Reza Monfared (Laval University)

Day: Friday, June 5 (9:40 a.m. - 10:10 a.m.)

Location: Symphonie 4

Dynamic Assessment (DA), grounded in Vygotskian Sociocultural Theory, integrates assessment and instruction to promote learner development. Despite its global recognition, its local enactment in EFL classrooms remains complex, raising questions about feasibility and ecological validity. This study investigated how EFL teachers perceive, interpret, and implement DA principles within their socio-educational contexts.

Using an explanatory sequential mixed-methods design, the study examined the applicability of interactionist DA operationalized through Feuerstein's Mediated Learning Experience (MLE), a set of 12 guiding attributes. Phase I involved workshops introducing DA, mediation, and MLE attributes via demonstrations, reflection, and collaborative lesson planning. In Phase II, teachers' classroom practices were observed with a structured MLE-aligned scheme, and questionnaires captured perceptions of DA applicability. Phase III employed focus-group interviews to examine teachers' experiences, challenges, successes, and adaptations. Qualitative data were coded in NVivo and triangulated with quantitative findings.

Results indicated teachers implemented MLE attributes they could enact independently, while those requiring high collaboration or institutional flexibility were more challenging. DA fostered participation, autonomy, and confidence, producing small but meaningful shifts toward interactive and reflective classroom dynamics. Teachers adapted DA selectively, integrating group-mediated interaction and L1 support, while contextual barriers, such as time constraints, class size, rigid curricula, and cultural sensitivities, limited consistent practice.

The study suggests that DA is feasible when applied flexibly and contextually. Broad adoption in EFL contexts requires sustained, practice-oriented teacher education, institutional support, and integration of L1 and technology-assisted mediation. Future research should explore long-term implementation and learner outcomes to inform sustainable classroom assessment practices.

Evaluating a Simulation-Based German for Medical Purposes Test: A Many-Facet Rasch Model Analysis

Author: Laura Mair (digiFORT Gesundheit und Pflege GmbH/University of Munich)

Day: Friday, June 5 (9:40 a.m. - 10:10 a.m.)

Location: Ovation

In Germany, internationally trained doctors must pass a German for Medical Purposes test to obtain professional licensure. This examination must include three components (doctor–doctor communication, doctor–patient communication, and the writing of a doctor's letter) and is commonly based on the Objective Structured Clinical Examination format („OSCE“; Harden et al. 1975; Patrício et al. 2013). The OSCE format enables authentic simulation of medical communication tasks, typically involving patient cases enacted by trained actors.

However, research on German for Medical Purposes testing remains scarce (Lenz et al. 2019; Jehle 2021a; Schön 2022), and no study to date has examined the role of OSCE cases within this context. If each OSCE case can be considered a task prompt eliciting a communicative performance, it is essential to ensure that these cases do not introduce construct-irrelevant variance in candidates' scores.

This study addresses this gap by applying the many-facet Rasch model („MFRM“; Linacre 1994) to data from FaMed, the first scientifically evaluated German for Medical Purposes test (Lenz et al. 2019). The analysis focuses on the doctor–doctor communication component (“patient presentation”), where candidates present clinical cases to an actor playing a senior physician. A total of 906 performances, based on 28 randomly selected OSCE cases and rated by 16 trained raters, were analyzed. Results indicate that the OSCE cases showed good model fit, though some rater misfit occurred. These findings are highly relevant not only to the context of German for Medical Purposes testing, but also to simulation-based LSP assessments more broadly.

Paper Session 4.2 (Day 2: 10:15 a.m. - 10:45 a.m.)

Defining and Operationalizing GenAI Application Skills in Academic Writing: An AI-Mediated Performance Task

Author: Shuhui Yin (Iowa State University)

Day: Friday, June 5 (10:15 a.m. - 10:45 a.m.)

Location: Opera AB

This study explores how the construct of GenAI application skills in academic writing could be conceptualized and operationalized through an AI-mediated writing assessment. GenAI tools are reshaping scholarly publishing and research practices in brainstorming, summarizing, drafting, and improving research writing. However, human-AI collaboration introduces complexities and ambiguities requiring AI literacy development among academic researchers. This study contributes to addressing the gap in AI-literacy assessment by moving beyond global definitions (AI knowledge, AI skills, AI ethics) and self-report assessment to context-specific measurement of GenAI application skills in academic research writing.

Framed in the argument-based validation framework (Kane, 2006; Chapelle, 2021), this study was conducted in two phases, each focusing on different validity arguments. Phase one described the domain of GenAI application skills in research article writing and developed a GenAI-mediated discussion writing task. Phase two collected validity evidence for the evaluation, generalization, and explanation inferences. A total of 30 test takers were tasked with composing a 120-word discussion paragraph with the help of ChatGPT within 50 minutes, with screen recordings capturing complete writing processes. Four trained raters evaluated performances using a five-category analytical rubric. Results will be reported in three main areas: (1) Many-facets Rasch measurement analyzed scores across raters, test-takers, and rubric categories, (2) descriptive statistics of survey results provided insights into test takers' perceptions of the test, and (3) thematic analysis of test takers' overall impression of the test and perceptions of test administration, authenticity, usefulness, as well as suggestions for improvements.

Meaningful Measures: How well can computer-based speaking assessments measure real-world communicative language proficiency?

Authors: Anastasia Ulicheva (Pearson), Ellen Barrow (Pearson), Sarah R. Hughes (Pearson)

Day: Friday, June 5 (10:15 a.m. - 10:45 a.m.)

Location: Symphonie 1

This presentation reports on mixed-methods validation research investigating two new speaking tasks introduced to a high-stakes computer-based English language proficiency test. The tasks, Summarise Group Discussion and Respond to a Situation, were designed to reflect real-world communicative demands, aligning with broader efforts to increase domain relevance in large-scale language assessments used for academic entry and skilled migration purposes. Computer-based tests generally do not offer face-to-face interaction opportunities; however, if designed well, tasks can elicit cognitively demanding and linguistically complex responses, targeting communicative language skills such as discourse management, pragmatic flexibility, interactional competence, and synthesis (Council of Europe, 2020).

Grounded in Weir's (2005) Socio-Cognitive Framework, this research combines three strands of enquiry: (1) investigating test-taker strategies and cognitive processes while responding to tasks, (2) exploring test-taker perceptions of task relevance to authentic language use contexts, and (3) analysing field test data (N = 998) to determine the extent to which new tasks enhance score differentiation compared to conventional speaking tasks.

By triangulating test taker perceptions, cognitive processes, and performance data, the findings from this research demonstrate the potential of computer-based speaking tasks to tap into communicative skills relevant to academic and skilled professional domains, refining measurement precision for advanced proficiency. These findings challenge assumptions about the limitations of computer-based speaking assessment, suggesting well-designed tasks can provide meaningful measures of real-world communicative language ability.

Multi-stage development of an L2 prosody scoring rubric for Chinese learners of English

Author: Yuanyue Hao (University of Oxford)

Day: Friday, June 5 (10:15 a.m. - 10:45 a.m.)

Location: Symphonie 2

Prosodic features, such as lexical stress, intonation, and rhythm, have been found to be significant predictors of second language (L2) English speech comprehensibility and communicative success. Some global pronunciation assessment criteria such as nativelikeness and comprehensibility are inherently biased and too vague to provide detailed feedback to language learners. Therefore, this paper attempted to develop an L2 English prosody rubric with specific criteria which can provide feedback to language learners to act upon. The study triangulated the criteria by using different methods, including meta-analysis, expert review, think-aloud protocols, and psychometric analysis.

First, meta-analysis was conducted to identify prosodic features that were significantly correlated with overall L2 speech quality. Effect sizes ($k = 494$ from 101 primary studies) were pooled using multi-level meta-analysis models for each prosodic feature. Seven features were identified and selected as the criteria, in reference to the autosegmental-metrical theory of intonational phonology. The drafted criteria were then subject to review by experts specializing in speech prosody. They agreed that these criteria are relevant to and representative of the prosody construct and raised some suggestions to revise the rubric. Finally, the criteria were used by 15 raters to assess L2 English prosody in 160 speech samples. Many-facets Rasch modelling was conducted to examine psychometric properties of each criterion. Results suggested that the proposed criteria are theoretically relevant and psychometrically fit. This study aims to develop an L2 English prosody scoring rubric with specific criteria to address limitations of commonly used global assessment standards.

Derivational Morphology Gaps Among EFL Learners: Implications for Population-Specific Vocabulary Assessment

Authors: Mastoor Al-Kaboody (Carleton University), Geoffrey G. Pinchbeck (Carleton University)

Day: Friday, June 5 (10:15 a.m. - 10:45 a.m.)

Location: Symphonie 3

Vocabulary assessment has long relied on the assumption that knowing a base word implies knowledge of its inflectional and derivational forms. This assumption is embedded in word-family frameworks and operationalized in widely used tests such as the Vocabulary Size Test and the Updated Vocabulary Levels Test. While efficient, it risks overstating lexical knowledge for learners whose first language morphology diverges from English. Arabic presents a test case, as its root-and-pattern system differs fundamentally from English affixation.

This study examined whether knowledge of a base form predicts knowledge of its derivational counterparts among Arabic-speaking learners of English and how proficiency moderates this relationship. A 76-item meaning-recall test (English->Arabic translation) was administered to 108 learners in the Middle East, spanning CEFR levels from Below A1 to B1+. Items included 30 base words and 46 derivational forms representing the most common English affix types (Laufer & Cobb, 2021; Sánchez-Gutiérrez et al., 2018). Responses were scored by four trained raters and analyzed using a generalized linear mixed-effects model with random intercepts for participants and items and a Rasch-calibrated index of item difficulty (Brysbaert et al., 2020). Learners were significantly more accurate on base forms than on derivations, $B = -1.32$, $SE = 0.28$, $OR = 0.27$, $p < .001$. Estimated marginal means showed predicted accuracies of 0.83 (base) versus 0.47 (derivation) for A1 learners and 0.99 versus 0.91 for B1+ learners. These results question the construct validity of word-family-based measures and highlight the need for population-specific lexical units in assessing vocabulary knowledge

Emerging Language Assessment Literacy in Preservice Teacher Practicum: Challenges and Insights from a Brazilian Context

Author: Vanessa Borges-Almeida (University of Brasilia)

Day: Friday, June 5 (10:15 a.m. - 10:45 a.m.)

Location: Symphonie 4

This presentation reports on an interpretive-qualitative ethnographic study that explores how language assessment literacy (LAL) (Fulcher, 2012; Inbar-Lourie, 2008) develops during pre-service language teacher education. Although LAL requires formal curricular initiatives, it is also a continuous, socially constructed process shaped by authentic contexts (Xu & Brown, 2016; Inbar-Lourie, 2008) rarely found in preservice education, except for the mandatory teaching practicum. Despite its relevance, research on LAL in practicum contexts remains scarce (Penna, 2024; Brum-Barreto & Borges-Almeida, 2023). This presentation discusses the nature of important challenges preservice teachers face when conducting assessment processes in real contexts and seeks to provide subsidies to refine the language assessment course that is offered prior to the practicum. The study involved eleven pre-service teachers enrolled in a public university who were engaged in a one-semester practicum. Data were collected through semi-structured interviews, self-assessment forms, practicum portfolios, and the tests designed by participants. Grounded theory (Hadley, 2017), thematic content analysis (Bardin, 1977), and reverse engineering (Carroll, 1993) were employed. Results revealed that a key area that requires attention comprises the knowledge, skills and principles which mostly affect validity: a clear comprehension of what to assess needs to emerge from a genuine embracement of language as use. Participants recognize their major difficulties as relating to selecting content, setting item difficulty, establishing assessment criteria, and addressing grading and scoring. These challenges were linked to difficulty in defining constructs. The study highlights the need to refine language assessment courses by integrating a tiered construct-focused approach to LAL.

Reducing Avoidable Research Waste in AI-Driven Language Assessment: Is Radical Transparency the Panacea?

Authors: Talia Isaacs (University College London), Hamish Chalmers (University of Oxford)

Day: Friday, June 5 (10:15 a.m. - 10:45 a.m.)

Location: Ovation

Language assessment scholarship has emphasized the importance of contextualizing assessment practices within their pedagogical, social, and political environments (Dimova et al., 2022). This shift calls for validation evidence that supports score interpretations across diverse settings to support responsible test use (Kane, 2012). A related but underexamined issue is the emergence of ‘research waste’—a concept from health sciences denoting inefficient research practices that squander resources yield minimal benefit (Chalmers & Glasziou, 2009). The research waste typology includes irrelevant or poorly justified research questions, flawed methodologies, unduly burdensome regulatory processes, and poor dissemination practices (e.g., selective reporting; Macleod et al., 2014), with applications for applied linguistics (Plonsky et al., 2025).

This presentation explores how generative AI (GenAI) may contribute to research waste in language assessment. While GenAI has already been transformational in test development and validation, with greater further potential, its rapid adoption raises concerns: the ‘Blackbox problem,’ biased training data, reproducibility issues, and machine-driven assessment, which may subordinate construct considerations. GenAI can also facilitate redundant or “soulless” outputs, including duplicate publications that distort the evidence base and strain limited editorial and reviewer resources.

Beyond technical limitations, GenAI use prompts deeper epistemological and ethical questions (Harding, 2025), including diminished human judgment and autonomy. In high-stakes contexts, these risks are amplified, potentially undermining trust. Addressing these challenges while maximising the benefits of GenAI requires a cultural shift toward radical transparency: clear reporting of AI use, meticulous documentation of prompt and model parameters, and engagement in open science practices. Such practices can mitigate drivers of avoidable research waste and uphold the integrity of language assessment scholarship.

Paper Session 4.3 (Day 2: 10:50 a.m. - 11:20 a.m.)

Diagnosing rater effects in writing assessment: Comparing three analytical frameworks

Authors: Mark Elliott^{1,2}, Nadir Zanini¹, Madeleine Austen¹

Organizations: 1: Cambridge University Press & Assessment, 2: Department of Computer Science & Technology, University of Cambridge

Day: Friday, June 5 (10:50 a.m. - 11:20 a.m.)

Location: Opera AB

Non-standard interpretations of assessment scales, or rater effects pose a threat to the validity of the assessment. Beyond severity and leniency, common non-uniform behaviours include central tendency (compressed use of rating scales) and halo effect (flattened score profiles). These behaviours can distort test scores, making their statistical identification crucial for maintaining assessment integrity.

This study investigates rater effects in large-scale writing assessments using three analytical frameworks. First, descriptive and inferential statistical methods (such as visualisations, non-parametric correlations, and regression) are applied to identify severity, central tendency, and halo effects. Second, the many-facet Rasch model (MFRM) is used to analyse rater behaviour from a measurement perspective. Third, extended MFRMs are employed to provide more nuanced rater profiles, distinguishing systematic from random error and modelling non-uniform behaviours more precisely.

Our findings reveal that extended MFRMs offer the most detailed insights, especially for detecting central tendency and halo effects. They outperform simpler methods by avoiding misleading conclusions based on macro-level indicators. However, they have more demanding data requirements, requiring multi-marking.

The study highlights that in diagnosing rater effects there is a trade-off between data requirements and inferential power. While under certain conditions, and especially in single-marking contexts, simpler statistical methods can offer useful preliminary insights, they may serve best as triggers for deeper analysis using MFRM or extended MFRM. This layered approach supports scalable, routine monitoring while enabling forensic investigation when needed, offering practical guidance for test providers seeking to balance operational feasibility with diagnostic precision.

When Emotion Meets Cognition: An Eye-Tracking Investigation in L2 Summary Writing

Authors: Yang Zhang (Zhejiang University), Lin Shi (Zhejiang University), Lianzhen He (Zhejiang University)

Day: Friday, June 5 (10:50 a.m. - 11:20 a.m.)

Location: Symphonie 1

This study investigates how emotional and cognitive factors jointly shape test-takers' performance in an integrated summary writing task. While previous validation research in language testing has primarily emphasized cognitive processes, affective influences—such as enjoyment and boredom—remain largely unexamined. Grounded in Pekrun's Control-Value Theory (2006), this study explores how emotions modulate cognitive engagement and subsequent writing outcomes, thereby extending the construct of validity to encompass both cognitive and affective dimensions.

Thirty-three Chinese EFL learners completed an English summary writing task while their eye movements were recorded using a Tobii TX300 eye tracker. Three areas of interest (AOIs)—the source text, notes, and writing area—were defined to capture distinct cognitive phases. Relative fixation durations were computed to represent attention distribution across these regions. Following task completion, participants rated their levels of enjoyment and boredom on standardized emotion scales (Li et al., 2023). English proficiency was controlled, and writing samples were double-rated (inter-rater reliability = .87).

Correlational analyses revealed that enjoyment and source-text fixation proportion were both positively related to writing scores, while boredom correlated negatively. Enjoyment was further associated with greater attention to source-text comprehension. Hierarchical regression showed that proficiency and source-text fixation proportion explained 40% of score variance. A mediation analysis (PROCESS Model 4; Hayes, 2022) indicated a marginal indirect effect of enjoyment on performance via source-text fixation proportion.

Findings highlight the intertwined roles of affect and cognition in test performance, offering empirical evidence for emotion–cognition integration in assessment validity. The study underscores the value of eye-tracking as a methodological bridge linking cognitive and affective processes in language testing research.

Uncovering Panelist Decision Strategies: Insights from Mapping CELPIP Test Scores to CEFR Standards

Authors: You-Min Lin (Prometric), Michelle Chen (Prometric)

Day: Friday, June 5 (10:50 a.m. - 11:20 a.m.)

Location: Symphonie 2

This study investigates expert panel decision strategies in mapping CELPIP test scores to CEFR proficiency levels. Data was drawn from four CELPIP standard-setting panels (Listening, Reading, Speaking, Writing), each setting cut scores at four CEFR levels (B1, B2, C1, C2) using either the Modified Angoff or Bookmark method. Panelists collaboratively defined the “minimally competent candidate” (MCC), independently judged test items or performances, then engaged in group discussion and consensus-building before finalizing decisions. Panel transcripts were analyzed to address three questions: (1) What strategies do panelists employ to conceptualize the MCC? (2) What specific strategies and reasoning processes guide panelists’ cut-score decisions? (3) How do panelists’ decision strategies differ in Modified Angoff and Bookmark standard-setting contexts? Thematic analysis was conducted using a multi-domain coding scheme developed inductively from the data, distinguishing MCC conceptualization, decision strategies, and interactional dynamics. Findings show that panelists negotiated between global CEFR descriptors and local assessment evidence, drawing on professional and practical test-taker knowledge to construct working MCC models. Cut-score decisions unfolded through layered reasoning spanning evidence appraisal, perspective-taking, and social and evidential calibration, reflecting the cognitive and social complexity of standard-setting discourse. Discussions improved consistency but revealed tensions between individual expert judgement and collective influences. Method-specific pathways were observed across Modified Angoff and Bookmark contexts, showing notable differences in panelist engagement with assessment content, judgement calibration, and decision negotiation. The study contributes a layered analytical framework for standard-setting process research, providing implications for configuring panel procedures that support alignment with external language standards.

Where measurement meets machine learning: Explaining item difficulty with examples from L2 vocabulary assessment

Authors: Karen Jeanette Dunn (British Council), Lucy Skidmore (British Council), Thomas Rogers (British Council)

Day: Friday, June 5 (10:50 a.m. - 11:20 a.m.)

Location: Symphonie 3

Estimation of item difficulty is essential in language test development, but recent attention has shifted toward the need also to explain and predict it. This has practical implications for item development, adaptive testing, and construct validation. Measurement specialists have traditionally explored factors contributing to item difficulty through explanatory item response models (EIRMs). Yet, in language assessment, explaining difficulty remains challenging due to the complex, context-sensitive nature of linguistic constructs. Advances in machine learning (ML) and Natural Language Processing (NLP) have expanded possibilities, offering scalable and flexible solutions, but may compromise interpretability, i.e., the capacity to link results to the underlying ability construct. In sensitive areas, such as immigration and citizenship, generating validation evidence is critical. A pressing need therefore exists to understand the implications of using ML models in this context.

This conceptual paper explores the meeting ground between measurement and machine learning, examining how these traditions converge and diverge in modelling item difficulty. Empirical illustrations are drawn from explanatory modelling of vocabulary test difficulty using the British Council’s Knowledge-based Vocabulary Lists (KVL; Schmitt et al., 2024), a crowd-sourced dataset of form-based English language word recall from Mandarin, German, and Spanish L1 groups. Explanatory IRT approaches (Wilson & de Boeck, 2004; Dunn, 2024) which employ observable item features are contrasted with NLP-based methods using transformer architectures (Skidmore et al., 2025). Trade-offs between construct-based rigour and scalable application are highlighted, and implications discussed in the light of the increasingly interdisciplinary nature of this field, including possibilities offered by hybrid IRT-ML solutions.

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

Author: Coral Yiwei Qin (University of Ottawa)

Day: Friday, June 5 (10:50 a.m. - 11:20 a.m.)

Location: Symphonie 4

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 Meta-Analysis of Reported Agreement Between Large Language Models and Human Raters in Essay Scoring

Authors: Hongli Li¹, Che Han Chen¹, Chiho Young-Johnson¹, Soyoung Lim¹, Yali Feng¹, Kevin Fan¹

Organization: 1: Georgia State University

Day: Friday, June 5 (10:50 a.m. - 11:20 a.m.)

Location: Ovation

Modern transformer-based large language models (LLMs) are increasingly used for automated essay scoring (AES), yet their agreement with human ratings remains uncertain. Huang et al. (2025) conducted an early review of 29 studies published between 2023 and July 2025. Based on qualitative observations, they concluded that general language models "hold promise to enhance [automatic writing evaluation]" (p. 1). However, no meta-analysis has yet quantified the average level of agreement between LLM and human ratings. We present a meta-analysis to synthesize LLM-human agreement in essay scoring. Specifically, we ask two research questions: (1) What is the pooled level of agreement between LLMs and human raters in essay scoring? (2) Which study- and design-related factors are associated with the level of agreement? Our primary outcome is quadratic-weighted kappa (QWK), the most consistently reported agreement metric in this literature. To account for the multiple outcome variables from a single study, we use robust variance estimation to adjust standard errors with R Package robumeta. We first estimate the pooled LLM-human agreement across studies and then conduct a moderator analysis to examine which study- and design-related factors are associated with higher or lower agreement. This meta-analysis provides an initial, quantitative summary of LLM-human agreement in AES. The findings also offer some early insights into where agreement is stronger or weaker, informing future research design and reporting.

Paper Session 5.1 (Day 2: 3:30 p.m. - 4:00 p.m.)

Diagnosing Language Ability with China's Standards of English Language Ability: A Cross-sectional Comparison across Educational Stages

Authors: Hui Wang (Zhejiang University), Shangchao Min (Zhejiang University)

Day: Friday, June 5 (3:30 p.m. - 4:00 p.m.)

Location: Opera AB

The present study employed cognitive diagnostic modeling to examine variations in language subskills between junior high school and college students in the Chinese EFL context, aiming to advance our understanding of developmental patterns in language ability. Specifically, we developed two cognitive diagnostic tests aligned with China's Standards of English Language Ability: a 54-item version for junior high school and a 59-item version for college students, linked by 12 anchor items. Both tests targeted five attributes, including organization (A1), cohesion (A2), phonology (A3), vocabulary (A4), and syntax (A5). A total of 1,102 Grade 8 students and 948 sophomores participated in the tests. Using the G-DINA package in R, we estimated item parameters, attribute prevalence, and mastery profiles, and then compared these indices between groups. We also conducted correlation analyses to examine the interrelationships among subskills. The results revealed a consistent pattern of attribute prevalence across both groups ($A1 > A2 > A4 > A3 > A5$). Sophomores outperformed Grade 8 students on all attributes, with significant differences observed in A1 and A3. Mastery profile comparisons showed that Grade 8 students exhibited a polarized distribution focused on non-mastery and near-complete mastery, while sophomores displayed more varied and sequential mastery patterns. Correlation analysis showed that sophomores exhibited systematically weaker correlations than Grade 8 students across all attribute pairs, suggesting a move from general-ability dominance to differentiated skills as proficiency rises. The findings of this study provide insights into subskill development and underscore the need for stage-sensitive instructional practices tailored to learners' developmental needs.

Language Assessment Literacy among Colloquial Arabic Instructors: Convergence and Divergence between Local Practices and Global Assessments

Author: Hadeel Arqbi (Georgetown University)

Day: Friday, June 5 (3:30 p.m. - 4:00 p.m.)

Location: Symphonie 1

This study aimed to explore LAL among Arabic instructors in US postsecondary institutions who teach Colloquial Arabic. While research on assessing Modern Standard Arabic (MSA) exists, there is limited discussion on evaluating Colloquial Arabic (CA) (Raish, 2021). The integration of CA with MSA in teaching has been extensively debated, yet the assessment of Colloquial Arabic is often overlooked in these discussions (Norris & Raish, 2017). To this end, this research focused specifically on Colloquial Arabic instructors' perceptions of their LAL and their classroom assessment practices. Using Taylor's (2013) language assessment literacy (LAL) framework, this research explores instructors' perceived competence in assessing CA. A mixed-methods approach was used, collecting data through a 66-item questionnaire, partially adapted from Kremmel and Harding's (2020) survey, followed by semi-structured interviews. Participants were recruited via convenience sampling and snowball sampling across various dialect groups.

CA instructors reported a high level of LAL, which aligns with their profiles as experienced, highly educated Arabic language teachers. They also mentioned using OPI in conjunction with other classroom assessments. Their professional OPI training affects how they evaluate CA. Interview data provided further insights into their assessment practices and demonstrated how their perception of CA influences their methods. Factors such as programming policies and student profiles also impact their decisions. These findings can inform best practices for assessing CA and influence assessment approaches in other spoken-language contexts across the U.S. and globally, leading to more accurate evaluations of Colloquial language proficiency.

Text Ease and Readability Assessment of Human-versus ChatGPT-Generated Listen-to-Summarize Conversation Scripts

Authors: Lixia Cheng (Purdue University), Rodrigo A. Rodríguez-Fuentes (Universidad del Norte)

Day: Friday, June 5 (3:30 p.m. - 4:00 p.m.)

Location: Symphonie 2

Locally developed language tests are often preferred in ESL/EFL teaching contexts due to their alignment with program goals, cost-effectiveness, and construct control. Integrated speaking tasks, essential for assessing productive language skills, are resource-intensive to design, presenting challenges for small programs with limited budget and staffing. Recent studies suggest that ChatGPT can accelerate test development, particularly in generating large amounts of language input. However, little research has examined the linguistic and textual properties of materials generated by ChatGPT versus those created by human item writers. This study compared the text “easibility” of conversation scripts—written by ChatGPT versus 41 CEFR C1-level upper-division English majors at a Colombian university—for a Listen-to-Summarize speaking task on a local English test. Text “easibility” was analyzed using Coh-Metrix, measuring narrativity, syntactic simplicity, word concreteness, referential cohesion, and deep cohesion. A repeated-measures MANOVA revealed significant differences between ChatGPT- and human-generated scripts across the combined readability measures. Follow-up repeated-measures ANOVAs showed that ChatGPT scripts included more concrete words and syntactically simple sentences but received lower narrativity and referential cohesion scores. Each of the four Bonferroni-corrected ANOVAs yielded large to very large effect sizes. However, no significant difference was found for deep cohesion between the two sources. The average Flesch-Kincaid grade level indicated that ChatGPT scripts were a grade level more difficult (6.83 vs. 5.70). Practical implications include confirming ChatGPT’s potential in test development—particularly for scalability and creation of parallel test forms—and highlighting the need for AI literacy among pre-service teachers for efficient and effective development of local language tests.

Global Standards, Local Practices: Rethinking NoteTaking Modes in Language Testing

Author: Jieun Kim (Soongsil University)

Day: Friday, June 5 (3:30 p.m. - 4:00 p.m.)

Location: Symphonie 3

Technological advancements have influenced note-taking practices in classrooms as well as their treatment in standardized language assessments. Many high-stakes English listening tests (e.g., TOEFL iBT, IELTS) include varying note-taking guidelines, often established without empirical support. This study examined the effects of note-taking modes on listening performance and note content. A total of 305 L1-Korean L2-English adults were randomly assigned to handwriting (n = 102), typing (n = 102), or no note-taking (n = 101) conditions and completed TOEFL iBT lecture comprehension tasks. Linear regression results revealed no significant differences in overall test scores across modes. Uniform DIF analyses indicated that one item favored handwriting and another favored typing, while non-uniform DIF emerged for the same items only among lower-ability participants. When examining note quality measures, no significant differences were found in word count or translanguageing across modes. However, handwritten notes contained more information units, verbatim transcription, and nonlinguistic elements. The overall lack of significant performance differences, together with finer-grained differences at the item and note-content levels, calls for test developers to reconsider note-taking policies in terms of test validity and fairness.

Language Assessment Literacy In Practice: EFL Teachers' Implementation of Classroom-Based Assessments in Chile

Author: Evelyn Diaz-Iturriaga (Georgetown University)

Day: Friday, June 5 (3:30 p.m. - 4:00 p.m.)

Location: Symphonie 4

This study investigates how in-service English as a Foreign Language (EFL) teachers in Chile enact classroom-based assessment (CBA) and how their practices reflect Language Assessment Literacy (LAL). While LAL has been widely conceptualized as comprising knowledge, skills, and principles, most research has relied on teachers' self-reports rather than observing assessment in action. Framed within Hill and McNamara's (2012) view of assessment as a sociocultural practice inseparable from teaching, the study positions teachers as assessment agents who interpret policy, make context-sensitive decisions, and act on classroom evidence.

To document assessment in context, the researcher developed the Language Assessment Practices Observation Tool (LAPOT), an instrument encompassing 28 observable behaviors across four domains—planning, framing, conducting, and using assessment. Six Chilean EFL teachers from public, private, and subsidized schools were each observed across five consecutive lessons, complemented by fieldnotes and interviews with teachers and principals. Data analysis combined structured coding and qualitative content analysis to triangulate reported beliefs, observed behaviors, and contextual influences.

Findings show teachers regularly planned lessons and monitored learning but rarely made success criteria explicit. Feedback was mainly corrective and delivered in students' L1, limiting language use opportunities. Inclusive assessment practices were scarce, suggesting partial enactment of LAL—strong in procedural knowledge but weaker in principles such as fairness and transparency. The observation process itself fostered teachers' reflection, demonstrating LAPOT's potential as both a research and professional development tool. Overall, the study contextualizes LAL within Chilean classrooms, highlighting teachers' agency and the systemic constraints shaping assessment practice.

Paper Session 5.2 (Day 2: 4:05 p.m. - 4:35 p.m.)

Validating the Elicited Imitation Task through Eye-Tracking Evidence

Authors: Yulin Pan (University of Illinois Urbana-Champaign), Huiying Cai (University of Illinois Urbana-Champaign), Amanda Li (University of Illinois Urbana-Champaign), Angie Lee (University of Illinois Urbana-Champaign)

Day: Friday, June 5 (4:05 p.m. - 4:35 p.m.)

Location: Opera AB

The Elicited Imitation Task (EIT) has gained popularity in local and institutional testing contexts due to its practicality and high reliability. Yet, a persistent challenge in validating the EIT lies in determining whether repetition performance reflects genuine language proficiency rather than short-term memory or simple parroting. To address this concern, the present study integrates process-level evidence using eye-tracking within a Visual World Paradigm (VWP) to examine whether test-takers engage in meaning-based comprehension during EIT performance.

Thirty-two participants (14 L1 Chinese speakers, 9 L2 learners, 8 heritage speakers) from a U.S. university completed a 50-item Chinese EIT developed for a local proficiency test. Items varied in sentence length, curricular level, and target-noun concreteness. During each trial, participants listened to a sentence while viewing one semantically related target and three distractors, then repeated the sentence aloud and later selected the picture matching its meaning. Total fixation duration was measured during sentence processing to assess attention allocation.

Mixed-effects regression models examined both online gaze behavior and offline comprehension outcomes. Participants showed significantly longer fixation durations on target pictures than on distractors, demonstrating meaning-driven attention consistent with linguistic processing. Longer fixations on distractors were linked to more abstract target nouns and incorrect comprehension responses, reflecting increased processing difficulty. The alignment between gaze behavior and comprehension accuracy provides process-product consistency, offering novel process-level validity evidence that EIT elicits intended cognitive operations.

Ethical Dimension in the Language Assessment Practices of University Language Teachers in Mexico

Authors: Hugo A. Andrade-Mayer (Universidad Autonoma del Estado de Mexico), Virna Velazquez (Universidad Autonoma del Estado de Mexico), Georgia Grondin (Universidad Autonoma del Estado de Mexico)

Day: Friday, June 5 (4:05 p.m. - 4:35 p.m.)

Location: Symphonie 1

This study explores the ethical dimensions that shape the language assessment practices of university language teachers in Mexico. It responds to McNamara and Roever's (2006) call for social responsibility in language testing, extending it to classroom contexts where teachers design and apply their own assessments. While studies in other regions have examined assessment literacy (Vogt & Tsagari, 2014; Wang, 2017), little is known about how ethics influences assessment in Mexican higher education. The study investigates the factors guiding teachers' decisions when planning, designing, and grading, and how ethical concerns are intertwined with pedagogical and technical aspects.

Using a qualitative grounded theory approach (Strauss & Corbin, 1998), semi-structured interviews were conducted with twelve language teachers at a state university. Data were transcribed, coded, and analyzed with Atlas.ti, leading to five emerging dimensions: linguistic, technical, pedagogical, personal, and ethical. The ethical dimension proved central, encompassing notions such as care for students, fairness, inclusion, and responsibility. Two subdimensions were identified: interpersonal ethics, relating to relationships and fairness, and value-based ethics, involving moral reasoning and professional judgment.

Findings highlight assessment as a socially situated act (Bachman & Palmer, 1996; McNamara & Roever, 2006) and echo Morin's (2006) view of ethics as interconnection among individuals and communities. The study argues that ethical reflection should be integral to language teacher education, promoting more responsible, humane, and context-sensitive assessment practices.

Which AI should we use? Using corpus linguistics to compare and evaluate ChatGPT-and Claude-generated texts for TOEFL iBT's Writing for an Academic Discussion task

Author: Che-Han Chen (Georgia State University)

Day: Friday, June 5 (4:05 p.m. - 4:35 p.m.)

Location: Symphonie 2

As language tests introduce new writing assessments like TOEFL iBT's Writing for Academic Discussion (WAD) task, test preparation requires sufficient practice materials to guide students' preparation process (He et al., 2025). GenAI offers promising approaches to creating assessment tasks and content (Runge et al., 2025), increasingly employed to generate stimulus materials for both instructional and large-scale testing purposes (Lee et al., 2024; Xi, 2023). However, the assumption that GenAI produces register-appropriate texts requires empirical validation, since the quality of stimulus texts supports domain description inferences of a test (Cumming et al., 2021). Given register's role in predicting TLU linguistic variation (Kyle et al., 2022), the success of AI-produced texts depends on the degree to which AI produces language appropriate to TLU registers.

This study compared three WAD stimulus text corpora: 50 human-authored texts were compiled from official TOEFL materials and test preparation resources. Texts from ChatGPT-4o (N=50) and Claude-opus 4.1 (N=50) were generated using few-shot prompt engineering through APIs in a Python program. Register alignment was evaluated through additive multi-dimensional analysis based on 67 lexico-grammatical features from Biber's original factor analysis (1988).

Results show that neither AI system fully captured human-authored register characteristics. Claude-opus 4.1 produced excessively formal, information-dense texts lacking the interactional and involved features typical of online academic discussions. ChatGPT-4o significantly underperformed in generating WAD texts with persuasive language. These findings indicate that current generative AI tools lack sufficient contextual awareness to produce register-appropriate WAD source texts. Implications for test preparation materials designers will be discussed.

Exploring the relationship between self-assessment and objective measures of reading and listening abilities among young Spanish learners of English

Authors: Renka Ohta (Educational Testing Service), Ching-Ni Hsieh (Educational Testing Service)

Day: Friday, June 5 (4:05 p.m. - 4:35 p.m.)

Location: Symphonie 3

Research has shown that self-assessment (SA) can be a useful alternative assessment tool to help young language learners (YLLs) develop their language skills, and that YLLs can meaningfully reflect on their own language abilities when provided with developmentally appropriate can-do descriptors and SA training and practices (Butler, 2018). This study focused on the use of SA among YLLs who are underrepresented in previous research—young Spanish learners of English at the primary level. We explored the relationship between SA and objective measures of reading and listening skills using a standardized language proficiency test. Participants were 110 fifth- and sixth-grade students, enrolled in an afterschool English language program in Mexico. The students completed the TOEFL Primary Reading and Listening test followed by the SA questionnaire that included 16 reading and listening can-do descriptors. The results showed that the students' SA ratings had satisfactory internal consistency for both reading and listening SA items. Results of descriptive statistics and one-sample t-tests indicated that students self-rated higher on basic comprehension tasks (e.g., understanding word definitions) and lower on more complex tasks (e.g., identifying main ideas or guessing word meanings). Spearman's rho correlation analyses showed statistically significant relationships between students' self-assessments and their TOEFL Primary test scores. Additionally, the students made better judgments of their abilities to perform more concrete and familiar tasks (e.g., Sentence Clue, Correspondence) than more advanced or complex ones (e.g., Expository Texts, Academic Monologue). We conclude with implications for the use of SA to support YLLs' language development.

From interactional data to measurement modelling: Investigating cognitive processing of L2 pragmatic competence in the design of a UK pre-sessional speaking test

Author: Shishi Zhang (University College London)

Day: Friday, June 5 (4:05 p.m. - 4:35 p.m.)

Location: Symphonie 4

This study investigates how second language (L2) learners cognitively process pragmatic competence in speaking through a mixed-methods approach. It is situated in the development of an EAP speaking test for UK pre-sessional programmes, targeting students progressing to postgraduate studies. The test comprises two task types: a group oral assessing pragmatic knowledge and use, and video analytical tasks assessing metapragmatic awareness, supported by task-type-specific analytic rating scales. Seventeen pre-sessional students' test-trial performance were qualitatively analysed using pragmatic- and metapragmatic-related episodes to capture enacted and reflective aspects of pragmatic competence. Interactional Sociolinguistics further illuminated how pragmatic meanings were negotiated and interpreted in discourse, offering a nuanced view of students' cognitive processing of pragmatic competence.

Complementing the qualitative results, Many-facet Rasch measurement was conducted on field testing scores from 52 students and three raters. Results showed satisfactory task and rating scale functioning and rater reliability, providing statistical evidence of cognitive validity of the test. Principal Component Analysis of Rasch residuals revealed that pragmatic knowledge and use and metapragmatic awareness were distinct yet related sub-dimensions of pragmatic competence. Pearson correlation between scores on the two tasks ($r = 0.322$, $p < 0.05$) suggested a significant but modest relationship, with only 10.37% shared variance.

The study highlights the complexity of translating L2 pragmatic competence into measurable test constructs, particularly in ensuring test design to reflect test-taker cognitive processing. Integrating qualitative and quantitative evidence, this study strengthened the test's cognitive validity and demonstrated a replicable approach to developing fair, valid, and contextually grounded EAP speaking assessments.

Can an LLM-powered spoken dialogue system accommodate L2 learners of varying proficiency levels?

Authors: Jing Xu¹, Masaki Eguchi², Yasin Karatay¹, Mao Saeki², Nicholas Glasson¹, Anastasia Ramjag², Annabelle Pinnington¹, Elliott Bruce², Evelina Galaczi¹, Max Thomson², Martin Robinson¹, Yoichi Matsuyama²

Organizations: 1: Cambridge University Press & Assessment, 2: Equimenopolis Inc.

Day: Friday, June 5 (4:05 p.m. - 4:35 p.m.)

Location: Ovation

Large language models (LLMs) have shown great promise in enabling human-AI interactions, as seen in LLM-powered spoken dialogue systems (SDSs) designed to play the role of interlocutors. While a large body of SDS research has focused on learner production, little is known about how well an SDS can accommodate L2 learners of varying proficiency.

This paper reports a small-scale study examining the extent to which an LLM-powered SDS adapts to learners' proficiency in an interactive speaking task. The study posed two research questions. First, how does the SDS provide interactional support in response to learners' understanding difficulties, and how might such support vary across different proficiency levels? Second, can the SDS adjust the complexity of its language based on learners' proficiency?

Two computer-based speaking tasks— one on monologic speaking and the other on interactive speaking with an SDS—were created. One hundred L2 English learners, whose English proficiency ranged from A2 to C2, completed both tasks. A smaller group ($n=15$) took part in follow-up interviews. Their oral responses were marked by four trained examiners who were also interviewed.

To address the first question, conversation analysis (CA) was conducted on 50 learner-SDS interactions, focusing on the SDS's supportive behaviours to resolve learners' understanding difficulties. To answer the second question, normalised linguistic complexity measures of the AI speech in all 100 interactions were compared across the five proficiency levels using ANOVA. In addition, the interview data were used to support, complement, and interpret the findings from the CA and statistical analysis.

Paper Session 5.3 (Day 2: 4:40 p.m. - 5:10 p.m.)

When Tests Meet Reality: Cognitive and Context Validity of the PTE Core for Canadian Immigration and Employment

Authors: Peiyu Wang (Queen's University), Liying Cheng (City University of Macau)

Day: Friday, June 5 (4:40 p.m. - 5:10 p.m.)

Location: Opera AB

PTE Core, recently approved by IRCC, is mapped to the CLB but developed with reference to the CEFR, raising concerns about its fit with Canadian immigration and workplace contexts. Its validity depends not only on psychometric reliability but also on whether tasks reflect authentic communicative demands (context validity) and elicit real-world cognitive processes (cognitive validity) (Weir, 2005). However, unlike academic testing, large-scale immigration testing has seldom been examined for these validity dimensions. Guided by Weir's socio-cognitive framework, this qualitative study examined how the PTE Core aligns with immigration and workplace communication across three occupational groups with differing English demands—high, mid, and low. Ten test-takers and twelve immigrant employees completed real-time think-aloud protocols on PTE Core tasks and one-on-one interviews reflecting on cognitive processing, linguistic demands, context relevance, and task alignment. Data were analyzed using inductive–deductive coding, cognitive process tracing (Ericsson & Simon, 1993), and comparative case analysis (Yin, 2014). Findings show that while PTE Core tasks were generally perceived as realistic, both context and cognitive alignment varied across occupational groups. High-demand professionals reported strong alignment with workplace communication, mid-demand participants noted that some reading and listening tasks appeared overly academic, and low-demand workers highlighted limited authenticity and absence of relevant tasks. These differences underscore the risks of “one-size-fits-all” interpretations and carry implications for test design, score interpretation, and immigration policy. The study provides validity evidence and a contextualized portrait of Canadian immigration and workplace communication, illustrating how assessments can better reflect diverse occupational language needs.

English Proficiency Test Exemption Policies in US Research-Intensive Universities

Authors: Nicholas Coney (University of Hawai'i at Mānoa), Daniel Isbell (University of Hawai'i at Mānoa)

Day: Friday, June 5 (4:40 p.m. - 5:10 p.m.)

Location: Symphonie 1

English proficiency tests are commonly used to make higher education admissions decisions for international students in Anglophone countries and in other contexts where English is the medium of instruction. However, not all international students are required to submit English proficiency test scores when applying to an English-medium university. It is common to find a number of exemption criteria, including (but not limited to) completion of English-medium secondary schooling or higher degree, citizenship linked to an Anglophone country, or being a ‘native’ (first language, L1) speaker of the language. Our motivation for the current study is to better understand English proficiency test exemption policies in US higher education, with a particular focus on country-based exemptions for citizenship and education. As a starting point, our investigation focuses on research-intensive universities (Carnegie R1 classification, $n = 146$). Preliminary results indicate that nearly all ($n = 137$) institutions offered explicit exemptions based on being a L1 speaker ($n = 12$), completing some amount of education through English ($n = 51$), or both ($n = 74$). While most of these institutions appear to handle L1 and education-based exemptions on a case-by-case basis, 61 institutions linked exemption criteria to specific countries/territories. Inner Circle (Kachru, 1985) countries were almost universally exempted and former colonies in the Caribbean frequently were, while former African colonies were less consistently exempted. Countries/territories in other regions were rarely exempted.

Surprisal as an Indicator of Linguistic Accuracy in L2 Writing

Author: Xuan Hu (University of Illinois at Urbana Champaign)

Day: Friday, June 5 (4:40 p.m. - 5:10 p.m.)

Location: Symphonie 2

Surprisal, a psycholinguistic metric quantifying the predictability of linguistic input, has been shown to distinguish L1 and L2 writing, capture developmental differences, and predict human-assigned writing scores (Cong, 2025). From an assessment perspective, linguistic accuracy reflects a writer's ability to produce fewer errors (Cumming et al., 2005) and to use language conforming to target-like norms (Plakans et al., 2019). Although surprisal has been useful in psycholinguistic research to measure L2 learners' control over linguistic accuracy, its potential to be applied in measuring accuracy in language assessment contexts remains less explored.

This study examined (1) whether surprisal predicts sentence-level error occurrence and frequency, and (2) which error types are associated with surprisal compared to those reflected in holistic writing scores. Data were drawn from the Corpus of Non-Native Written English (ETS). A random sample of 149 essays (2,075 sentences) was analyzed using an LLM-based pipeline combining a T5 GEC model, the ERRANT error annotation toolkit, and GPT-2 surprisal estimates. Mixed-effects modeling and correlation analyses were conducted to examine the associations among surprisal, error type and frequency, and writing proficiency level.

Results showed that surprisal significantly predicted both the likelihood and number of errors. Higher surprisal values were associated with more morphological, grammatical, lexical, and spelling errors. The pattern largely paralleled human raters' weighting of error gravity, except for spelling. The "GEC tool + surprisal" pipeline can provide evidence for scoring validity and has potential for integration into automated essay scoring.

Beyond the 'Benefit of the Doubt': An Eye-Tracking Study of Assessment-modality Effects on Rater Behaviour and Cognition

Author: Daniel Yu-Sheng Chang (University of Bristol)

Day: Friday, June 5 (4:40 p.m. - 5:10 p.m.)

Location: Symphonie 3

As digital technologies reshape language assessment globally, the format in which test-takers' responses are produced (handwritten vs. typed) and scored (human vs. automated) has become a pressing issue. However, many local examinations continue to rely on handwritten essays, with only gradual transitions toward typed responses (Coombe et al., 2020), and such scoring remains largely conducted by raters. This highlights the need to understand how assessment modalities influence local-cohort raters scoring.

This study extended prior rater research by explicitly linking rater behaviour and cognition through triangulated data sources. Raters analytically scored graph-integrated writing essays, with assessment modalities experimentally controlled. Data included their eye-movement data, scoring logs, responses of post-rating questionnaires, and interviews.

Consistent modality effects were found. Typed essays were significantly scored lower than handwritten counterparts for all four criteria, with severity varying across criteria. Substantial differences in severity also emerged across raters, reflecting rater-specific rather than systematic behaviours. Cognitively, raters invested significantly greater load when scoring handwritten essays. Assessment modalities nudged rater attention: handwritten essays elicited more focus on test-takers' responses, while typed essays prompted more frequent checks against the source use.

By linking behaviour and cognition, the study showed that typed essays led raters to begin scoring from salient features (errors, impressions). Moreover, rater severity related to the score adjustments per criterion. Beyond the long-standing 'benefit of the doubt' explanation, the study further extended: some raters indeed scored illegible essays leniently, while others penalised them more severely, reflecting greater cognitive load in handwritten responses.

Contextualising constructs: Applying Systemic Functional Linguistics to language assessment design and validation

Authors: Mark Elliott (Cambridge University Press & Assessment and University of Cambridge), Hye-won Lee (Cambridge University Press & Assessment)

Day: Friday, June 5 (4:40 p.m. - 5:10 p.m.)

Location: Symphonie 4

Recent scholarship in second language assessment has highlighted the need for contextually sensitive yet theoretically principled construct frameworks. Researchers have called for approaches that balance global comparability with local relevance; this tension has become a defining issue for the future of test design and validation.

Assessment design and validation can be greatly facilitated by a coherent framework within which to orient the features of the theoretical construct. This may be particularly relevant when developing tests for more specific purposes or contexts of use where certain aspects have greater or lesser salience than is the case for global, standardised assessments.

We draw from Systemic Functional Linguistics and pragmatics to develop a flexible high-level framework that permits an analysis of constructs that places local concerns within a universal framework of language use, and provides a tool for developing and validating content. We present a flexible framework for defining aspects of linguistic choice that may be used to help define or validate theoretical constructs.

We consider various use cases for the framework: for proficiency and summative assessment, for defining and validating assessment constructs and developing rating scales; for providing feedback to learners in learning-oriented assessment, and assisting in the development of syllabus design and pedagogy, and incorporating positive washback by design into assessments. This proposal reflects the field's recognition that universal constructs alone cannot fully capture the contextual realities of language use. By applying SFL to construct definition, the framework offers a principled way to localise assessments without compromising comparability or theoretical coherence.

Revisiting the Social Dimension of Foreign Language Tests: The Case of English Gaokao from 1950-1966 in China

Author: Jinjie Wu (Shanghai International Studies University), Mingwei Pan (Shanghai International Studies University), Xini Liao (Shanghai International Studies University)

Day: Friday, June 5 (4:40 p.m. - 5:10 p.m.)

Location: Ovation

While existing research has globally examined the interconnectedness of test items and educational policies (Shohamy, 2007), the historical evolution of high-stakes national-level tests in response to major political shifts remains underexplored. One example of such a test is China's National College Entrance Examination, or gaokao. English gaokao, for instance, had maintained a precarious position at its infancy from 1950-1966, when the vacillating diplomatic ties with Soviet Union and the West led to constant policy changes. Against this backdrop, this study investigated how English gaokao navigated this complex geopolitical landscape during this period.

Employing a document analysis approach, the study drew on two sets of materials: the annual English gaokao exam papers from 1950-1966 and relevant foreign language education policies. By integrating these sources, the current study aimed to answer one research question: What was the interaction between foreign language policy and English gaokao during 1950-1966 in China?

The findings demonstrated a direct, though lagged correlation between national policy and English gaokao. The Sino-Soviet alignment was characterised by the marginalisation of English, reflected in a flawed test design and saturated socialist ideology. The subsequent Sino-Soviet split elevated English as the primary foreign language and yielded a more comprehensive test design despite significantly constrained by the ongoing existence of ideological indoctrination. As the political climate relaxed in the mid-1960s, test developers displayed more initiative in developing an exam of better discriminatory power. This study offers insights into the social dimension of gaokao through a diachronic lens in a distinctive context.

Paper Session 5.4 (Day 2: 5:15 p.m. - 5:45 p.m.)

Bridging Proficiency and Pedagogy: Rethinking the validity of Chinese Placement Tests in Canada

Authors: Shujiao Wang (McGill University), Zhi Li (University of Saskatchewan), Min Sun (Université du Québec à Montréal)

Day: Friday, June 5 (5:15 p.m. - 5:45 p.m.)

Location: Opera AB

This study investigates placement testing in Chinese language programs at Canadian universities, an area that has grown rapidly yet remains under-researched. While effective placement is crucial for language learning, teaching, and program administration, little empirical work has examined the validity of placement procedures in this specific context.

Guided by Kane's validity framework, which conceptualizes test validity as an argument-based evaluation across scoring, generalization, extrapolation, and implications, this study employs a sequential explanatory mixed methods design. Data were collected from institutional websites, program documents, social media forums, and semi-structured interviews with program directors, instructors and learners.

Findings show that most Chinese programs report placement mechanisms, typically combining written tests, oral interviews, and background questionnaires. This multi-source approach enhances validity by addressing multiple dimensions of learner ability. However, challenges persist, including the lack of test versions tailored to different learner profiles, concerns over rater reliability, limited alignment with international benchmarks, and questions about the validity of online testing.

The study contributes to both local and global discussions on language assessment. Conceptually, it applies Kane's framework to a less commonly taught language in a multilingual Canadian context. Methodologically, it illustrates the value of combining institutional, interview, and digital sources. Practically, it identifies areas for improvement in placement design and administration.

By highlighting these issues, the research not only informs Chinese language education in Canada but also offers broader implications for placement testing in other less commonly researched languages worldwide.

Building a Policy Literacy Survey

Authors: Laura Schildt (Ghent University), Bart Deygers (Ghent University)

Day: Friday, June 5 (5:15 p.m. - 5:45 p.m.)

Location: Symphonie 1

Policy refers to the frameworks that shape language assessment, including government policy, education policy, university regulations, and professional or industry standards. Policy literacy (Lo Bianco, 2001, 2021) is understood as being contextually prepared to engage with the interface between language testing and policy. As language tests are increasingly used as policy instruments that regulate educational and professional decision-making, language assessment professionals are expected to navigate complex policy environments. This raises the question of the forms of policy literacy required in the language assessment profession to meaningfully influence policy decisions.

This presentation addresses two questions: (1) what types of policy knowledge and skills are deemed necessary in the language assessment profession, and (2) how are these dimensions applied in practice. To investigate these questions, we developed and piloted a structured survey consisting of Likert-scale items and a ranking task. The instrument was administered internationally to educators, researchers, test developers, consultants, and other professionals working with language tests. The survey elicited views on how core dimensions of policy literacy should ideally be applied in the language testing field, how they are currently applied in practice, and how these dimensions are prioritized. Based on data from this large-scale administration, an exploratory factor analysis was conducted to identify underlying dimensions of policy literacy. The presentation reports the results of this analysis and introduces a validated instrument for measuring policy literacy in language assessment. The findings contribute to a clearer conceptualization of policy literacy and provide a tool for future research and professional capacity-building in the field.

Modeling Human-AI Collaboration Dynamics in EFL Writing: Hidden Markov Model and Process Mining Approach

Author: Shasha Xu (Zhejiang University of Finance & Economics)

Day: Friday, June 5 (5:15 p.m. - 5:45 p.m.)

Location: Symphonie 2

Current cognitive models (e.g., Hayes, 2012) depict writing as recursive planning, formulating, and reviewing, but were not designed to account for GenAI's affordances. Process-oriented studies are thus needed to reveal how GenAI reshapes writing dynamics and redistributes cognitive effort. This study analyzed 996 minutes of screen-recorded behavioral data from 18 Chinese EFL undergraduates engaged in an AI-assisted writing task. A total of 4,158 micro-level behaviors across 42 types were coded in ELAN. Using hidden Markov modeling and process mining, five latent states were identified: Content Pasting, Reading, Content Copying, Searching and Prompting, and Reviewing and Editing, each representing a distinct engagement mode with GenAI tools. Two recurrent loops emerged, one reflecting surface-level operations and the other deeper recursive engagement, along with three dominant patterns of sustained interaction, iterative prompting, and task-oriented regulation. Sequence clustering and process mining were conducted to further explore variation across learners. A two-cluster model best fit the data, revealing two regulatory patterns. Cluster 1 featured flexible strategic adaptation, where learners dynamically applied diverse prompting strategies and refined outputs locally. Cluster 2 reflected pragmatic global management, emphasizing holistic verification, integration, and global guidance of AI use. This study refines and extends previous frameworks through a data-driven five-state model among EFL undergraduates, highlighting the interplay between collaboration patterns and learner characteristics such as AI literacy and language proficiency. These insights have practical and methodological implications for the design, implementation, and research of human-AI collaborative writing.

Binary and CEFR Classification Consistency of the Duolingo English Test

Authors: Ramsey Lee Cardwell (Duolingo), Allison Michalowski (Duolingo)

Day: Friday, June 5 (5:15 p.m. - 5:45 p.m.)

Location: Symphonie 3

Admissions decisions translate global proficiency frameworks such as the CEFR into local score requirements. This study quantifies how consistently the Duolingo English Test (DET) supports those decisions at commonly used cut scores and across CEFR bands, with emphasis on the binary “meets/does not meet” choices institutions implement. Using operational data from August 2024–July 2025, we analyze 325,902 first-time test takers and 43,199 repeaters retesting within 30 days. We estimate classification consistency via (a) test–retest percent agreement and Cohen's kappa, (b) a population-weighted test–retest variant to mitigate self-selection, and (c) Livingston–Lewis single-administration estimates leveraging DET's overall reliability ($r=.96$).

At single cut scores aligned with B1 (60–95) and B2 (100–125), all methods yield $\geq .87$ agreement and kappas of .68–.83 (TRT .680–.755; TRT-Adj .781–.830; LL .708–.822). Percent agreement is highest at DET 60, while kappa peaks between 110–125, consistent with greater chance agreement at the distribution tails. For CEFR classifications, overall TRT kappa=.697 (weighted kappa=.776); 83.5% of repeaters remain in the same level and <0.1% move by more than one level. Conditional agreement is strongest at B2 (.904) and C1 (.870), with modest upward movement from B1→B2 on retest. Convergence between the weighted test–retest and Livingston–Lewis approaches suggests well-specified single-administration models are informative when retest data are sparse. Taken together, results indicate robust decision consistency that supports both CEFR-aligned interpretations and local score policies. Nevertheless, institutions should account for inevitable retest cut-score crossings when developing score requirement policies.

“Brevity is the soul of wit” but why is it untrue in writing assessment?

Authors: Bosco Zeng (University of Toronto), Angelie Ignacio (University of Toronto), Hajung Kim (University of Toronto)

Day: Friday, June 5 (5:15 p.m. - 5:45 p.m.)

Location: Symphonie 4

Automated essay scoring (AES) systems have advanced language assessment by providing efficiency, scalability, and consistency. However, their black-box nature raises concerns about interpretability, fairness, and construct validity. Addressing these issues, this study applies SHAP (SHapley Additive exPlanations) to audit BalanceAI’s AES model used in a formative assessment platform supporting students in Grades 3–8. Using 661 student essays (347 female, 314 male) scored on a four-point scale, the study examined whether linguistic features driving the model align with the intended writing construct and remain free from bias. Controlling for developmental stage (grade), learning differences (IEP/ESL), and cognitive ability (reading-ability quartiles), we extracted 33 linguistic features combining performance checks using SHAP with trained models to identify features most influencing score disparities.

Across all strata, gender gaps persist (Cohen’s $d \approx 0.33$ – 0.40), indicating differences not fully explained by controlled layers. SHAP attribution reveals that word count (24.68%) dominates one fourth of predictions. This indicates a structural length bias favoring verbose writing style, typically associated with female students. A scoring system that rewards length can advantage one subgroup while weakening construct alignment with writing quality.

This over-reliance on surface features such as length shows a broader tension between globally scalable proxies (length) and local pedagogies. Methodologically, we contribute a reproducible approach combining stratified fairness tests with feature-level explanations to disentangle educational variation from algorithmic bias. Practically, it proposes mitigations: normalizing length, shifting toward quality features, and human-AI teaming to promote more equitable, transparent, and pedagogically aligned language assessment.

Picture Prompt Design and Interactional Competence in Paired Speaking Assessment: Evidence from a Low-Stakes Online English Speaking Test in China

Author: Miao Gao (Beijing Normal University)

Day: Friday, June 5 (5:15 p.m. - 5:45 p.m.)

Location: Ovation

Task characteristics play a critical role in how speaking performance is elicited and interpreted. Yet, little attention has been given to the design of visual prompts in paired speaking tasks, despite their widespread use. This study investigates how the density of information in picture prompts influences interactional competence (IC) in the context of a Chinese online diagnostic oral assessment.

Three pairs of first-year English majors completed two versions of the same paired discussion task: one prompted by a black-and-white image with minimal details, and the other by a colored image enriched with multiple speech bubbles and visual cues. Conversation Analysis was employed to examine three dimensions of IC: turn-taking, topic management, and contingency on prior talk.

The findings highlight distinct contrasts between the two prompt formats. Sparse prompts led to rigid question–answer exchanges and fewer opportunities for elaboration, while richer prompts stimulated more frequent initiations, extended topical development, and greater balance between interlocutors. The multiple viewpoints embedded in the colored image further supported stepwise topic transitions and contingent responses.

The study also revealed that lower-proficiency candidates were able to demonstrate strong collaboration, while more proficient candidates sometimes produced less co-constructed interaction, underscoring that IC does not presuppose linguistic complexity. Personality traits also appeared to shape interactional performance.

By applying globally recognized IC frameworks to a local Chinese assessment context, this study illustrates how prompt design can bridge global constructs with local practices. It underscores the importance of task design in eliciting authentic interaction as well.

Paper Session 6.1 (Day 3: 9:40 a.m. - 10:10 a.m.)

Assessing metapragmatic awareness for intercultural communication: Test localisation targeting UK pre-sessional students

Author: Shishi Zhang (University College London)

Day: Saturday, June 6 (9:40 a.m. - 10:10 a.m.)

Location: Opera AB

Despite the past decades' efforts in assessing second language pragmatic competence, challenges remain, including assessing skills more transferrable mirroring test takers' real-life language needs, and designing a purpose-built assessment to operationalise such a construct. Addressing this research gap, this study proposes the assessment of metapragmatic awareness (Kecskes, 2014; McConachy, 2018). Within the field of applied linguistics, diverse interpretations of metapragmatic awareness exist, albeit using the same "metapragmatic" label. A few studies trying to assess metapragmatic awareness are essentially assessing elements more in the realm of pragmatic awareness without emphasising reflectivity and reflexivity of test takers during communication. Therefore, this study is the first to assess "meta"pragmatic awareness. Specifically, the study contextualises the assessment to UK pre-sessional students.

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

Test Taker Perceptions of Fairness, Validity, and Impact in a Local Standardized English Proficiency Test

Author: I-Chun Vera Hsiao (University of Iowa)

Day: Saturday, June 6 (9:40 a.m. - 10:10 a.m.)

Location: Symphonie 1

While English proficiency tests such as TOEFL, IELTS, and the Duolingo English Test have been extensively studied globally, locally developed assessments remain underexamined, particularly in non-English-dominant academic contexts. This mixed-methods study addresses this gap by investigating test takers' perceptions of the BEST Test of English Proficiency (BESTEP), Taiwan's first English for Academic Purposes (EAP) assessment, introduced under the Bilingual 2030 policy. Specifically, the study examined perceptions of BESTEP's fairness, alignment with students' academic and linguistic needs, and its broader impact on test takers' learning experiences and future opportunities.

The study extends Green's (2014, 2020) Framework for Effective Assessment Systems by proposing a Quality of Standardized Language Tests model, which conceptualizes test quality as a circular, interconnected system comprising Fairness, Validity, Impact, Reliability, and Practicality. Fairness is elevated to the same conceptual level as validity, drawing on Wallace's (2018) Language Assessment Fairness Model, while validity is conceptualized in accordance with the Standards for Educational and Psychological Testing (AERA et al., 2014). Impact is conceptualized through Shih's (2007) Washback Model and related scholarship (Alderson & Wall, 1993; Bachman & Palmer, 1996), involving both intended and unintended consequences at individual and systemic levels.

By centering test takers' perspectives, this study contributes to a more holistic understanding of locally developed high-stakes assessments and provides a validated instrument suitable for large-scale standardized testing. The findings underscore the importance of fairness, contextual relevance, and stakeholder perspectives in test design, offering guidance for more equitable and context-sensitive assessment practices and supporting the implementation of Taiwan's English education policies. While focused on BESTEP, the results have broader implications for language assessment and education policy across the Asia-Pacific region.

Attention Allocation in L2 Summary Writing: Insights from Eye-Movement Data

Authors: Lin Shi (Zhejiang University), Yang Zhang (Zhejiang University), Lianzhen He (Zhejiang University)

Day: Saturday, June 6 (9:40 a.m. - 10:10 a.m.)

Location: Symphonie 2

Summary writing has gained prominence in language testing due to its inclusion in large-scale, high-stakes assessments, yet the cognitive mechanisms underlying this task remain underexplored. Drawing on attentional resource theory (Kahneman, 1973), this study investigates how Chinese EFL learners allocate cognitive resources between comprehending source materials and producing text. As traditional methods like think-alouds may intrude upon processing and undermine authenticity, eye-tracking was employed to capture attention allocation patterns and investigate how these patterns relate to L2 summary writing performance.

Forty Chinese EFL learners (B1–C1) completed an argumentative summary-writing task on a computer interface with three areas: the source text, a note-taking space, and a writing space, while their eye movements were recorded using a Tobii TX300 tracker. Writing performance was assessed by two experienced raters using a standardized rubric. Areas of Interest (AOIs) were defined in two schemes: (1) the three task areas to examine attention–score associations, and (2) Toulmin units to investigate how attention to argumentative components relates to summary quality. Eye-movement measures were computed for each AOI. Finally, retrospective stimulated-recall interviews were analyzed using grounded theory.

Results indicated that attention allocation significantly influenced L2 summary writing performance. High-quality summaries emerged when learners balanced attention between reading and writing and engaged more deeply with the source text. Eye-tracking and recall data further revealed proficiency-based differences in how learners distributed attention across argumentative components. These findings elucidate cognitive mechanisms, offer pedagogical insights, and highlight the potential of eye-tracking to enhance task authenticity and refine writing assessment constructs.

Global to Local: Examining the Predictive Validity of the DET Interactive Listening Tasks

Authors: Noriko Iwashita¹, Hong Giang Nguyen², Anh Nguyen¹, Van-Trao Nguyen², Trung Quang Doan², Hanh Hong Kieu²

Organizations: 1: The University of Queensland, 2: Hanoi University

Day: Saturday, June 6 (9:40 a.m. - 10:10 a.m.)

Location: Symphonie 3

The study examined the validity of the DET Interactive Listening Tasks as an assessment of Interactional Competence. Data were collected from 125 participants enrolled in two English courses at a large university in Vietnam. Participants completed the DET, one of the course speaking tests, and a practice version of the VSTEP (Vietnamese Standardised Test of English Proficiency). Both course tests and the VSTEP Speaking test have a substantial component of IC measures. Using FACETS (Linacre, 2014), we analysed participants' performances on the course-based tests and the VSTEP Speaking Test through multi-faceted Rasch measurement. The results showed small to moderate levels of significant correlation between DET sub-scores and both the course tests and the VSTEP speaking task scores. For the analysis conducted separately for C1 and B2 levels, we found a stronger significant correlation with the course test and the VSTEP speaking task for the higher level (C1) than the lower level (B2). Not all DET IL-related sub-scores were significantly correlated with B2 course test and VSTEP speaking task (B1). The conversation analysis of a small portion of the course tests and VSTEP test performance revealed some interactionally relevant features were influenced by speaking proficiency, while others were independent. The findings support the use of the IL task as an IC assessment in a CMT. Finally, the study underscores the importance of aligning global computer-mediated assessments with local, performance-based frameworks to ensure fairness and contextual validity.

Does the Rating Make the Score? Comparing how Human Ratings affect Automated Essay Scoring in a Large-Scale Dutch Writing Assessment

Authors: Joni Kruijsbergen¹, Marieke Vanbuel^{1,2}, Amaury Van Parys¹, Bart Deygers¹, Orphée De Clercq¹

Organizations: 1: Ghent University, 2: KU Leuven

Day: Saturday, June 6 (9:40 a.m. - 10:10 a.m.)

Location: Symphonie 4

Automated essay scoring (AES) is increasingly used to complement or replace human raters in large-scale testing, yet little is known about how different rating modes align as input for such systems, particularly in languages other than English. This study examines the relationship between analytic human scoring, comparative pairwise judgment, and AES in the context of large-scale Dutch writing assessments for students in the second year of secondary education (Grade 8). A total of 1,776 persuasive essays were rated under two human rating conditions: (a) analytic scoring on criteria such as structure, grammar, and vocabulary, and (b) comparative judgment based on pairwise ranking of overall quality. The two scoring methods were compared for reliability, consistency, and construct overlap. Results show a strong correlation between analytic and comparative scores (Spearman $\rho = .77$, $p < .001$), though slight discrepancies may suggest differences in construct emphasis. Feature-based AES pass-fail models were then trained separately on both rating types using linguistic indicators extracted with T-Scan, covering lexical sophistication, syntactic complexity, cohesion, and fluency. Models based on both human rating types achieved robust predictive accuracy (Quadratic Weighted Kappa $> .75$). However, certain higher-level dimensions (e.g., communicative adequacy, register, and content) were not fully captured by linguistic features alone. By integrating analytic and comparative human judgments into automated scoring, the study aims to show how AES can be adapted and validated for a specific linguistic and educational context, and how locally developed systems can operate within an English-dominant technological ecosystem.

Invisible Assessment & Transparent Evidence with LACES

Author: Olivia Seow (Harvard University)

Day: Saturday, June 6 (9:40 a.m. - 10:10 a.m.)

Location: Ovation

Language assessment seeks to balance psychometric rigor with pedagogical value, yet capturing authentic performance while providing actionable feedback remains challenging. LACES (Language Assessment via Conversational Exchange with Scaffolds) addresses this challenge by embedding multi-perspective observation within natural conversation. Every assessment claim links to specific conversational moments, transforming evaluation into personalized coaching.

LACES deploys specialized large language model observers to capture complementary phenomena, building holistic competence profiles and personalized learning paths. The modular architecture adapts to local pedagogical priorities, with observers reconfigurable for specific contexts. We demonstrate LACES through LingoLand (Seow, 2023), an immersive quest-based game where authentic interactions enable formative assessment without disrupting communicative flow. Currently, LACES analyzes text and transcribed conversations, with direct audio processing capabilities in development. The system uses Bayesian skill tracking with conservative priors to maintain calibrated confidence estimates, triggering interventions only when expected benefit exceeds interruption cost. Validation studies with intermediate learners demonstrated strong assessment value while reducing communication anxiety and increasing metacognitive awareness.

LACES adapts to local contexts. For example, in Québec scenarios, it recognizes code-switching between French and English as translanguaging competence rather than error. This configurability extends to regional varieties, professional registers, and evolving linguistic norms.

By making assessment invisible, LACES enables authentic performance. By making evidence transparent, it transforms evaluation into personalized learning. While validated in a game context, the modular architecture of LACES readily adapts to classrooms, professional training, or online exchanges, offering assessment support through meaningful language use.

Paper Session 6.2 (Day 3: 10:15 a.m. - 10:45 a.m.)

A mixed methods study of rater judgements of pronunciation in a paired discussion task

Author: Liberato Silva dos Santos (Iowa State University and Instituto Federal de Goias)

Day: Saturday, June 6 (10:15 a.m. - 10:45 a.m.)

Location: Opera AB

Pronunciation is crucial to oral communication but remains poorly conceptualized and operationalized for assessment purposes (Choi & Kang, 2023; Isaacs & Trofimovich, 2017), especially with more communicatively authentic and interactive tasks (Galaczi & French, 2011; Isaacs, 2018). This study addressed this gap in an EFL context in Brazil. The goal was to determine to what extent pronunciation could be effectively measured with an L2 paired discussion task – PDT (Davis 2009, Ockey et al., 2021). Rater judgements of pronunciation ability were examined using an operational construct developed for this study. The study investigated (1) how raters judged pronunciation ability in a PDT, (2) what effect examinee pairings based on pronunciation ability had on rater judgements, and (3) what were the raters' perceptions of the construct, rating scale, and task. A mixed-methods study (Creswell & Plano Clark, 2018) examined Brazilian raters (N = 9) judging adult Brazilian examinees (N = 78) using a 4-point pronunciation rating scale with five features: rhythm, intonation, word stress, vowels, and consonants. Examinee ability, rater severity, and scale functioning were analyzed using Rasch measurement. Rater behavior while scoring same versus different pronunciation ability examinee pairs was examined using ANOVAs. Rater perceptions of construct, task, and scale were analyzed using content analysis. Results indicated raters effectively used the construct and rating scale to judge pronunciation ability. No statistically significant effects of pairings on ratings were found. Implications include contributions to understanding rater behavior, construct and rating scale development, and could support test developers, rater trainers, and language educators.

Prompt design based on communication strategies to facilitate spoken interaction between generative AI and EFL learners, and validation of its effectiveness

Author: Yutaka Yamauchi (Soka University)

Day: Saturday, June 6 (10:15 a.m. - 10:45 a.m.)

Location: Symphonie 1

With the rapid advancement of generative AI, language learners can now engage in interactive speaking practice through sophisticated voice-based conversational systems. However, previous studies have reported that such interactions often lack balance, as generative AI tends to dominate the dialogue while learners produce limited output. Despite this issue, few studies have examined how different AI prompt designs can foster more effective and balanced conversational practice across varying proficiency levels.

Grounded in communication strategy theory, this study developed an innovative AI-based speaking practice and evaluation system. The system incorporates two distinct prompting styles for generative AI. Prompt A, drawing on Abe (2023) and Seong (2014), encourages topic expansion through elicitation, clarification, and exemplification strategies (e.g., What do you think?, Can you tell me more?, Can you give me an example?).

In contrast, Prompt B, informed by Dietl et al. (2023) and Ting et al. (2017), prioritizes psychological support over topic development, employing strategies to reduce learner anxiety and linguistic load (e.g., Take your time., You don't have to say a full sentence., You can say it in Japanese first.).

These results suggest that in AI-assisted speaking practice, topic-expanding prompts enhance engagement among advanced learners, while psychologically supportive prompts foster greater speech production among lower-proficiency learners. The study underscores the importance of tailoring AI prompt design to learner proficiency to promote smoother, more effective conversational interaction.

Effects of task type on the variability pattern of syntactic complexity in second language writing

Author: Minh Nhut Nguyen (Macquarie University)

Day: Saturday, June 6 (10:15 a.m. - 10:45 a.m.)

Location: Symphonie 2

Research on second language (L2) writing has increasingly examined how task types shape learners' performance, particularly in terms of syntactic complexity. Most previous studies reported that argumentative texts elicit more complex language than descriptive or narrative texts, yet the variability patterns among specific complexity metrics remain inconsistent. In addition, limited attention has been paid to how descriptive and narrative tasks compare, despite their relevance in assessment contexts.

This study draws on Robinson's Cognition Hypothesis, which proposes that tasks with greater cognitive demand lead to greater linguistic complexity. It investigates whether syntactic complexity varies systematically across three task types—descriptive, narrative, and argumentative—produced under the same conditions. Forty-eight L2 English learners studying at Australian institutions completed two of the three task types, generating 96 texts. Fourteen syntactic complexity metrics were analysed.

The findings show that argumentative tasks generally prompted higher syntactic complexity than both descriptive and narrative tasks. Meanwhile, the descriptive–narrative comparisons revealed mixed patterns: narrative tasks outperformed descriptive tasks on seven metrics, while the remaining seven metrics showed the opposite variability pattern. Interestingly, only two metrics proved sensitive in distinguishing the written texts across three different task types.

By identifying which complexity metrics best capture systematic variability, this study offers new insights into how the findings can inform the assessment and testing of L2 writing performance. Moreover, greater emphasis should be placed on incorporating a balanced mix of task types, including underexplored ones, to better understand how complexity metrics reflect variability in learner writing.

Global Standards, Local Impact: Educational and Social Consequences of EXAVER, Mexico's English Certification

Author: Marion Alain Meunier Colula (Universidad Veracruzana)

Day: Saturday, June 6 (10:15 a.m. - 10:45 a.m.)

Location: Symphonie 3

This paper examines the educational and social consequences of the Certificación de Lengua Inglesa EXAVER, the only English language proficiency exam fully designed, validated, and administered in Mexico, and officially recognized by the Ministry of Education. While global frameworks such as the Common European Framework of Reference for Languages (CEFR) shape language assessment worldwide, little is known about how locally developed certification systems function within their own contexts. This study addresses that gap by investigating how EXAVER balances international standards with local needs and the impacts it generates for students, teachers, and institutions.

Drawing on leading theories of validation (Messick, 1989; Kane, 2013; Weir, 2005) and Shohamy's (2001) critical perspective on the power of language tests, the study applies an expanded model of consequential validity across eight dimensions, including fairness, equity, usability, and policy alignment. A mixed-methods design combined data from 492 test-taker surveys with a focus group of eight English teachers, whose perspectives provided insights into teaching practices and institutional policies.

Findings show that EXAVER is perceived as fair, accessible, and pedagogically valuable, supporting program accreditation, student mobility, and curriculum design. Its affordability compared to international certifications emerged as a key factor for equity and inclusion. Challenges include limited recognition outside Veracruz and cases of negative washback in classrooms.

By situating a local certification system within global debates, this paper demonstrates how context-sensitive assessment can promote fairness, social justice, and sustainable language policy, offering lessons relevant to Mexico and other regions worldwide.

A Language-for-Specific-Purposes Approach to University Admission: Meeting Local Needs through Needs Analysis and Test Innovation

Authors: Ashlee Farnsworth (Brigham Young University), Shantu Mir (Brigham Young University), Troy L. Cox (Brigham Young University)

Day: Saturday, June 6 (10:15 a.m. - 10:45 a.m.)

Location: Symphonie 4

Global English proficiency tests such as TOEFL and IELTS have long shaped university admissions but often excluded or disadvantaged applicants due to cost, logistics, and documentation requirements. Moreover, these global assessments frequently overlook the specific linguistic and academic practices required within individual institutions. This study examined how a private university in the western United States could design a context-specific English admissions test grounded in Language-for-Specific-Purposes (LSP) principles through a detailed needs and content analysis.

Researchers first identified the twenty most common first-year courses, spanning both general education and introductory major classes in the sciences and humanities. Syllabi and course materials were analyzed to identify the reading and writing tasks students were expected to perform, revealing a range of genres such as analytical essays, lab reports, and research syntheses. Faculty interviews provided further insight into the linguistic abilities and rhetorical competencies most critical for student success, while focus groups with international students highlighted the real-world challenges multilingual learners faced when completing these tasks.

Together, these data sources described the institution's target language use (TLU) domain, demonstrating that a locally designed assessment can more accurately represent the authentic communicative demands of the university. The alignment between test tasks and real academic activities suggests increased validity and potential for positive washback, as preparation for the test parallels preparation for coursework. This study offers a replicable model for institutions seeking to develop local assessments that are equitable, authentic, and contextually grounded.

Response Patterns of Native and Non-Native English Speakers on PIAAC Literacy Items: A Latent Profile Analysis

Authors: Hongli Li (Georgia State University), Boshi Wang (Georgia State University), Natalie Pruitt (Georgia State University)

Day: Saturday, June 6 (10:15 a.m. - 10:45 a.m.)

Location: Ovation

The Programme for the International Assessment of Adult Competencies (PIAAC) literacy assessment measures how well adults can access, understand, evaluate, and reflect on written texts to accomplish goals, participate in society, and develop knowledge (OECD, 2021). During its Cycle 2 data collection, 3,765 U.S. adults participated. Of these, 458 reported a native language other than English (i.e., non-native English speakers), and 3,273 reported English as their native language (i.e., native English speakers). Non-native language speakers are not only lagging in their proficiency level but also show different response processes. Prior research suggests that non-native speakers show lower engagement (Goldhammer et al., 2017) and more rapid guessing (Rios & Soland, 2022). In this study, drawing on the U.S. PIAAC Cycle 2 literacy assessment data, we examine response-process patterns using latent profile analysis (LPA) and then test whether profile membership differs for native-language vs. non-native language test takers. We fit a series of LPA models with full maximum-likelihood estimation in Mplus. Model selection considered AIC, BIC, entropy, and likelihood-ratio tests, balancing model fit and interpretability. A three-profile solution provides the best combination of model fit and interpretability. Findings also suggest that language status relates to profile membership, with non-native speakers more often showing high-effort/less-efficient behaviors. These patterns can inform more targeted instructional and intervention programs for non-native English-speaking adult learners.

Paper Session 6.3 (Day 3: 10:50 a.m. - 11:20 a.m.)

Human Judgment or Algorithmic Ease? Score Comparability and Student Perceptions in Adolescent EFL Writing Assessment

Author: Chia-Hsin Yin (The Ohio State University)

Day: Saturday, June 6 (10:50 a.m. - 11:20 a.m.)

Location: Opera AB

Generative AI (GenAI) is reshaping second language writing (SLW) assessment, yet few studies have examined how its scoring compares to that of human teachers in real classroom contexts. This explanatory sequential mixed-methods study investigates the comparability of Microsoft 365 Copilot and teacher ratings in Taiwanese high school English as a Foreign Language (EFL) classrooms, and how students interpret the feedback they receive from each.

A total of 278 Grade 12 essays, spanning narrative and argumentative genres, were independently rated using the Taiwan General Scholastic Ability Test (GSAT) English Composition Rubric, which includes four analytic subscales and a holistic score. Two-way (rater $\times$ genre) and within-genre analyses revealed systematic differences: Copilot consistently assigned higher scores for content, organization, and holistic performance, while aligning more closely with teachers on grammar and vocabulary. These patterns suggest GenAI may be more consistent in rating lower-order linguistic features than higher-order rhetorical elements.

To explore student perspectives, semi-structured interviews were conducted. Students found Copilot helpful for surface-level editing, but valued teacher feedback for its specificity, nuance, and motivational tone, especially for revision. They also raised concerns about AI's impersonality and limited contextual awareness.

Findings underscore GenAI's potential as a complementary tool in formative writing assessment while reaffirming the irreplaceable role of teachers in providing pedagogically meaningful and contextually grounded feedback. This study contributes to the LTRC 2026 theme by advocating for ethically responsible, context-sensitive AI integration in adolescent EFL writing assessment.

Supporting the Validity Argument of a High-Stakes Language Assessment Through Evidence of Test-Taker Identity Verification

Authors: Ramsey Lee Cardwell¹, William Belzak¹, Manqian Liao¹, Geoffrey T. LaFlair¹

Organization: 1: Duolingo

Day: Saturday, June 6 (10:50 a.m. - 11:20 a.m.)

Location: Symphonie 1

Test score interpretations are defensible only when performance is attributable to the intended examinee. We make that inference explicit via the Attribution inference. Attribution includes three warrants: no impersonation, no item pre-knowledge, and no external resources. This session focuses on "no impersonation" and its backing, including cases where helpers feed answers. In international high-stakes testing, impersonation risk intersects with local realities (ID formats, infrastructure, fraud markets). Using an English language admissions test accepted worldwide, we show how security must adapt to local constraints while preserving equity and access. Conceptually, we pair an argument-based validity structure with tangible evidence across people, processes, and data, and we offer a portable method for specifying warrants and communicating evidence so stakeholders can judge suitability.

We outline a multi-layer framework that underwrites the "no impersonation" warrant through three complementary categories. First, prevention and detection jointly block attempts and surface anomalies: locked-down software, multi-factor ID checks, dual-camera recording, plus AI signals for keystrokes, collusion, gaze, and copy-typing. Second, human-in-the-loop review adjudicates flags through asynchronous, multi-round, partially anonymized proctoring, where trained proctors contextualize alerts and make determinations. Third, procedures and evidence are made transparent through technical documentation. We report measure-specific indicators (e.g., alert precision/recall) and show how proctor training and weekly feedback sustain high agreement while minimizing false positives. Aligning concrete procedures and results to an explicit Attribution inference provides a replicable template for making security warrants visible, balancing integrity with access, and evaluating test security in high-stakes contexts such as international admissions.

Does Input Modality Affect L2 Integrated Writing Consistently? Examining the Moderating Effect of English Proficiency, Modality Preference, and Visual-spatial Working Memory

Author: Jiayi Dong (Zhejiang University)

Day: Saturday, June 6 (10:50 a.m. - 11:20 a.m.)

Location: Symphonie 2

In real-world academic contexts, L2 learners interact with abundant multimodal resources, where visual information plays a pivotal role in emotion transmission and meaning construction. While extensive research has examined visual input's impact on L2 listening comprehension, its role in L2 writing remains understudied. Existing studies mainly verify whether input modality affects writing performance, with little attention to how test-takers engage with visual information or how personal traits moderate such modality effects. Building on prior findings that visual scaffolds enhance L2 writing, this study explored how visual information interacts with three key learner traits (English proficiency, modality preference, and visual-spatial working memory (VWM)) to influence L2 integrated writing.

51 Chinese college L2 English learners participated in a counterbalanced within-subjects design. They first completed an English proficiency test, a Corsi task (i.e. a measure to assess VWM capacity), and the VARK questionnaire (i.e. a measure to identify modality preference), then two TOEFL iBT-adapted writing tasks: a reading-listening-writing task (with an audio-only lecture) vs. a reading-viewing-writing task (with a video-based lecture), rigorously matched except for input modality. Correlation analyses revealed positive links between input modality, proficiency, and writing performance, but no significant correlations for modality preference or VWM. Mixed-effects regression showed modality preference and VWM significantly moderated the modality effect: video input improved performance more for learners with higher VWM or visual preference. Proficiency did not interact with input modality. These findings deepen understanding the interplay of traits and modality in L2 writing, informing multimodal task design and cultivation of learners' multimodal adaptability.

From Global Standards to Local Realities: Rethinking Vocabulary Testing for Chinese and Arabic Learners

Authors: Geoffrey Pinchbeck (Carleton University), Mastoor Al Kaboody (Carleton University)

Day: Saturday, June 6 (10:50 a.m. - 11:20 a.m.)

Location: Symphonie 3

Vocabulary assessment used in research and pedagogy have long relied on corpus-frequency-based sampling approaches, but are not the most accurate predictors of learner word knowledge (Hashimoto & Egbert, 2019). In contrast, knowledge-based word lists (KBLs), compiled from extensive vocabulary testing in specific L1 English-learning contexts, have been proposed as superior to frequency lists when used as resources in text-readability models and when sampling for diagnostic vocabulary tests (Schmitt et al., 2021). This study investigates whether L1-Arabic and L1-Chinese learners of English exhibit sufficiently distinct vocabulary knowledge profiles to justify the development of separate KBLs. Data were collected from 530 English for Academic Purposes learners (including 425 L1-Chinese and 105 L1-Arabic speakers) at a Canadian university using a matrix-sampled 187-item multiple-choice vocabulary test. Rasch modeling estimated item difficulty and learner ability, while Differential Item Functioning (DIF) analysis identified item-level differences between the two L1 groups. 29 items displayed statistically significant and large effect size DIF (13 favoring Arabic, 16 favoring Chinese learners), revealing group-specific lexical strengths. In a post-hoc qualitative follow-up, eight expert informants (MA, Applied Linguistics), four representing each L1-group, offered explanations that converged within each L1 group of experts. DIF patterns were inferred to have stemmed from L1-English loanwords (Arabic), differences in culturally embedded exposure (Arabic), and test-preparation practices (Chinese). These findings underscore the limitations of a single lexical standard for diverse learners and highlight the potential of L1-sensitive KBLs to enhance vocabulary assessment fairness and instructional targeting. Implications for test design, pedagogy, and cross-linguistic lexical research are discussed.

Assessing a Localized Construct of Writing Ability: A Validity Argument for an NGSS-Aligned Scenario-Based Language Assessment

Author: Soo Hyoung Joo (Teachers College, Columbia University)

Day: Saturday, June 6 (10:50 a.m. - 11:20 a.m.)

Location: Symphonie 4

In Language Assessment for Specific Purposes (LASP), longstanding debates on how to distinguish specific target language use (TLU) domains from general proficiency underscore that communicative practices often overlap across contexts (O'Sullivan, 2012; Knoch & Macqueen, 2019). This recognition challenges views of language ability as a decontextualized skill set and instead frames it as contextually realized performance integrating linguistic, cognitive, and situational resources (Douglas, 2000; Hamp-Lyons & Lumley, 2001). Purpura (2021, 2024) extends this perspective by calling for the assessment of situated S/FL proficiency, best elicited through Scenario-Based Language Assessment (SBLA).

Responding to this call, this study builds a validity argument for a Scenario-Based Language Assessment for Science Purposes (SBLA-SP) designed to measure situated specific purpose writing ability in STEM contexts among secondary EFL learners in South Korea (N = 176). Localization was achieved by aligning the assessment with both the Next Generation Science Standards (NGSS) and the Korean national science standards, ensuring that task design reflected competencies essential for disciplinary communication.

Guided by the Argument-Based Validation (ABV) framework (Kane, 2013; Chapelle, 2020), evidence was gathered for three inferences: domain definition (alignment analysis confirming correspondence between task features and disciplinary competencies), evaluation (many-facet Rasch measurement showing construct-relevant scoring and minimal rater bias), and generalization (multifaceted generalizability theory indicating acceptable dependability). Findings suggest that the NGSS-aligned SBLA-SP operationalizes a localized construct of writing ability, demonstrating how localized validity arguments can bridge global constructs of language proficiency with contextually grounded disciplinary practices.

Yup'ik Students' Voices: Alaska Native Language (Yugtun) assessment

Authors: Catherine Moses (Lower Kuskokwim School District), Rosalie Lincoln (Lower Kuskokwim School District), Sally Samson (University of Alaska - Fairbanks), Rosalie Grant (Australian National University)

Day: Saturday, June 6 (10:50 a.m. - 11:20 a.m.)

Location: Ovation

The theme of the LTRC 2026 addresses the importance of considering the sociocultural and sociolinguistic contexts within which language assessment activities are conducted. This is particularly critical for Indigenous language test development.

This paper discusses feedback from grades 7-12 students regarding items developed for an Alaska Native language (Yugtun) assessment, the *Yugtun Piciryaraneq Qaneryaraneq-Ilu Cuqyun* (aka *Yup'ik Culture and Language Measurement*). The project is based on enacting Community-Based Participatory Research principles. As part of that development, items were piloted in six village schools in the Lower Kuskokwim School District, Central Alaska, U.S.A.

Students provided feedback on items for Oracy, Literacy and Cultural Understanding – 175 responses were received in the first pilot and 140 in the second pilot. The students' views provided indicative evidence of achieving two project aims, namely, that: (a) the assessment be embedded in Yup'ik culture and language; and (b) Yup'ik students understand the importance of who they are in the world and that their language and culture are to be embraced and respected.

The feedback shows that many students enjoyed doing the tests even when some parts were challenging. They valued the opportunity to learn Yugtun; to become more culturally aware; and to play their role in preserving their culture, language and Yup'ik identity. The feedback suggests that the tests themselves kindled Yup'ik students' interest and pride in their language (Yugtun) and in Yup'ik culture. In other words, the tests themselves had a positive outcome on students.

Paper Session 7.1 (Day 3: 12:30 p.m. - 1:00 p.m.)

A Framework for Validating Hybrid-Marking Models of L2 Speech & Writing

Authors: Mark Brenchley (Cambridge University Press & Assessment), Trevor Breakspear (Cambridge University Press & Assessment), Yan Huang (Cambridge University Press & Assessment)

Day: Saturday, June 6 (12:30 p.m. - 1:00 p.m.)

Location: Opera AB

As AI rapidly transforms the assessment of constructed responses, it is important to design approaches that maximise the complementary strengths of AI and human judgment. To this end, we leverage a hybrid-marking approach based on confidence models. These are ancillary models, trained to provide a quantitative estimate of a particular automarker's scoring accuracy. Using this estimate, we set a threshold that determines whether a particular automarker score is accurate enough to send for results determination or whether the response should instead be sent for human marking.

In this presentation, we detail a validation framework for evaluating this approach to hybrid-marking, illustrating it with a case study that draws on a large-scale, multi-mark dataset of operationally-scored responses from a high-stakes, second language writing exam. We show how this framework enables better deployment decisions in two ways. Firstly, it enables better selection among available hybrid-marking models, achieving a combined model with lower score variation and that more notably outperformed operational examiners across a wider range of thresholds. Secondly, we show how the approach enables us to incorporate information about the testing context that helps identify which model is more appropriate for that context; for example, a hybrid-marking model that achieves better performance for tests of particular CEFR levels. Consequently, test providers can more effectively combine the consistency, reliability and efficiency of automated scoring with the targeted quality assurance of human judgement.

Local writing placement tests and generative AI assistance: A domain analysis of an evolving academic construct

Authors: Rebecca Yeager¹, Rurik Tywoniw¹, Ha Ram Kim², Melissa Meisterheim³, Shireen Baghestani⁴, Jim Ranalli⁴, Sara Cushing⁵

Organizations: 1: University of Illinois Urbana-Champaign, 2: University of California-Irvine, 3: University of Iowa, 4: Iowa State University, 5: Georgia State University

Day: Saturday, June 6 (12:30 p.m. - 1:00 p.m.)

Location: Symphonie 1

Local domain analysis is critical to the development of valid academic placement tests. In the wake of generative AI, we expand upon Knoch and Macqueen's (2020) domain analysis framework, which explores task, language, processing, criteria, and standards in a target language use (TLU) domain, by introducing the concept of domain resource analysis to understand the tools that test-takers may access in the TLU domain. We build on groundwork by Seguis and Lim (2020), who justified changes to writing assessment standards for medical workers on the basis of technological changes in the workplace. Our study seeks to apply this framework to understand the technological resources available to student writers in university contexts.

We report on a Fall 2024 survey across three U.S. universities, asking faculty to describe their teaching context and AI expectations for writing assignments using the AI Assessment Scale (AIAS; Perkins et al., 2023). These data, along with follow-up interviews and a mini round of surveys in Spring 2025, were analyzed descriptively and qualitatively to observe shifting expectations over time.

Responses indicate significant heterogeneity in AI expectations across institutions, disciplines, and class sizes, as instructors balance innovation and disciplinary standards. This heterogeneity largely persists across semesters. Our findings caution that international student writers cannot currently assume access to generative AI for all academic writing tasks, warranting continued assessment of unassisted writing ability on language placement tests. However, ongoing longitudinal, discipline-specific, and locally-contextualized domain resource analysis is necessary to monitor the evolving landscape of academic writing.

Test-taker agency in writing assessment: Balancing universal proficiency constructs with context-sensitive design through topic choice

Authors: Hye-won Lee (Cambridge University Press & Assessment), Carla Pastorino-Campos (Cambridge University Press & Assessment)

Day: Saturday, June 6 (12:30 p.m. - 1:00 p.m.)

Location: Symphonie 2

In writing assessment, even minor design features can shape test-takers' experiences of agency, fairness, and validity. Recent scholarship highlights learners as active stakeholders in assessment design, raising questions about how large-scale systems can embed learner agency without compromising comparability. Topic choice exemplifies this tension: while universal proficiency models prioritise standardisation, local systems should address situated learner needs.

Drawing on sociocognitive (Bandura, 2001) and ecological (van Lier, 2008) perspectives, this study conceptualises agency as both an internal capacity (intentionality, self-regulation, self-efficacy) and an emergent relation to environmental affordances. Within this framing, topic familiarity and personal relevance may introduce construct-irrelevant variance (CIV), whereas constrained choice could mitigate such effects by engaging test-takers with meaningful prompts.

An experiment introduced constrained topic choice into a two-task writing test typically offering none. Adult English learners ($n = 321$) completed two tasks with a choice of two topics each, followed by a questionnaire on anxiety (modified STAI; Marteau & Bekker, 1992) and engagement (Henrie et al., 2015). Writing tasks were scored by trained examiners, and behavioural indicators of engagement (e.g., text length, sentence complexity) were analysed alongside self-reports using linear mixed-effects models.

Preliminary results show that topic choice did not affect scores but influenced affective and behavioural engagement, with participant characteristics moderating these relationships. The findings advance validity arguments by examining how constrained choice can support meaningful engagement while maintaining score comparability, illustrating how writing assessments can balance fairness, agency, and standardisation in large-scale proficiency frameworks.

Can you prototype in the field?: Revisiting the tripartite model of preoperational testing

Authors: Yena Park (Duolingo), Xiaowan Zhang (Duolingo)

Day: Saturday, June 6 (12:30 p.m. - 1:00 p.m.)

Location: Symphonie 3

The prototyping stage in the tripartite model of preoperational testing has mostly been limited to small-scale process-based qualitative data (Fulcher, 2010). Data collected at the field testing stage is not typically available at the prototyping stage, limiting the types and range of evidence that can be used to support decisions at the prototyping stage. We challenge this somewhat disparate nature of the three-stage preoperational testing in practice and introduce the idea of prototyping in the field, enabled by an experimentation platform for rapid iteration and deployment.

We present one case study that illustrates prototyping in the field, where data collected encompassed all three stages. The case study involved a series of experiments to introduce design interface changes on the test without impacting test-taker performance or behaviors, allowing the test to continue making the same interpretative claims.

We find that clear distinctions between prototyping, piloting, and field testing are not tenable on a digital-first test with a platform that allows for prototyping directly in the field. We argue that evidence gathered in each stage is not limited to that stage – item and test statistics, typically reserved for the piloting and field testing stages, can be obtained and used to make decisions at the prototyping stage, allowing for a truly iterative, cyclical, nonlinear, and dynamic loop of preoperational testing (Fulcher, 2010). Types of data collected for prototyping can be expanded, beyond small-scale interpretation-based qualitative data to quantitative measures like timestamps and keystroke data, resulting in stronger and richer evidence for validity arguments

Bridging Policy and Practice: A multi-method evaluation of a highstakes language test's fitness-for-purpose in the Canadian migration context

Authors: Ellen Barrow (Pearson), Sarah Hughes (Pearson)

Day: Saturday, June 6 (12:30 p.m. - 1:00 p.m.)

Location: Symphonie 4

This presentation reports on two complementary validity research projects for a computer-based English language proficiency test designed to support Canadian migration. Prior to its launch in 2024, extensive validation activities established the test's alignment with the Canadian Language Benchmarks (CLB), confirming its suitability for high-stakes decisions by Immigration, Refugees, and Citizenship Canada (IRCC). However, in recognising that language assessments must remain relevant as contexts and populations evolve, ongoing validation activities are essential.

A post-launch standards verification study engaged expert Canadian judges to review extended Speaking and Writing responses from a representative sample of test takers. The study found over 90% agreement between CLB classifications based on test scores and those based on expert judgement, with qualitative analysis confirming that characteristics of test taker performances closely reflected CLB standards. To complement this technical evidence, a mixed-methods study explored test takers' perceptions and experiences, combining survey data (N=280) with qualitative insights from interviews (N=9) and focus groups (N=10). Findings highlighted that test takers viewed the assessment as authentic and relevant to their daily lives, reinforcing test alignment with the CLB and its appropriateness for immigration and citizenship pathways.

Overall, while policy alignment is a critical foundation for test validity, this research demonstrates that understanding the real-world communicative needs and lived experiences of test takers is equally vital. By integrating empirical alignment to the CLB with insights from test takers themselves, we establish a more holistic understanding of test validation, grounded in both policy and practice.

Exploring the potential of AI-generated reformulation and model texts in a resourceconstrained assessment context: An eye tracking and stimulated recall study

Authors: Atta Gebril (The American University in Cairo), Khaled ElEbyary (University of York), Ramy Shabara (University of Hertfordshire, hosted by Global Academic Foundation)

Day: Saturday, June 6 (12:30 p.m. - 1:00 p.m.)

Location: Ovation

The rapid emergence of generative artificial intelligence (AI) has introduced new forms of automated support for second language (L2) writing revision, particularly through AI-driven reformulation and the generation of model texts. While both strategies have long been associated with facilitating noticing and uptake in L2 writing development, little is known about how learners cognitively engage with AI-generated input or how attentional processes translate into revision behaviour.

Participants, university L2 writers at B1 & B2, were assigned to either a reformulation or a model-text group and completed an initial essay before receiving AI-generated input (reformulation or model texts) to guide their editing processes. The study investigated what learners notice when comparing their original texts with AI-reformulated versions or AI-generated model essays, and how language proficiency moderates noticing and uptake behaviour as evidenced through eye-tracking data and consequent stimulated recall. Results reveal distinct attentional and revision patterns across conditions. AI-reformulated texts primarily triggered form-focused noticing characterised by narrow fixation clusters and clause-local regressions, leading predominantly to surface-level and microstructural revisions. In contrast, AI-generated model texts prompted broader fixation distributions and paragraph-spanning regressions, reflecting rhetorical comparison and discourse-level evaluation. Language proficiency moderated the noticing-uptake relationship: while B1 learners mainly incorporated lexical or phrasal refinements, B2 learners demonstrated greater capacity to transform rhetorical noticing into macrostructural revisions, including counterargument integration and argument reorganisation. These findings demonstrate that AI-reformulation and AI-generated models afford distinct cognitive pathways for L2 writing revision and highlight the role of proficiency in mediating how learners utilise AI-generated input.

Paper Session 7.2 (Day 3: 1:05 p.m. - 1:35 p.m.)

Qualitative Patterns of Change in Argument Development in a Test of L2 Argumentative Speaking Ability

Author: Jorge Luis Beltran Zuniga (Universidad Autonoma de Guadalajara)

Day: Saturday, June 6 (1:05 p.m. - 1:35 p.m.)

Location: Opera AB

This paper reports on qualitative findings from a study examining how EFL learners' argumentative speaking performance changed across tasks within a scenario-based assessment (SBA). Using 75 paired responses from learners at B1–C1 proficiency levels, the study compared pre-scenario and post-scenario (culminating) tasks on the topic of data tracking. The transcribed responses were coded in three analytic stages to explore changes in argument quality. First, open coding was used to categorize overall change in argument quality. This was followed by a trait-based analysis focusing on displayed elaboration, audience awareness, clarity and grounding of the main claim, and specificity of evidence. Finally, profiles of examinee performance were identified.

Seventy-one percent of examinees produced higher-quality arguments after engaging with the scenario, showing richer elaboration, more specific evidence, and stronger awareness of the simulated audience. Forty-four percent articulated clearer and more grounded argumentative claims. Nevertheless, 24% of examinees displayed no meaningful change in argument quality, and 5% showed weakened quality in their culminating task response (e.g., due to disengagement or topic fatigue). Five performance profiles were identified. Together, these profiles provide insight into how scenario-based tasks can both support and reveal learners' argumentation processes. Overall, the study highlights the value of integrating contextualized, meaning-oriented tasks into language assessment to better capture the interaction between language ability, topical knowledge, and affect. Moreover, findings from the study contribute to growing evidence that scenario-based tests elicit more authentic displays of L2 language use by situating assessment in contexts that engage reasoning within a situated social context.

Bridging Traditional and AI-Assisted Writing Constructs in the Chinese EFL Context

Authors: Jianda Liu (Guangdong University of Foreign Studies), Xuan Zhou (Guangdong University of Foreign Studies)

Day: Saturday, June 6 (1:05 p.m. - 1:35 p.m.)

Location: Symphonie 1

Generative AI (GenAI) is reshaping how students produce written work worldwide. Yet large-scale language tests and classroom tasks continue to rely on pre-AI constructs, and institutional AI policies, particularly in China's EFL higher education context, are underdeveloped. This creates tension between technological realities and local assessment practices. In Chinese universities, the question is whether to restrict AI use or adapt assessment frameworks to reflect AI-enhanced writing competence.

This case study examines how AI-assisted writing differs from traditional writing in Chinese EFL classrooms. Five university students completed two authentic, untimed argumentative tasks: traditional computer-based writing (online dictionary and web search allowed) and AI-assisted writing (with GenAI tools). Data included screen recordings, think-aloud protocols, retrospective interviews, keystroke logs, final products, and AI interaction threads.

Building on established writing process frameworks, we reconceptualize AI as a dynamic task environment element influencing resource use and decision-making. Findings show core process architecture remained consistent across modes. However, AI-assisted writing often took longer, with extended pauses linked to evaluating AI-generated content. AI support eased writer's block and shifted help-seeking from word to phrase and sentence level. Lower-proficiency students revised less with AI, while higher-proficiency students maintained revision habits. AI-assisted products were longer, more accurate, lexically complex, with richer content, though structural organization and argumentative rigor remained similar.

Results indicate AI-assisted writing represents a construct shift requiring new competencies: strategic prompt design, critical evaluation of AI output, and selective integration. Updating assessment frameworks to include these competencies will better align China's EFL writing assessment with technological realities.

Assessing the Impact of Note-Taking Methods on Specialized Vocabulary Knowledge and Lecture Comprehension in EMI Graduate Classroom

Authors: Nataliya Borkovska (University of Toronto), Pelin Irgin (Brock University)

Day: Saturday, June 6 (1:05 p.m. - 1:35 p.m.)

Location: Symphonie 2

Traditional approaches to assessing academic lecture comprehension often treat vocabulary coverage thresholds as key predictors of understanding (Nation, 2006; Webb & Nation, 2017). However, lecture-based studies suggest that comprehension is also mediated by learners' note-taking practices, which influence how information is processed beyond vocabulary knowledge alone (Jin & Webb, 2020; Mežek, 2013). This study investigates the effect of three note-taking conditions on international graduate students' specialized vocabulary learning and lecture comprehension at an English-medium post-secondary institution in Canada. International graduate students ($n = 75$) enrolled in a TESOL program were invited to participate during Intersession 2025, of whom 61 volunteered. Participants were assigned to structured note-taking using the Cornell method ($n = 21$), conventional note-taking ($n = 20$), or no note-taking (control, $n = 20$). Data collection included vocabulary tests, a lecture comprehension test, and an analysis of students' notes. Results showed that the conventional note-taking group had higher baseline knowledge of specialized vocabulary compared to the structured and no-notes groups at the start of the lecture. One-way ANOVA and ANCOVA were used for the quantitative analysis. Results showed that participants who took notes, regardless of note-taking format, achieved higher lecture comprehension scores than those who listened without notes. However, no evidence indicated that structured note-taking led to greater vocabulary gains or comprehension outcomes compared to conventional note-taking. Additionally, the frequency of specialized vocabulary in students' notes did not predict comprehension performance. These findings suggest that the act of note-taking itself, rather than the specific structure or vocabulary captured, is the key factor contributing to lecture comprehension.

Failing Forward in Local Contexts: Marginal Gains and the Infinite Game of Language Learning

Author: Troy L. Cox (Brigham Young University)

Day: Saturday, June 6 (1:05 p.m. - 1:35 p.m.)

Location: Symphonie 3

This study examines how well-known global proficiency frameworks such as the CEFR and the ACTFL Proficiency Guidelines can inform locally adapted, learner-centered assessment. Drawing on data from more than 2,600 university students, it explores how grit, willingness to communicate (WTC), and metacognitive calibration vary across proficiency levels using ACTFL-aligned measures. Participants first completed a locally developed self-assessment diagnostic followed by the Oral Proficiency Interview–Computer (OPIC). Grit increased modestly with proficiency ($F(3, 1900)=14.51, p < .001$), WTC showed stronger relationships ($F(6, 2599)=39.99, p < .001$), and calibration accuracy improved consistently ($F(6, 2609)=15.77, p < .001$). Together, these results suggest that persistence, communicative engagement, and self-awareness develop hand-in-hand as learners advance. Drawing on Carse's infinite game and Brailsford's marginal gains, the paper argues that assessment can help learners fail forward—using reflection, feedback, and small, sustained improvements to support long-term growth. The presentation offers a practical model for implementing proficiency-based formative assessments that balance global comparability with local learning needs.

Context Matters: Proctor Language and Procedural Equity in Remote English Testing

Author: Jieun Kim (Soongsil University)

Day: Saturday, June 6 (1:05 p.m. - 1:35 p.m.)

Location: Symphonie 4

Remote proctoring is increasingly common in English language testing (Isbell & Kremmel, 2020), yet it often requires communication in the target language, unlike in-person centers that may offer support in test-takers' first language. Although prior studies have noted communication challenges in remote testing environments (Green & Lung, 2021; Purpura et al., 2021), few have examined how proctor language affects test validity and fairness. This study investigated the impact of proctor language on test scores and perceptions among English learners in an at-home testing context.

Sixty-seven adult L1-Korean L2-English speakers completed reading and listening tasks adapted from an IELTS practice test, with scripted disruptions (e.g., noise alerts, audio stoppages) introduced to simulate technical issues. Participants were randomly assigned to Korean- or English-speaking proctors. They were then debriefed and interviewed.

Although scores did not differ significantly, English-proctored participants needed more conversational turns to resolve issues, indicating greater communication burden. Interview data revealed concerns about procedural equity, accountability for technical problems, and potential for miscommunication or wrongful accusations. Some participants distrusted at-home testing due to the possibility of undetected malpractice or the fear of being unfairly penalized. The findings call for a more just and holistic approach to remote test design.

Revision and learning transfer effects of actionable high-level suggestions on writing drafts

Author: Yigal Attali¹, Kai-Ling Lo¹, Yena Park¹, Jacqueline Church¹, Andrew Runge¹

Organization: 1: Duolingo

Day: Saturday, June 6 (1:05 p.m. - 1:35 p.m.)

Location: Ovation

Automated written feedback (AWF) holds promise for supporting formative assessment and learning, yet most systems show effects limited to within-task revisions rather than transfer to new writing. A key reason is that existing feedback focuses on surface-level corrections rather than actionable, process-oriented guidance that helps writers clarify ideas, strengthen arguments, and reflect on their choices.

We developed an AI-based feedback agent that provides high-level, actionable suggestions on persuasiveness/engagement, idea development, and coherence. After submitting a draft, users receive a targeted suggestion and revise their text; the system then evaluates feedback uptake, providing encouragement or further pointers. This feedback–revision loop aims to foster sustained writing development.

In an empirical study with 782 participants completing two feedback sessions (one persuasive or narrative prompt each), writing scores were analyzed using a mixed-effects model. Results showed significant revision effects within sessions and a significant learning effect across sessions: second-session drafts were stronger than first-session drafts, suggesting learners transferred strategies to new tasks.

These findings address a long-standing limitation in AWF research by demonstrating learning transfer across tasks. Consistent with Bachman's (2007) task/context-focused perspective, writing ability and feedback engagement are inherently context-bound. Effective AWF must therefore be evaluated not only through global performance measures but through how learners act on feedback within specific writing tasks and communicative goals.

Paper Session 8.1 (Day 3: 4:25 p.m. - 4:55 p.m.)

Can a process-oriented approach to listening assessment and instruction meet the needs of standards-based L2 listening assessment?

Author: Jianling Liao (Arizona State Uni.)

Day: Saturday, June 6 (4:25 p.m. - 4:55 p.m.)

Location: Opera AB

While the process-oriented learning approach is widely recognized, its applications in L2 listening assessment and instruction remain limited. Top-down processing and compensatory strategies continue to dominate L2 listening assessment and teaching. Consequently, product-oriented training practices have often resulted in unsatisfactory assessment outcomes, heightened learner anxiety, and a lack of effective strategies for improvement (Newton 2024; Nunan 2002). This study investigates how a process-oriented approach to learning and assessment influences college L2 learners' listening development and their performance in standardized L2 listening tests. The following two research questions were investigated:

- 1) What L2 listening sub-skills are facilitated by a process-oriented training approach among advanced L2 learners of Chinese, and how do these sub-skills develop over a semester?
- 2) How does this process-oriented training approach affect learners' performance on standards-based listening tests, and how are these test results correlated with students' classroom-based listening assessment outcomes?

Fifteen advanced-level college learners of Chinese participated in the study. The process-oriented approach to learning and assessment was built upon Anderson's (1995) and Field's (2019) cognitive listening processing model and processing instruction framework. Each week, students completed in-class and out-of-class listening activities, during which they practiced lower-level processing skills and higher-level discourse-construction skills. Students also completed: (1) Four monthly authentic assessment tasks; (2) A standardized listening test at the beginning and end of the semester; and (3) A pre- and post-semester self-assessment of listening processing skills. The results suggest that a process-oriented approach to listening assessment and instruction may tap into key constructs of L2 listening skills that are evaluated in standardized listening assessments.

The Role of Linguistic Knowledge and Input Text Modality in an Integrated Reading-to-Speaking Task

Authors: Shungo Suzuki (Nagoya University), Judit Kormos (Lancaster University)

Day: Saturday, June 6 (4:25 p.m. - 4:55 p.m.)

Location: Symphonie 1

Integrated speaking tasks have gained increasing attention in second language (L2) assessment due to their authenticity and alignment with academic communication skills. Among them, Reading-to-Speaking (RtoS) tasks require test-takers to reproduce key information from a source text, highlighting the importance of content-related assessment criteria. However, limited research has examined the validity evidence underlying such criteria in relation to linguistic knowledge, or how input modality (e.g., reading-only vs. reading-while-listening) may influence task performance. This study, therefore, investigated the extent to which learners' linguistic knowledge predicts the successful reproduction of source text information in RtoS tasks and whether this relationship varies across input modes.

A total of 128 intermediate-level Japanese learners of English completed linguistic knowledge tests measuring productive vocabulary size, sentence construction speed, and grammatical judgement speed and accuracy, followed by two RtoS tasks under counterbalanced modality conditions. Learners' speech was analysed for the accurate reproduction of 14 key idea units per text. Generalised Linear Mixed-Effects Models revealed significant main effects of vocabulary size and grammatical judgement accuracy, indicating that both productive lexical and controlled grammatical knowledge facilitated successful idea reproduction. A marginal interaction further suggested that bimodal input presentation (reading-while-listening) attenuated the role of receptive grammatical skills. These findings offer construct validity evidence for assessing integrated speaking performance through content-related dimensions and underscore the need to account for modality effects in defining and interpreting the construct of RtoS tasks.

Picture-based narrative writing tasks in language assessments for young learners: Effects of task design on test performance and implications for validity

Author: Haeun {Hannah} Kim (The University of Melbourne)

Day: Saturday, June 6 (4:25 p.m. - 4:55 p.m.)

Location: Symphonie 2

This study investigates whether picture prompt features known to influence spoken narratives similarly affect written narratives, and whether linguistic differences in responses lead to score differences. The study focuses on picture-based story writing tasks in ACCESS for ELLs. In phase one, eleven tasks were qualitatively analyzed to identify major differences in picture prompt characteristics. From this analysis, two prompts per grade were selected that contrasted most clearly in potential complexity features. In phase two, responses and scores from 800 students—400 in grade 5 and 400 in grade 7 (200 per task)—were quantitatively analyzed. In the final phase, stimulated recall interviews with eight children provided further insight into how learners engaged with the prompts.

Findings showed that grade 5 tasks differed in accuracy, repeated content lemmas and pronouns, and index of narrative complexity (INC), while grade 7 tasks differed in content word abundance, syntactic complexity, repeated content lemmas and pronouns, and INC. However, only some of these linguistic differences translated into score differences. Regression analyses revealed that for grade 5, repeated content lemmas and pronouns and INC interacted with task type to influence scores, while for grade 7, syntactic complexity and accuracy showed interaction effects. Stimulated recall interviews corroborated these findings, indicating that prompt characteristics influenced how children constructed their narratives. The study highlights the challenges of designing comparable multimodal assessments and provides guidance for developing picture prompts that support valid interpretations of test scores in the context of large-scale standardized tests for young learners.

Insights from Primary Reading Assessments in Southeast Asia: The Tiger and the Tapir

Author: Louise Courtney (Australian Council for Education Research)

Day: Saturday, June 6 (4:25 p.m. - 4:55 p.m.)

Location: Symphonie 3

SEA-PLM is a joint effort by the Southeast Asian Ministers of Education Organization (SEAMEO) and UNICEF, with the Australian Council for Educational Research (ACER) providing test development support. It represents the first cross-national assessment of its kind focused on Grade 5 students in Southeast Asia. SEA-PLM provides robust comparative data on how children perform in reading, writing, and mathematics near the end of primary school, offering critical insights into the strengths and gaps in foundational learning.

This paper focuses on how local contexts were at the forefront of consideration when developing the large-scale assessment across Vietnam, Malaysia, The Philippines, Myanmar, Cambodia, Laos PDR and Timor Leste. To develop a valid assessment, we needed to balance tensions between regional and local concerns; to find a sweet spot where each country could assess students using the same reading texts and items, while being satisfied that the assessment was accessible, accurately reflecting the cultural and religious contexts of their own students. The main study was in 2019, with the second main study in 2024.

We discuss examples of the methods used to develop a shared context for the regional assessment and share the 2024 Reading assessment data from the regional report. By demonstrating SEA-PLM's unique approach to assessing literacy skills this paper may provide models for future success in large-scale regional assessments, where local culture and language contexts are key considerations.

AI-Facilitated Evaluation of Chinese as a Second Language Integrated Speaking Performance: Measurement Features, Human Agreement, and Criterion-Relatedness

Author: Xiaozhu Wang (Beijing Language and Culture University)

Day: Saturday, June 6 (4:25 p.m. - 4:55 p.m.)

Location: Symphonie 4

Automated speech evaluation has greatly improved the efficiency and scalability of speaking assessments since the early 21st century. With the rapid advancement of large language models (LLMs) and generative artificial intelligence (AI), their applications in language education have expanded considerably. However, research on AI-based speech evaluation remains limited, particularly for languages other than English.

To examine AI's capability in assessing the integrated speaking performance of Chinese as a second language and to provide practical insights for classroom assessment, this study investigated the potential of AI in rating integrated L2 speaking tasks using both holistic and analytic scoring modes. The analysis focused on AI's measurement characteristics, its agreement with human raters, and its correlation with independent speaking and general proficiency measures.

Results indicated that the analytic rating mode better supports explainable and interpretable AI evaluation. AI scoring demonstrates stable and meaningful measurement, particularly in discourse-related dimensions of integrated speaking performance, while challenges remain in evaluating natural speech and ensuring score accuracy for classroom use. The validity and construct representation of AI facilitated speaking evaluation are discussed.

Cognitive Diagnostic Modelling (CDM) for Construct Validity Evidence in e-TEP Reading Comprehension Tasks: A Mixed-Methods Approach to Test Validation

Authors: Fahri Yilmaz (Independent), Secil Alaca (Hacettepe University), Merve Selçuk (Bilgi University), Zeynep Urkun (Independent)

Day: Saturday, June 6 (4:25 p.m. - 4:55 p.m.)

Location: Ovation

This study presents a comprehensive validity argument for the reading comprehension (RC) section of the e-TEP, a high-stakes, CEFR-aligned EFL exam in Türkiye. The study employs a four-phase, mixed-methods design, triangulating quantitative data from Cognitive Diagnostic Modeling (CDM) on 686 test-takers with qualitative evidence from stimulated recalls (B1/C1 levels) and expert judgments. A final phase investigates the washback of these findings by presenting them to the test's item writers in a professional development workshop. Findings provide strong evidence for the test's construct validity, confirming a clear progression of cognitive demand across proficiency levels, with C1 tasks engaging more complex attributes than B1 tasks. The methodological approach proposed also uncovers opportunities for enhancement by identifying divergences between intended and actual cognitive processes used by test-takers. Furthermore, the CDM analysis provides empirical evidence to guide test refinement by identifying foundational attributes that do not significantly correlate with overall reading proficiency. By bridging a global validation framework with a specific local context, this study offers a validation model that demonstrates the importance of using data-based evidence from test-takers. It concludes by highlighting how creating a direct feedback loop to item writers improves validity research into a practical tool for enhancing assessment quality and fairness.

Paper Session 8.2 (Day 3: 5:00 p.m. - 5:30 p.m.)

Accent Effects of Text-to-speech Voices in Listening Assessment

Authors: Xiaowan Zhang¹, Jacqueline Church¹, Andrew Runge¹, Geoff LaFlair¹

Organization: 1: Duolingo

Day: Saturday, June 6 (5:00 p.m. - 5:30 p.m.)

Location: Opera AB

Accent is an essential feature of listening assessments that may interact with the local context of test use. This study examines how three inner-circle English accents (American, British, and Australian), produced by text-to-speech (TTS) voices, affect performance in a large-scale, high-stakes English dictation task. It also investigates how accent interacts with test-taker variables (proficiency, accent familiarity, country of testing) and item characteristics (item difficulty and speech rate).

Six TTS voices representing the three accents and two genders were evaluated by experts and achieved average ratings of 4.0 or above on a 1–5 Likert scale across six quality dimensions: enunciation, prosody, naturalness, audio quality, accentedness, and likability. Their average transcription accuracy was at or above 90%, confirming high intelligibility.

The main study used 30 dictation items varying in difficulty, divided into five parallel forms, each recorded in a different TTS voice and administered to about 300 test takers. In total, 8,916 test takers from more than 160 countries participated. Generalized mixed-effects modeling showed that Australian-accented TTS voices produced lower scores than North American voices after controlling for all other variables. These effects were moderated by proficiency, item difficulty, and speech rate, but not by country of testing or self-reported familiarity. The strongest moderation was observed for item difficulty: accent effects were largest on easier items and diminished on harder ones. Survey results indicated limited exposure to Australian accents. These findings align in part with prior research on human voices, suggesting that Australian accents present greater challenges across delivery modes.

Systematizing local concerns about a large-scale, standardized speaking test for young language learners

Authors: David MacGregor (WIDA at the University of Wisconsin), Mark Chapman (WIDA at the University of Wisconsin), Fabiana MacMillan (WIDA at the University of Wisconsin), Meg Montee (Georgetown University), Burak Sene (Iowa State University)

Day: Saturday, June 6 (5:00 p.m. - 5:30 p.m.)

Location: Symphonie 1

We report on a project to address local users' concerns with student performance on a large-scale computerized assessment of English-speaking proficiency for English learners in grades 1-12 in public schools in the United States while maintaining the efficiencies afforded by computer delivery. Specifically, we study the feasibility of an enhancement to allow students to respond twice to a speaking prompt, and the effects of such an enhancement on student performance and educator perception of assessment results.

Results from a pilot study showed that students could navigate the revised tasks and use the enhancement when desired. Interviews with students and educators indicated a strong desire to apply it to the test. In a second phase, 422 English learners responded to four speaking tasks. These students were given the opportunity to respond once or twice to each task. Students' responses were captured by the online test platform and scored by trained raters. Students' responses were transcribed using an automated speech-to-text program and analyzed for word count and duration. Additionally, researchers visited six study sites to observe test administrations, then interviewed test administrators (TAs) and students. Finally, all TAs were asked to complete a survey regarding their impressions of the impact of the enhancement on student performance.

We focus on how the study design and findings allowed us to address a complex problem that arose from the local context of educators drawing inferences from students' scores about student performance, along with the meaningfulness of a computer-mediated speaking test for a marginalized population.

Human versus Machine Rater: Investigating the Influence of Response Length in Essay Scoring

Authors: Dellannia Segreti (University of Toronto), Jia Guo (University of Toronto), Athena Tassis (University of Toronto)

Day: Saturday, June 6 (5:00 p.m. - 5:30 p.m.)

Location: Symphonie 2

This study examined the extent to which response length influences both LLM-based and human scores. Situated in the APLUS context, the study analyzed a corpus of post-secondary students' writing samples that were double-scored: once by trained human raters using a validated analytic rubric and a fine-tuned LLM model. Statistical and qualitative comparisons were conducted to explore whether and how essay length correlates with overall scores and performance on analytic traits. Preliminary findings revealed a significant positive correlation between response length and assigned scores for both LLMs and human raters. Findings suggest that while both rating methods are influenced by essay length, LLM-based scoring is more prone to over-reward verbosity, raising concerns about construct underrepresentation and fairness.

Transparent to Whom? Localizing the Principles of Language Assessment Transparency

Authors: Nathaniel Owen (Oxford University Press), Ananda Senel (Oxford University Press)

Day: Saturday, June 6 (5:00 p.m. - 5:30 p.m.)

Location: Symphonie 3

In high-stakes language assessment, transparency is often treated as a universal principle, essential for fairness and validity. This approach, however, frequently relies on a 'techno-rational' model where transparency is equated simply with disclosure of technical information, overlooking the critical role of local context and assuming a single model of assessment transparency will be understood and experienced uniformly by diverse stakeholder groups worldwide. This study addresses this global-local tension, arguing that our understanding of transparency remains fragmented because it has not adequately accounted for how it is enacted and perceived locally.

We employed a mixed-methods approach to investigate this tension. First, a systematic review of 61 articles (2000–2024) mapped the evolution of transparency as a global concept. Second, we gathered empirical evidence from local stakeholders of a high-stakes computer-adaptive English language proficiency test (299 test takers and 41 test center managers) analyzing their experiences through factor and cluster analysis.

Our synthesis revealed a critical disconnect. While the literature showed a conceptual evolution of transparency from a technical to a social concern, our empirical data demonstrates it is not experienced universally. Factor analysis identified six dimensions of assessment transparency. The cluster analysis revealed four distinct response patterns with significant variations across local demographic contexts, specifically country of origin and gender.

Our findings challenge the notion of a monolithic, one-size-fits-all assessment transparency model. We propose a 'layered transparency' framework that accommodates differentiated stakeholder needs, offering a more equitable path forward by being responsive to the specific, context-bound assessment needs of local stakeholders.

Test impact on teachers' wellbeing—a case study of NMET in China

Authors: Xiao Cheng (East China Normal University), Xunyi Pan (Zhejiang University), Feier Ji (East China Normal University)

Day: Saturday, June 6 (5:00 p.m. - 5:30 p.m.)

Location: Symphonie 4

The impact of high-stakes exams have long been a popular topic in language testing research. However, psychological impact of tests has been under researched, with the few studies touching upon students' wellbeing (Macqueen et al., 2019; Winke, 2011; Zhang, 2021). The current study explores the impact of the National Matriculation English Test (NMET), the most important high-stakes English exam in China, on teachers' wellbeing. Combining the frameworks of test impact/washback (Alderson and Wall, 1993; Wall and Horak, 2008; Saville, 2009) and Van Dick and Wagner's (2001) model of teacher stress, we used exploratory interviews with 45 teachers to identify stressors related to the implementation of NMET reform, perceived mental, physical and behavioural impact of the test reform, and the coping strategies at the societal, institutional and individual levels. A stress questionnaire with three scales—stress, impact and coping strategy— was designed on the basis of the interview results and tried out with 117 senior high school English teachers in Shanghai. In so doing, we attempted to validate the questionnaire and to explore the impact of the latest NMET reform on Shanghai English teachers' wellbeing.

The small scale survey showed that the instrument we designed showed a good validity and reliability; the high stress level of Shanghai English teachers indicated that the test reform imposed a severe negative impact on teachers' mental and physical wellbeing, while the behavioural impact was relatively minor; the increasing workload caused by the test reform appeared to be the primary stressor, followed by the increasing students' mental problems and the tension in family-work balance; teachers tended to rely on individual resources and informal support from colleagues to cope the stress, while the institutional and societal support was insufficient; differences were found among teachers of different age.

Modeling Classical Item Difficulty in a Medical English Test Using Semantic Embeddings, Retrieval–Entailment Features, and Answer-Choice Interactions

Author: Peter Kim (Cambridge Boxhill Language Assessment)

Day: Saturday, June 6 (5:00 p.m. - 5:30 p.m.)

Location: Ovation

Accurately estimating item difficulty is fundamental for maintaining validity and fairness in global and local language assessment. And in recent years, advances in NLP, machine learning, and AI have enabled exciting, state-of-the-art approaches to model and predict what makes test items difficult (for example, He & von Davier, 2024; Jin et al., 2023; Lu et al., 2024). This study provides initial but promising evidence that a multi-domain semantic modeling approach combining embedding, retrieval–entailment, and answer-choice interaction features can serve as effective models for predicting classical item difficulty (p-value) in English assessments for specific context. The dataset comprised of reading and listening comprehension items from an English proficiency test for medical professionals, OET (Occupational English Test) Pulse. Each included a passage text, question stem, answer choices, and CEFR metadata. We engineered features in three novel domains: 1) retrieval-based features which capture the degree to which the relevant information for answering a question can be directly retrieved from the passage based on semantic similarity; 2) entailment features quantify how strongly the passage logically supports the correct answer compared to distractors, reflecting inferential reasoning demands; 3) question and answer features represent the confidence of an automated question-answering model in selecting the correct answer, indicating how unambiguous or model-accessible the item is (Benedetto 2023; Winklerová 2024; Ben Abacha & Demner-Fushman 2019; Shen et al. 2023). The results demonstrate that combining semantic embeddings with retrieval–entailment and answer-choice interaction features can meaningfully predict item facility index, even with limited data.

Demos Summaries

Day: Thursday, June 4 (9:40 a.m.-10:45 a.m.)

Location: Ovation

Facilitator: Claudia Harsch

Assessing multilingual learners in kindergarten: incorporating age-appropriate design features into a high-stakes test

Authors: Mark Derek Chapman¹, Gordon Blaine West², Marcy Olson¹, Beth DeVito¹, Christy Shepherd³, Regina DiBella³

Organizations: 1: WIDA at UW-Madison, 2: National Institute of Education, Nanyang Technological University, 3: Center for Applied Linguistics

This demo introduces a newly developed kindergarten English language proficiency (ELP) assessment designed for multilingual learners in their first year of school. Finalized in 2025 through collaboration between WIDA and the Center for Applied Linguistics, WIDA ACCESS for Kindergarten satisfies U.S. federal regulations requiring annual measurement of students' academic English development beginning in kindergarten. Because very young learners (ages 5–6) present unique developmental and affective needs, the test was designed to be developmentally appropriate, engaging, and supportive. Drawing on 33 hours of video recordings from cognitive labs, pilot, and field tests, the presenters will illustrate how the assessment balances linguistic and cognitive demands with scaffolding and motivational supports. Key features include a one-on-one, paper-and-pencil design that allows test administrators to interact closely with students using scripted verbal scaffolding, as well as a storybook-based theme that contextualizes test tasks in familiar ways. Listening and speaking tasks are integrated for more natural interaction, followed by writing and reading tasks reflecting typical classroom activities. Tasks are supported by colorful images, card-based interactive activities, and other multimodal inputs to foster engagement. The test also incorporates stopping rules and off-ramping protocols to prevent frustration and support positive affective responses. Participants in this session will view student videos to understand how key test design features function and examine speaking and writing samples that highlight the range of productive language young learners can demonstrate. The session will offer concrete insights into designing valid and supportive ELP assessments for young learners.

Context-Aware Adaptive Assessment Model (CAAM): Bridging Global Frameworks and Local Realities in LLM-Driven Language Testing

Author: Kamaledine Tabine (Foreign Service Institute)

This demonstration showcases the Context-Aware Adaptive Assessment Model (CAAM), an innovative AI- and LLM-driven framework designed to integrate contextual adaptation, adaptive proficiency scaling, and algorithmic transparency into language testing. CAAM bridges the gap between global proficiency frameworks and the local linguistic, cultural, and pedagogical realities of test-takers, offering a scalable yet contextually grounded approach to language assessment.

The session will feature a live simulation of CAAM's adaptive system in action. The presenter will select from more than twenty languages, including English, French, Arabic, Mandarin, Spanish, and Swahili, and choose a communicative domain such as a formal business exchange, academic seminar, casual street conversation, or idiomatic discourse. Once the language and domain are selected, CAAM dynamically generates authentic test content through Dynamic Context Mapping and adjusts task difficulty and scoring trajectories in real time through Adaptive Proficiency Scaling.

Throughout the demonstration, the Ethical AI Governance Log will display the rationale behind each scoring decision, making the model's reasoning transparent and auditable. This feature exemplifies how explainable AI principles can enhance fairness, accountability, and validity in assessment.

By embedding contextual awareness and ethical transparency within adaptive testing, CAAM reimagines language assessment as an equitable, interpretable, and context-sensitive process that honors both global standards and local communicative realities.

Improving Proctoring Consistency in Remote English Language Assessment: A Feedback and Training Tool for Proctor Decision Making

Authors: William Belzak (Duolingo), Yigal Attali (Duolingo)

Remote proctoring plays an important role in safeguarding the validity and integrity of large-scale online English assessments (Isbell et al., 2023). However, proctoring judgments often vary across individuals and contexts, creating challenges for consistency and fairness (Belzak et al., 2024). This demonstration introduces the Proctor Calibration Tool (PCT), a system designed to improve decision consistency and fairness in remote test security operations. The PCT measures each proctor's alignment with expert judgments, provides real-time feedback during the proctoring process, and generates targeted training materials from authentic test sessions. The demonstration will introduce the challenge of maintaining consistent proctoring quality in a globally distributed testing system and present a guided overview of the PCT's interface and functionality. Attendees will see how proctors receive immediate feedback on their decisions and access a training library of expert-reviewed cases. Empirical results show a 9% improvement in decision consistency after two months of tool use, illustrating the potential of continuous feedback loops between proctors and experts to promote equitable test administration. Grounded in naturalistic decision-making theory (Klein, 1998) and heuristics and biases research (Tversky & Kahneman, 1974), the PCT allows for proctor calibration to be an ongoing, feedback-driven process. By combining global standards with sensitivity to local contexts, this approach demonstrates how technology can strengthen human judgment, enhancing fairness and reliability in high-stakes remote language testing.

A Fully AI-mediated Learning-Oriented Assessment of Interactional Competence: A Demonstration from an Operational Test in Japan

Authors: Anastasia Ramjag¹, Masaki Eguchi¹, Elliott Bruce¹, Mao Saeki¹, Hiroaki Takatsu¹, Max Thomson¹, Daniel Travin¹, Yuta Ishida¹, Sadahiro Yoshikawa¹, Yoichi Matsuyama¹

Organization: 1: Equimenopolis, Inc.

Interactional competence (IC)—the ability to co-construct meaning with interlocutors—has become a key focus in language assessment, yet scalable and context-sensitive approaches remain limited. Recent research shows that conversational AI can elicit IC features such as negotiation and repair under standardized yet authentic conditions (Litman et al., 2018; Eguchi et al., 2025; Timpe-Laughlin et al., 2025). However, creating operational, learning-oriented assessments that integrate elicitation, scoring, and feedback remains a largely uncharted area.

This demonstration presents the Intelligent Language Learning Assistant (IntelLA), a multimodal virtual agent implemented in the operational LANGX Speaking test. Used for mid-stakes purposes such as placement and progress checks in Japanese secondary and tertiary contexts, LANGX Speaking represents a concrete application of AI in assessing real-world communicative ability. Drawing on frameworks of interactional competence (Galaczi & Taylor, 2018), task-based performance assessment, and formative feedback (Lee, 2015), IntelLA conducts interactive elicited conversation tasks; and our scoring system evaluates the interaction and generates checklist-based feedback on interactional features (cf. Nakatsuhara et al., 2018).

The session includes a demonstration and analytic walkthrough illustrating how design principles grounded in the construct of IC are localized for Japanese educational needs. By linking automated elicitation, scoring, and feedback within an operational test, this project showcases a model for scalable, learning-oriented assessment of IC. The demonstration highlights how adaptive AI can bridge theories and local practices, contributing to more contextually grounded language assessment. The talk concludes with implications for the use of (generative) AI technologies in developing localized assessments.

WIPs Summaries

Day: Thursday, June 4 (10:50 a.m.-12:20 p.m.)

Location: Ovation

Facilitator: Jee Wha Dakin

1. Localizing the “global” CEFR framework for a national assessment project in Cuba

Authors: Yoan Martínez Márquez (Universidad de las Ciencias Informáticas), Claudia Harsch (University of Bremen)

The Ministry of Higher Education (MES) in Cuba initiated an improvement policy for the teaching and learning of the English language in 2015. MES introduced the Common European Framework of Reference as a globally accepted proficiency framework, with level B1 being set as the exit standard for BA studies. A certification system to teach and assess students' English proficiency became a demand.

We report on the design and development of the high-stakes exam at the end of the certification system. The project commenced in 2017, with partners from the MES, Cuban local universities, international support from the British Council and the Belgian VLIR network, an external advisor, and a workshop grant from ILTA. The exam is to be aligned with the CEFR to achieve international recognition of the certification results. We will present our approaches to addressing the challenges and tensions that arose from localizing the CEFR categories and descriptors to adapt them to the Cuban HE context.

Concerning theoretical frameworks, the CEFR served as a “global” proficiency framework for this endeavor. In addition, the manuscript for aligning exams to the CEFR (Council of Europe, 2009) and the handbook for aligning language education to the CEFR (British Council et al., 2022) have served as guiding frameworks for local test development and alignment. We will present the outcomes and data collected during the first three alignment stages: familiarization, specification, and standardization. The final two stages of standard setting and validation are yet to come in the next phase of the project.

2. From Humans to Chatbots: Emotional Mediation between Task Mode and Performance in L2 Speaking Assessment

Authors: Lin Shi (Zhejiang University), Shuting Zhang (Zhejiang University), Yang Zhang (Zhejiang University)

As GenAI chatbots gain prominence, comparing human–AI and human–human interactions becomes crucial, since learners' emotional responses may differ across interlocutor types. Despite growing interest, empirical comparisons remain scarce, and most existing studies have been conducted in classroom contexts rather than assessment settings, where time pressure and evaluative stakes may elicit distinct affective responses. These considerations highlight the need for empirical evidence on GenAI's impact on test-takers' emotions and the feasibility of AI-mediated speaking assessment.

Drawing on Positive Psychology (e.g., Dewaele & MacIntyre, 2014), this quasi-experimental study investigates: (1) how human–human and human–AI speaking tasks in a testing context affect Chinese EFL learners' state emotions (FLE, FLA, WTC), and examines (2) to what extent these emotions mediate speaking performance. It also discusses the feasibility of GenAI-mediated speaking assessment within localized EFL testing contexts.

To address these aims, sixty Chinese university students of varying English proficiency will complete two paired-discussion tasks (human–human and human–AI) in a counterbalanced design, with a brief emotional washout between tasks. Pre- and post-task questionnaires will measure FLE, FLA, and WTC, complemented by an AI-use Perception Scale in the human–AI condition. Multimodal data (speech, facial expressions, gestures) and stimulated recall interviews will provide additional insight into learners' emotional experiences.

As a work in progress, feedback is welcome on the planned analyses and on how affective data can best be captured. Insights from the language testing community will help refine the analytical design and strengthen the study's implications for GenAI-mediated speaking assessment.

3. Cheating is in the eye of the beholder: exploring views on academic integrity and cheating in contemporary language assessment

Author: Anna Soltyska (Ruhr-Universität Bochum)

Academic integrity and the prevention of cheating seem to have come to the fore in post-pandemic language assessment. However, the assessment literacy and training needs of different assessment stakeholders in this regard have yet to be explored to be adequately addressed.

This mixed-methods study aims to determine whether testing and assessing foreign language skills online or in technology-mediated contexts has impacted stakeholders' understanding of academic integrity. In particular, it will identify aspects of cheating, test security and academic integrity that assessment research has recently focused on. Furthermore, it will analyse how various stakeholder groups understand the concepts of academic integrity and cheating and in what ways these are addressed in foreign language classes at university level. Finally, the study will examine to what extent the current approach to language assessment literacy embraces academic integrity and related concepts.

To this end, a scoping review of articles published in selected language assessment journals, as well as papers presented at major conferences on language assessment worldwide, will identify key areas of concern relating to cheating and test security. A nested qualitative analysis will then be conducted, collecting data through an online questionnaire for language teachers and students (test-takers), based on earlier assessment literacy surveys and issues emerging from the scoping review, and conducting semi-structured interviews with selected teachers and representatives of standardised test providers. Case studies will offer further insight into academic integrity curricula and classroom procedures.

4. Comparing BERT, LLM, and Human Ratings of Elementary Students' Writing Strategies

Authors: Angelie Ignacio (University of Toronto), Hajung Kim (University of Toronto), Bosco Zeng (University of Toronto), Han Lai (University of Toronto), Eunice Eunhee Jang (University of Toronto)

Automated essay scoring (AES) is increasingly integrated into classrooms, yet questions remain about whether such systems capture meaningful aspects of student writing or rely on superficial cues. While transformer-based models such as BERT and large language models (LLMs) offer scalability, their interpretability is limited, raising concerns about validity and construct representation (Bennett & Zhang, 2015). In contrast, human raters use rubric-aligned criteria but face challenges of bias, inconsistency, and resource demands (Bridgeman, 2013; Knoch et al., 2021).

This study will investigate whether AES models (BERT and LLM) of student self-regulated writing strategies, specifically student-generated questions (SGQ) and planning, align with human judgements. SGQ and planning are essential for critical thinking and coherence in writing (Chai, 2006, Chin & Brown, 2002; Hayes & Flower, 1980). Using data from 534 Ontario students in Grades 3–8, SGQs was rated analytically for quality, vocabulary, grammar, and spelling, while planning will be rated holistically for appropriateness and depth.

Two phases are planned. Phase 1 will examine agreement among human raters, BERT, and LLM using κ coefficients, correlations, and Many-Facet Rasch Modeling to explore rater and model facets. Phase 2 will apply SHAP (SHapley Additive exPlanations; Lundberg & Lee, 2017) to interpret BERT predictions and identify linguistic features that most strongly influence scoring.

By combining rater modeling and model interpretability, this study contributes to construct-focused validation of AES. Aligned with the LTRC theme, it explores how global AI technologies can be responsibly localized to support, rather than replace, human judgment in formative writing assessment.

5. Language Classroom Assessment and its Psychological Consequences: Examining ESL Learners' Perceptions of Assessment, Mental Health, and Subjective Wellbeing

Authors: Wei Yan (Thompson Rivers University), Glenda Christou (Queen's University), Peiyu Wang (Queen's University)

Although consequences of language assessments have been widely examined from pedagogical and policy perspectives, relatively little is known about their psychological dimensions (Wen & Liu, 2024). Studies show that assessment situations can provoke stress and anxiety (von der Embse et al., 2018; Wass et al., 2020), which in turn influence students' subjective wellbeing and academic achievement (Vogl & Pekrun, 2016). Furthermore, research on students' perceptions of assessment (Brown & Hirschfeld, 2008) has seldom examined how these perceptions interact with wellbeing in English as a Second Language (ESL) learning contexts.

This study aims to investigate ESL learners' knowledge and perceptions of mental health and illness, their beliefs about subjective wellbeing and assessment, and how these factors are related to their academic performance. This study will be situated at a mid-sized, primarily undergraduate, university in Western Canada. A quantitative approach will be adopted, and questionnaires will be collected from ESL students at this university.

The questionnaire will include four main sections: 1) knowledge and attitudes towards mental health, 2) subjective wellbeing, 3) perceptions of assessment, and 4) demographic information including students' ESL course grade. A total of approximately 500 survey responses is expected to be collected and analyzed using descriptive statistics, CFA, and SEM. The analyses will examine relationships among knowledge and attitudes toward mental health, perceptions of assessment, subjective wellbeing, and assessment outcomes. This study contributes to the field of language assessment by integrating psychological factors—mental health literacy, subjective wellbeing, and assessment perceptions—into the examination of assessment consequences.

6. From Global Models to Local Uptake: Examining AI-Mediated Feedback and Instructor Mediation in Japanese Students' Academic English Writing

Authors: Johanathan Woodworth (MSVU), Darlene Yamauchi (Gunma University)

This Work-in-Progress study explores how Japanese students and instructors in English for Academic Purposes courses interpret and mediate feedback produced by generative AI systems trained on global English data. Conducted at Gunma University, Japan, the project compares three feedback conditions, AI-only, instructor-only, and hybrid (AI + instructor-curated), to examine equity, legitimacy, and instructor agency in multilingual academic writing. Approximately 50 students are randomly assigned by draft to one feedback condition. Instructors participate in the hybrid and instructor groups, providing human feedback and reflecting on their experiences through interviews and short surveys. Data include writing drafts, feedback texts, revision logs, and bilingual interviews with both students and instructors. Analyses compare pre- and post-revision writing quality and trace how perceptions of fairness, authority, and workload influence feedback uptake and classroom practices. Preliminary findings highlight clear contrasts in linguistic focus between AI and instructor feedback and reveal instructor ambivalence, valuing efficiency yet expressing concern about loss of pedagogical control and the cultural appropriateness of AI comments. The study contributes to the LTRC 2026 theme "Global to Local Language Assessment" by showing how globally trained AI systems acquire new meanings through instructor mediation and local pedagogical negotiation in a Japanese university context.

7. L2 Test Takers' Cognitive Engagement in Video-Based Listening Assessments: An EEG Study

Authors: Shanshan He (University of Western Ontario), Ruslan Suvorov (University of Western Ontario)

The integration of video in L2 listening assessment seeks to enhance test authenticity by mirroring real-life multimodal communication. However, research findings on the effects of visual input on test performance remain inconclusive. These inconsistencies have been associated with differences in the degree of visual support provided by the videos and the focus of test items (global vs. local items). However, little is known about how L2 test takers cognitively engage with video input during listening assessment. Understanding this engagement is essential for establishing evidence of cognitive validity based on test takers' response processes.

To address this gap, the present study employs electroencephalography (EEG) to capture real-time neural activity during L2 listening comprehension and multiple-choice item completion under three conditions: content-rich video, content-deficient video, and audio-only. Using a within-participants design, 45 L2 English learners at a Canadian university will complete 20 short academic listening passages and 40 multiple-choice items targeting global and local information. EEG data will be recorded with a 32-channel BioSemi ActiveTwo system, supplemented by a visual digit span test and a post-test questionnaire. Time-frequency analyses will extract theta and alpha power as indicators of cognitive effort. Mixed-effects models will examine differences across conditions and relationships between EEG and performance. This work-in-progress presentation will invite discussion on data analysis, result interpretation, and methodological considerations to link EEG-based findings with cognitive validation frameworks, thereby deepening understanding of test takers' cognitive engagement and informing the development of theoretically grounded, valid L2 listening assessments.

8. Tracking L2 writing development with China's Standards of English Language Ability: Application of longitudinal cognitive diagnosis models

Authors: Hui Wang (Zhejiang University), Shangchao Min (Zhejiang University)

Despite the growing interest in diagnostic assessment, few studies have employed cognitive diagnosis models (CDMs) to diagnose writing ability. Furthermore, the existing studies are cross-sectional, limiting their ability to fully capture the dynamic nature of skill development and individual differences in developmental trajectories. Therefore, this study utilized longitudinal CDMs to track the developmental pathways of Chinese EFL learners' writing subskills and further explore how English proficiency influences the developmental patterns. We recruited 350 first-year undergraduates from a prestigious university in China. Students were required to complete three parallel argumentative diagnostic writing tasks (each at least 160 words within 30 minutes) across three waves: T1 (completed in August 2025), T2 in November 2025, and T3 in January 2026. We employ a 36-item checklist developed using the descriptors of China's Standards of English Language Ability to measure students' writing performance. Six attributes are targeted: Argumentation (A1), Organization (A2), Vocabulary (A3), Syntax (A4), Mechanics (A5), and Writing Strategy (A6). Q-matrices have been developed to specify item-attribute relationships. Items of A1-A5 will be rated by trained raters, whereas A6 will be operationalized using keystroke-based indicators. Regarding data analysis, we will use the TDCM package in R to fit five longitudinal CDMs. The best-fitting model will be used to estimate attribute-level mastery probabilities, growth rates, and transition probabilities to track subskill development. To explore proficiency effects, students will be grouped into three levels using scores from the university's English diagnostic test. We will then fit multi-group longitudinal CDMs and compare these parameters across proficiency levels.

9. Comparing Dependency-Based Collocation Indices and LLM-Derived Surprisals and Semantic Similarity Measures in Predicting L2 Writing Scores

Authors: Rui Zheng (Zhejiang University), Hang Li (Zhejiang University)

Collocation use remains challenging for L2 writers, who often produce non-target-like combinations. Traditional frequency-based indices (e.g., MI) quantify association strength to capture "target-likeness" but fail to reflect contextual appropriateness. Recent advances in Large Language Models (LLMs) enable a more nuanced evaluation of collocation quality through two distinct computational dimensions: surprisal and semantic similarity. Surprisal, derived from the negative log probability of a word given its context, serves as a measure of processing load and expectancy. Meanwhile, semantic similarity leverages context-dependent semantic representations to assess whether a word choice constitutes a contextually appropriate selection within its specific semantic environment.

This study compares corpus-based dependency collocation indices with LLM-derived measures of surprisal and semantic similarity to explore their predictive power regarding human-rated holistic scores. Using 578 argumentative essays from the ICNALE corpus, three dependency relations (adjective–noun, verb–adverb, and verb–direct object) were extracted. Association strengths were computed via SpaCy, while surprisal and semantic similarity values were derived using BERT-based architectures. Hierarchical regression analyses will be conducted to examine the relative and joint predictive power of these indices, evaluating whether the inclusion of LLM-based processing and representation measures provides incremental validity beyond traditional corpus-based association strength.

By jointly considering target-likeness, processing load, and contextual appropriateness, the study aims to offer a more comprehensive account of how collocational use shapes L2 writing quality. The results are expected to advance quantitative modeling of collocation use and highlight the potential of combining corpus-linguistic and computational approaches for a more context-sensitive analysis of L2 learner writing.

10. Examining the Reliability of LLMs Scoring for Integrated Writing Tasks

Author: Yijia Luo (Zhejiang University)

Integrated writing tasks, such as TOEFL iBT, engage multiple subconstructs such as comprehension, organization, and source integration, which makes reliable scoring difficult for human raters, as raters may differ in judging strategies such as paraphrasing and integration. The recent advent of large language models (LLMs) such as GPT-5 offer new possibilities for automated evaluation. Although previous studies have shown promising consistency between LLM and human scorings in independent writing tasks, little is known about their reliability in complex integrated contexts.

This work-in-progress examines the reliability of GPT-5 scoring in integrated summary writing and whether its scoring patterns align with those of human raters. Approximately 200 Chinese EFL undergraduates will complete a TOEFL iBT integrated writing task. Human scoring follows the TOEFL rubric, with two raters and an adjudicator providing both holistic scores and rationales. GPT-5 will rate the essays under three prompt conditions: (1) rubric + sample essays, (2) rubric + chain-of-thought, and (3) rubric + sample essays + chain-of-thought.

Quantitative analyses will include consistency measures between GPT-assigned and human-assigned scores, and stepwise regression on 119 linguistic indices (covering text length, lexical sophistication, syntactic complexity, cohesion, and source integration). Qualitative thematic analysis will compare human raters' and GPT's scoring rationales.

This work-in-progress aims to provide evidence of the reliability of GPT-based scoring in integrated writing and to inform the practical feasibility and limitations of GPT-based evaluation in large-scale assessment contexts. Additionally, the prompt design offers insights into optimizing GPT prompting for integrated writing assessment.

11. To what extent do computer-based speaking performances relate to interactive performances in group oral discussions?

Authors: Yujia Zhou (Tokyo University of Foreign studies), Masashi Negishi (Tokyo University of Foreign studies), Asako Yoshitomi (Tokyo University of Foreign studies), James Carpenter (Tsurumi University)

Linking test performance to real-world outcomes (“extrapolation inference”) is essential for meaningful score interpretation (Chapelle, 2020). This inference is particularly critical for computer-delivered tests that rely solely on asynchronous communication, leading to potential construct underrepresentation of interactional competence (IC) (Ockey et al., 2015).

This study investigates the extrapolation inference for a computer-delivered speaking test for Japanese university admissions: the British Council–Tokyo University of Foreign Studies Speaking Test for Japanese Universities (BCT-S). It explores the extent to which BCT-S performance reflects IC measured in group discussions by examining (RQ1) associations between BCT-S scores and IC scores on a group oral test (GOT) and (RQ2) relationships between BCT-S scores and IC features in GOT utterances, and (RQ3) variation in IC among test-takers with the same BCT-S level.

Ninety-eight first-year Japanese university students took the BCT-S and a GOT including a convergent and free discussion task. The BCT-S provides a scaled score from 0 to 50; most participants were B1 or B2. Four raters assessed IC using analytic scales for turn and topic management, interactive listening, breakdown repair, and non-verbal behavior. Interactional features were coded using a scheme adapted from Galaczi (2014). Person measures for IC will be estimated with many-facet Rasch measurement. Pearson correlation and multiple regression analyses will be performed to explore RQ1 and RQ2. For RQ3, IC score distributions across BCT-S levels and selected GOT samples will be analyzed to illustrate interactional differences. Findings will inform BCT-S score interpretation and provide insight into the relationship between proficiency and IC.

12. Exploring Raters’ Cognitive Processes in Summary Content Evaluation: A Preliminary Think-Aloud Study

Authors: Yutaka Ishii (Chiba University), Yasuyo Sawaki (Waseda University), Yumi Fujita (Graduate School of Education, Waseda University)

Despite the importance of summary writing skills in the academic domain (Rosenfeld et al., 2001), the multifaceted construct of summarization often poses challenges in scoring and providing quality feedback to learners. Previous rating process studies have offered insights into rater behaviors in scoring integrated writing task responses (Gebriel & Plakans, 2014). Yet, relatively little work exists concerning the process of scoring source text content representation, a critical assessment criterion for summarization.

The present mixed-methods study explores the cognitive processes of raters to explicate their understanding of a content scoring rubric, developed locally for instruction and formative assessment purposes, and the ways in which they apply it to scoring summaries written by Japanese university EFL learners. Participants are six to eight experienced EFL writing instructors in Japan, who will attend a two-day training and scoring session. On Day 1, they will complete a training session to score student summaries using the four-point summary content rubric. They will then score 60 summaries independently. On Day 2, the raters will complete a think-aloud training session (Barkaoui, 2011; Lumley, 2002), score a new set of 20 summaries while thinking aloud, and participate in a follow-up semi-structured interview. Results from the coding of the think-aloud protocols and a thematic analysis of the interview data will be synthesized with many-facet Rasch analysis results to explore the relationship of rater behaviors to rater severity and consistency toward enhancing the transparency of the scoring rubric. The authors will present preliminary results for the discussion of potential analysis refinements.

13. Leveraging LLM to Identify Writing Strengths and Weaknesses Across Achievement Levels for Instructional Support

Authors: Maggie Dunlop (Education Quality and Accountability Office), Hyunah Kim (Education Quality and Accountability Office), Zhimei Gu (Education Quality and Accountability Office), Jennifer Hove (Education Quality and Accountability Office)

Large-scale assessments typically offer an overall score of ability that is evidence-based and robust. However, the information is generally not designed to address instructional needs. This creates tension for educators who often use instructional time to have students write assessments, but then receive limited instructionally relevant information in return. Our study addresses this tension by using large language models (LLM) to analyze student writing responses from an Ontario curriculum-based large-scale assessment. The study aims to develop methods for highlighting common strengths and weaknesses that explicitly reference the Ontario language curriculum, and thus meet Ontario educators' need for support in monitoring their students' progress against curriculum standards.

This study is situated in Ontario but is relevant to international efforts to make large-scale language assessments more instructionally useful. It contributes to the conversation about the role of artificial intelligence computing in language assessment by expanding the conversation toward instructional applications. We anticipate three key contributions: (1) a new, accessible method for generating empirically grounded profiles of student writing strengths and weaknesses, (2) new insight into effective feedback communication practices to Ontario teachers, and (3) an analysis framework to support future AWS validation and fairness research.

14. Aligning the content and scores of a large-scale K-12 test to an international framework

Authors: Meg Malone¹, Camelot Marshall¹, Reuben Vyn¹, Abbie Finnegan¹, Carly Favero¹, Aliza Price¹

Organization: 1: ACTFL

To meet growing international demands for test score comparability across proficiency frameworks while ensuring its test content is representative of the diverse pedagogical contexts and curricula of its users, ACTFL has launched a standard setting initiative to align its K–12 world language assessment, the AAPPL (ACTFL Assessment of Performance toward Proficiency in Languages), with the CEFR (Common European Framework of Reference for Languages). This work-in-progress presentation will share the methods and early findings of this empirical alignment process.

The methodology includes comprehensive rater orientation to the AAPPL constructs and content, as well as the CEFR scale, followed by separate mapping activities in the receptive and productive skills. For the interpretive content mapping, a cohort of trained experts evaluated the listening passages and reading texts of full AAPPL test forms, assigning CEFR levels to each item and reaching multi-rater consensus through evidence-based discussion. For productive skills, the same cohort evaluated more than 80 double-rate test-taker responses in presentational writing and interpersonal listening and speaking sections of the test. Raters will assign a CEFR level to the evidence in each complete sample, again working toward multi-rater consensus.

This presentation will reflect on the challenges and potential of cross-framework alignment, including implications for score interpretation, teaching and learning expectations, and curriculum development, particularly in bilingual and international education contexts.

15. Contextualizing AI Ethics for Local Language Assessment: A Mixed-Methods Study of University Policies and Stakeholders' Ethical Awareness in China

Authors: Qi Wang (Shanghai Jiao Tong University), Yan Jin (Shanghai Jiao Tong University)

The advancement of Artificial Intelligence (AI) in education brings potential benefits and ethical implications. Global initiatives have established ethical AI documents, yet AI ethics for language assessment remain scarce (Galaczi & Pastorino-Campos, 2025). Following DeepSeek's launch in China (2025), many Chinese universities issued AI policies that vary widely in attitude and scope, highlighting the need for research on local language assessment that requires context-specific guidance due to their distinct characteristics in the AI era (Yan, 2025).

This study analyzes university AI policies for language assessment in Chinese mainland universities, explores stakeholder perceptions, and develops actionable ethical guidance for local language assessment. Three research questions are proposed: (1) How do universities address ethical considerations of AI use in language assessment within their institutional policies? (2) How do key stakeholders—learners, educators, and administrators—interpret and respond to assessment-related AI policies? (3) What context-sensitive factors should inform a framework for ethical AI use in local language assessment?

A three-phase exploratory sequential design is adopted. Phase I, applying mixed methods combining a student survey across 12 universities (n=760) and teacher interviews (n=10), revealed broad support for AI integration but limited policy awareness, with perceptions varying across institution and discipline. Phase II (ongoing) focuses on language assessment ethics, through institutional assessment-related policies analysis, stakeholder survey and interviews. Phase III (ongoing) investigates context-sensitive factors for an ethical AI framework in local language assessment.

This study remains grounded in authentic educational practice, ultimately contributing a context-sensitive ethical framework to guide AI use in local university language assessment.

16. Examining L2 Learners' Cognitive Processing During GenAI-Mediated Multimodal Speaking Tasks

Authors: Sha Liu (British Council), Richard Spiby (British Council), Carolyn Westbrook (British Council)

Despite exponential growth in AI-assisted speaking assessment research, the real-time cognitive mechanisms underlying learners' interaction with generative AI during multimodal speaking tasks remain unexplored. Without understanding learners' moment-to-moment cognitive allocation patterns, optimal GenAI-mediated speaking task design cannot be established.

Drawing upon Cognitive Load Theory (Sweller, 1988; Sweller et al., 2011) and Multimedia Learning Theory (Mayer, 2005, 2024), this mixed-methods study examines how L2 learners across proficiency levels (CEFR B1-C1) process different multimodal configurations during GenAI-mediated speaking tasks. Thirty participants representing diverse L1 backgrounds will complete interactive roleplay tasks across three configurations: text with static visuals, text with audio, and text with video. Triangulating eye-tracking, stimulated recall with gaze replay, and cognitive load scales, we investigate whether and when cognitive overload occurs, its type, and causes.

Findings will inform the design of cognitively accessible AI-mediated assessments by revealing how processing patterns and cognitive load thresholds vary across proficiency levels and L1 backgrounds.

17. Glocalizing Language Assessment: Designing an Integrated Performance Assessment for Montreal's Sociolinguistic Reality

Author: Fanny Mace (Concordia University)

In today's world of increasing mobility, learners need not only grammatical accuracy but the pragmatic and sociolinguistic agility to succeed across diverse environments (Bachman & Adrian, 2022). In Montreal, new provincial legislation (Bill 96; Loi 14) requires newcomers to achieve a B2 level in French (or Échelle Québécoise Level 8), assessed through global European-based tests such as TCF(Q), TEF(AQ), and DELF. Yet, these tests often fail to capture the localized realities of Quebec French and the North American workplace pragmatics essential for successful social and professional integration.

Drawing on sociocultural and validity theories (DeLuca et al., 2019), I propose a Montreal-based Integrated Performance Assessment (IPA; called ÉPI in French or Évaluation de Performance Intégrée). This localized test would integrate linguistic, sociolinguistic, pragmatics, and transferable professional skills through authentic scenarios, such as team meetings, customer service interactions, and community participation, reflecting genuine Montreal language practices. The goal is to try and bridge the disconnection between test performance and professional success, promoting assessment as a tool for equity and belonging rather than exclusion (Shohamy, 2001). This proposal invites dialogue on how glocal frameworks can harmonize universal scales with local sociolinguistic specificities, ensuring that assessment serves both rigor and relevance.

18. Developing a Partial Interpretation Argument for a Learning-Oriented, AI-Enhanced, Scenario-Based Language Assessment for Business Case Study

Author: Daniel Eskin (Teachers College Columbia University)

Second and foreign language (SFL) speakers in most professional and academic contexts need adequate SFL proficiency and content knowledge to use language effectively (Douglas, 2000; Llosa, 2017). What sets academic contexts apart, however, is how speakers use language to learn content. In business education, an important disciplinary practice involving language use and content learning is the 'Case Study Method', in which students read a narrative about a problem illustrating a disciplinary concept then learn about the concept further through case analysis in reports and discussion (Basturkmen, 1999; Ellet, 2018; Nathan, 2016). Despite their pedagogical importance, few Business English exams use case studies (e.g., O'Sullivan, 2006) and no exam has incorporated their learning-oriented purpose. To address this gap, an AI-Enhanced Scenario-Based Language Assessment (SBLA) for Business Case Study was developed to measure examinees' reading, writing, and speaking abilities and their learning of disciplinary concepts from a Project Management case (Kerzner, 2017). The SBLA simulates a seminar context to investigate examinees' performance and the role of assistance via AI agents in moderating performance (Banerjee, 2019; Purpura, 2021). To guide the validation of the assessment, this work-in-progress outlines an Interpretation Argument (i.e., claims, warrants, assumptions) for the first two inferences in the Argument-Based Validation framework (Chapelle, 2020; Kane, 2021). For Domain Definition, assumptions regarding test development procedures are proposed based on Evidence-Centered Design (Mislevy & Yin, 2013) and the Learning-Oriented Assessment framework (Purpura, 2024). For the Evaluation Inference, assumptions pertaining to test administration conditions are proposed, focusing the technology use during the assessment.

19. The Alignment Between Language Frameworks and Curriculum Standards: The Case of English Speaking Standards in China

Authors: Shuai Zou (Shanghai Jiao Tong University), Yan Jin (Shanghai Jiao Tong University)

Language frameworks such as the CEFR and ACTFL are widely used as reference benchmarks, yet limited efforts have been made to align local curricula with these frameworks (British Council et al., 2022). The lack of communication between the fields of assessment and education has led to longstanding divisions regarding language ability constructs and developmental progressions. A similar issue pertains to China's Standards of English Language Ability (CSE). The CSE was released in 2018 as China's national framework, while a series of English curriculum standards have been regulating curricula and assessments since the 2000s. Both systems are currently in effect, each with its own descriptions of language proficiency, which creates confusion among learners and educators.

This study aims to examine the alignment between the CSE and curriculum standards. Specifically, this study investigates the English speaking scales at high-school and undergraduate levels and follows a multi-stage design: 1) a scoping review of 47 studies; 2) construct alignment regarding construct definitions, descriptive schemes and descriptor structures; 3) level alignment calibrating descriptor difficulty through Comparative Judgment tasks, where 20 judges make pairwise comparisons to scale descriptors; 4) claim validation through procedural, internal and external validity evidence. The findings are expected to provide empirical evidence on the alignment between standards and frameworks, as well as identifying gaps, overlaps or inconsistencies in the requirements. Overall, this study could provide insights into the coherence of China's English language education, while also supporting global score recognition and fostering synergies across education systems.

20. Customising GPT for learning-oriented test preparation in second language writing

Authors: Carrie Xin Peng (University of Melbourne), Xingcheng Wang (University of Melbourne)

This study aims to synergise with the accessible and personalised nature of ChatGPT and maximise L2 writers' learning opportunities as they engage in L2 writing test preparation. We have developed a customised GPT model named "WriteWithMe" that offers scaffolded support for L2 writers of varying proficiency levels, especially during pre-task planning and post-task revision stages. WriteWithMe is configured not to generate complete essay outlines or essay drafts upon request from users but to offer probing questions and step-by-step guidance to support L2 writers' planning and revision activities, tailored to their proficiency levels. This model hence aims to assist L2 writers' learning-oriented writing test preparation by providing tailored and structured guidance on writing planning and revision, grounded in the sociocultural theory of learning (Vygotsky, 1978).

This works-in-progress talk will present preliminary findings on a pilot study comparing the effects of WriteWithMe with the standard ChatGPT on L2 writers' argumentative writing pre-task planning. This presentation will also share the customised GPT model with the audience and seek advice on further adjusting its configuration. Research findings from this study will inform possible integration of GenAI-powered chatbots into L2 writing test preparation and instruction in general, as well as offer methodological implications for future research on customising GPT for specific language teaching, learning, and assessment purposes.

22. Needs and Content Analysis for the Brigham Young University (BYU) English Language Admissions Test: Listening and Speaking

Author: Shantona Mir (Brigham Young University)

This project examines how a needs and content analysis can inform the creation of a Brigham Young University (BYU) specific English admissions test for listening and speaking. Standardized English proficiency tests like TOEFL, IELTS, and Duolingo are designed for a global audience, but their generalized nature often fails to align with the linguistic realities of specific institutions. BYU offers a unique case because of its religious mission and academic culture, which involve exposure to U.S. faculty and mainstream American English in lectures and classroom interactions.

The study emphasizes local vs. general testing, recognizing that a single test suitable for everyone is rarely ideal for any one institution. Additionally, BYU's verification processes may offer more accessible pathways for students who face barriers with ID-based testing.

Using a mixed-methods design, the study will analyze course syllabi, recorded lectures, and classroom activity descriptions to identify target language demands, while surveys and interviews with NNES students will reveal present abilities and challenges. This dual analysis will clarify what specific listening and speaking skills students need to succeed academically at BYU. The results will inform the design of institution-specific test content and tasks that more accurately reflect BYU's linguistic expectations. This research aims to advance conversations on localized testing, offering a model for more valid, equitable, and mission-aligned language assessments.

23. How Do Interactional Patterns Differ in Human–AI and Human–Human Speaking Tasks? An Alignment Perspective

Authors: Yang Zhang (Zhejiang University), Lin Shi (Zhejiang University), Shuting Zhang (Zhejiang University)

The rapid integration of AI interlocutors into language assessment raises critical questions about how interactional competence is elicited and interpreted. Grounded in the Interactive Alignment Model (Pickering & Garrod, 2004, 2021), this study examines whether alignment—the dynamic adaptation between interlocutors across linguistic and affective levels—differs in human–AI and human–human speaking tasks. Alignment is conceptualized as a process underpinning mutual understanding and engagement, offering a promising psycholinguistic indicator of communicative ability and test authenticity.

Using a within-subject design, 20 EFL learners (target $n = 60$) completed three components: a baseline oral task, a human–human discussion, and an equivalent human–AI conversation with a self-built system. All sessions were transcribed for computational analysis. Alignment was measured through Latent Semantic Similarity (LSS) for semantic overlap, Language Style Matching (LSM) for stylistic convergence, and Emotional Alignment (EA) for affective convergence using VADER sentiment analysis. Two trained raters assessed fluency, accuracy, complexity, intelligibility, and interactional engagement, with baseline proficiency included as a covariate.

Preliminary analyses reveal no significant differences across task types, suggesting that AI interlocutors can elicit interactional patterns comparable to human peers. By linking psycholinguistic theory with test validation, this ongoing study offers early empirical insight into how alignment operates in AI-mediated oral tasks, informing the design, scoring, and construct interpretation of next-generation speaking assessments.

Posters Summaries

Day: Saturday, June 6 (3:20 p.m.-4:20 p.m.)

Location: Opera C

1. Localizing the “global” CEFR framework for a rating scale in a high-stakes exam in Cuba

Authors: Claudia Harsch (University of Bremen), Yoan Martinez Márquez (Universidad de las Ciencias Informáticas Habana)

In Cuba, the Ministry of Higher Education (MES) introduced the Common European Framework of Reference (CEFR) as a globally accepted proficiency framework, setting level B1 as exit standard for BA studies. To evaluate the proposed exit standard, the MES wanted to introduce a certification system to foster and assess students' English proficiency.

We report on the design and development of the high-stakes exam that is posited at the end of said certification system, with a specific focus on the development and validation of the rating scale for speaking assessment. The scale development underwent several stages. Starting with a descriptor-based approach, we used the CEFR as basis and adapted its categories and descriptors to suit the local context. We also took other CEFR-based scales and the local curriculum into account. Next, we added an intuitive approach, conducting a sorting exercise with local trained teachers to confirm criteria and levels, and to revise descriptors where necessary. Finally, we used an empirical approach during rater training and benchmarking sessions, which served to calibrate local interpretations of the CEFR. We present rating data from the different benchmarking sessions to illustrate the growing rater agreement.

For each of these stages, we summarize the main challenges and tensions that arose from localizing the CEFR categories and descriptors in order to adapt them for the localized speaking rating scale; we will present our approaches of dealing with these challenges. We end our presentation with exploring wider implications for adapting a global framework for local assessment contexts.

2. Unpacking the Role of Visual Aids in L2 Integrated Writing: A Dual-Focus Investigation of Written Products and Composing Processes

Author: Jiayi Dong (Zhejiang University)

Visual aids, particularly videos, are ubiquitous in L2 teaching and learning but underutilized in L2 assessment. Though extensive research has revealed facilitative effects of visual aids on enhancing L2 listening comprehension, very few studies have examined the impact of visual aids in L2 writing assessments. Among these studies, visual aids are predominately contextual, and their effects are typically measured through score comparisons, leaving two critical gaps: the combined effect of context and content visuals on test scores and textual features, and test-takers' perception, comprehension, and strategic utilization of these visuals.

To address these gaps, this study systematically investigated how visual aids affect test-takers' writing performance and process in L2 integrated writing tasks. In a counterbalanced within-subjects design, 51 tertiary-level EFL learner completed two tasks: a reading-listening-writing (RLW) task with audio paired with a still image and a reading-viewing-writing (RVW) task with video displaying a lecturer and PowerPoint slides. Qualitative and quantitative analyses were conducted on writing products (essays and notes) and process data (questionnaires and interviews). Results revealed test-takers' superior performance in the RVW task for content coverage, organization, and source integration, with the RVW task eliciting greater syntactic complexity, text fluency, and cohesion. In contrast, the RLW task manifested higher lexical complexity. Additionally, test-takers' strategic use of visual aids was significantly more frequent in the RVW task, and the relationship between writing strategies and scores varied across tasks. Notably, most learners preferred the RVW task for its role in enhancing listening comprehension, supporting writing production, and increasing task engagement.

3. AI-Assisted Language Assessment in Post-ChatGPT Era: A Scoping Review (2023–2025)

Authors: Jieun Kim (Soongsil University), Haeun Kim (The University of Melbourne), Hitoshi Nishizawa (Reitaku University)

The advancement of artificial intelligence (AI) has introduced both opportunities and challenges for language assessment. While recent reviews have explored AI in language learning more broadly (e.g., Law, 2024; Zhu et al., 2025), no systematic synthesis has focused specifically on AI-assisted assessment in second language (L2) contexts. This study presents a scoping review of research published between 2023 and 2025, designed to identify patterns, gaps, and implications for both global and local assessment practices.

Following PRISMA guidelines (PRISMA, 2020), comprehensive searches were conducted in ERIC, Web of Science, Scopus, and LLBA, yielding 2,314 entries. Studies were included if they were peer-reviewed, empirical in nature, focused on L2 learners, and explicitly framed as assessment. Exclusion criteria eliminated commentaries, theoretical articles, dissertations, book chapters, and conference proceedings. After multiple rounds of coding and calibration across three researchers, 89 studies were retained for final analysis.

The coding framework captured contextual and methodological variables, including the type of AI technology employed, the language skills targeted, assessment type and purpose, educational context, scale, stakes, reasons for AI adoption, and validity considerations.

Preliminary findings indicate that writing remains the most frequently assessed skill, particularly in automated writing evaluation, though interest in AI-mediated speaking assessment and interactional competence is emerging. ChatGPT is by far the most widely employed tool, primarily for scoring, though studies are increasingly investigating its use in item generation. This review consolidates current research, identifies key gaps, and points to future directions for valid and contextually appropriate AI use in L2 assessment.

4. Towards Context-Specific Measures: Developing an ESAP Test for Business Majors in EMI Programmes

Authors: Peng Wang (University of Bath), Daniel Yu-sheng Chang (University of Bristol)

The rapid expansion of English-medium instruction (EMI) in non-anglophone universities has intensified the need for assessments that reflect discipline-specific demands of academic English (McKinley, 2024). Yet most EMI studies rely on general proficiency measures from standardised tests or self-reports. Such measures are problematic because EMI requires competence in both disciplinary content and English. Proficiency assessment should therefore capture learners' ability to participate in field-specific communicative practices.

To address this gap and respond to calls for English for Specific Academic Purposes (ESAP) (Galloway et al., 2024), this study developed and validated an ESAP test for postgraduate business students in China. Guided by Weir's (2005) socio-cognitive model, the computer-based test assessed integrated skills across four equally weighted tasks: three listening–reading–writing tasks and one listening–reading–speaking task. For instance, in the writing task, students produced a strategy proposal based on a stock market scenario and meeting brief; in the speaking task, they critiqued a case study after engaging with a lecture on supply chain management.

The test and rating criteria were reviewed by subject and language specialists and piloted with 62 postgraduate students. Two instructors applied the rating scales, and 15 students were interviewed. Many-Facet Rasch Measurement (MFRM) showed acceptable rater fit, inter-rater agreement, and student ability spread. Writing tasks displayed comparable difficulty, whereas the speaking task was more challenging, reflecting students' limited oral practice in EMI contexts. This study advances EMI assessment by offering a valid ESAP measure.

5. A Comparative Study of the Scoring Quality between DeepSeek and Human Raters in Chinese as a Second Language Writing Assessment

Authors: Fuyu Kong (Shanghai International Studies University), Haifeng Qi (Shanghai International Studies University)

This study explores the scoring performance of DeepSeek, an AI language model, in assessing Chinese as a Second Language (CSL) writing, and compares its scoring results with those of novice and experienced human raters. The research aims to evaluate whether DeepSeek can provide reliable, valid, and human-comparable scoring in CSL writing assessment.

The dataset consists of 84 handwritten essays written by intermediate-to-advanced CSL learners. All essays were evaluated based on a four-dimensional analytic rubric adapted from the HSK writing criteria, covering content, vocabulary, grammar, and structure. Six human raters participated, divided into experienced teachers (over 15 years of CSL teaching experience) and novice teachers (one year of experience). DeepSeek was trained using the same rubric and evaluated the essays twice to ensure scoring reliability.

Results show that DeepSeek's total scores were moderately correlated with key linguistic indices such as lexical accuracy and diversity, suggesting that the AI model can capture essential aspects of writing quality. When compared with human raters, DeepSeek's overall scores were highly consistent with those of novice teachers but significantly higher than those of experienced teachers, who tended to assign lower and stricter scores.

Although all raters demonstrated a similar evaluation pattern—in which structure received the highest scores, followed by content, and then vocabulary and grammar—significant differences were found in the grammar dimension. DeepSeek and novice teachers consistently rated grammar higher than experienced teachers, indicating variation in how grammatical quality is perceived across rater types.

Overall, the findings indicate that DeepSeek can provide stable and rubric-aligned assessments comparable to novice raters, but further calibration is needed to achieve alignment with the more rigorous standards applied by experienced human raters.

6. Gaps between global and local policy: Implications for English learner identification

Authors: Ahyoung Alicia Kim (WIDA, Univ. of Wisconsin-Madison), Mark Chapman (WIDA, Univ. of Wisconsin-Madison), Haeun Kim (University of Melbourne), Jeanne Beck (ETS), Pauline Ho (Reed University)

In the United States, state education agencies and their school districts have a legal obligation of providing appropriate language support services to Kindergarten to Grade 12 English learners (ELs). According to Every Student Succeed Act (2015), districts are responsible for accurately identifying ELs in a timely manner (U.S. Department of Education, 2016). The federal government suggests using a home language survey and other tools, such as English language proficiency screeners, for this purpose. One commonly used tool for Kindergarten EL identification is the WIDA Screener for Kindergarten (hereafter K Screener), which was released in 2021. By referring to the language policy framework by Johnson and Johnson (2014), we describe EL identification policy and implementation of the K Screener at the federal/state (global) and school district (local) levels.

The findings are from a two-phase, mixed-methods study, which examined how the K Screener was implemented in one US state. In Phase I, 133 educators (from 118 districts) completed an online survey exploring EL identification practices. In Phase II, a case study approach was used to explore the EL identification practices of three different districts (with small, medium, and one large EL enrollment). Educator survey and interview data, and students' home language survey responses, and screener scores were analyzed. We discuss identification of Kindergarten ELs by exploring the decision makers, criteria, and procedures with specific attention to the K Screener. Findings indicate variation in EL identification across districts, suggesting gap between state-level EL identification policy and actual identification practices in districts.

7. Global Standards, Local Profiles: Implementing the Updated Vocabulary Level Test for Diagnostic Assessment of Thai Graduate Students

Authors: Sumanee Pinweha (Chulalongkorn University), Sutthirak Sapsirin (Chulalongkorn University)

This study connects global standards in academic English to local contexts by conducting a diagnostic assessment of the vocabulary levels of 217 Thai graduate students. Using the Updated Vocabulary Levels Test (UVLT), the research develops detailed local profiles of receptive vocabulary knowledge across the first 5,000 word families, comparing students in science and non-science disciplines. The findings reveal a significant gap between students' current vocabulary profiles and the global standards needed for academic reading, with average scores falling below the 80% mastery threshold. Additionally, science majors significantly outperformed their non-science peers at several word levels ($p < .05$), highlighting important disciplinary differences. This presentation emphasizes the value of using a psychometrically robust test, such as the UVLT, to generate reliable local data. These evidence-based profiles enable Thai higher education institutions to move beyond generic EAP instruction and create targeted, discipline-specific curricula that better prepare students for participation in the global academic environment.

8. Formative Assessment in Practice: Exploring How Chinese Teachers Support Learning in Elementary ESL Classroom

Authors: Liying Cheng (City University of Macau), Peiyu Wang (Queen's University), Liyan Huang (City University of Macau)

Despite national educational policy reforms promoting learning-centered assessment in China, the enactment of formative assessment (FA) in elementary classrooms remains uneven, constrained by public examination pressures, limited professional training in FA, and societal beliefs in traditional assessment. Recent research highlights the need for large-scale, empirically grounded investigations to better understand how teachers and students co-construct FA to enhance interaction, feedback, and learner participation across contexts. Addressing this gap requires a coherent framework that captures the dynamic nature of FA as it unfolds in classroom practice. Building on these insights, this study applies the Integrated Model of Assessment for Learning Strategies and Use of Feedback (DeLuca et al., 2018) to conceptualize FA as an interactive, evidence-informed process involving learners, peers, and teachers. Using a large-scale teacher questionnaire ($N = 673$) grounded in DeLuca's framework, the study examined elementary teachers' self-reported FA practices through descriptive statistics, EFA, and SEM to validate the underlying structure of FA and explore relationships among its dimensions and teacher variables (e.g., training, literacy, and professional training in FA). The analyses confirmed a robust three-factor model—(1) clarifying and sharing learning goals, (2) providing feedback to advance learning, and (3) engaging students in self-assessment and progress monitoring—with high reliability and excellent model fit. Teachers reported frequent use of goal sharing and feedback but less consistent implementation of student self-assessment; professional training and administrative dialogue significantly predicted overall FA use. The study advances understanding of FA as a multidimensional, integrated process linking teachers' pedagogical intentions with classroom practices.

9. Commonalities and Localized Differences in How Language Testers Teach Language Assessment around the World

Authors: Benjamin Kremmel (University of Innsbruck), Paula Winke (Michigan State University)

Recent developments in the field of language testing emphasize greater inclusivity, transparency, and ethical practices. This includes promoting open-access research and making evidence for test score interpretations publicly available. Concurrently, scholars have focused on defining language assessment literacy (LAL)—the essential knowledge and skills for designing and evaluating language assessments—which has helped define key pedagogical components for language testing instruction. Despite these advancements, there remains a gap in research regarding the teaching practices of language testing itself, particularly the procedural and contextual aspects of how LAL, ethics, and test design are taught. To address this, we collected over 30 syllabi and detailed course descriptions from selected instructors worldwide to foster openness and collaboration in syllabus design and pedagogy. These will be published in an Open-Access book by “Applied Linguistics Press.” The final chapter of the book is a synthesis of the chapter contents, which we summarize, preliminarily, here in our poster.

On our poster we explore two main questions: (1) What are the shared and different approaches to teaching language testing and assessment globally? (2) How do factors like instructional level, curriculum structure, and local contexts influence these approaches?

The resulting synthesis highlights core LAL components, effective teaching strategies, and the importance of localizing instruction. We also showcase regional efforts to integrate ethical considerations and disabilities accommodations into language testing and assessment pedagogy.

10. Agile for Equity: Leveraging Agile Project Management Methodologies to Advance Fairness in Local Language Assessment

Author: Sharareh Taghizadeh Vahed (Center for Applied Linguistics)

Fairness is a core principle in language assessment, ensuring that tests produce valid, reliable, and unbiased results across diverse populations (Kunnan, 2014). However, promoting fairness is not solely a psychometric concern, but also a project management challenge. This poster will demonstrate how Agile project management approaches can actively support fairness in local language assessment by enhancing stakeholder engagement, increasing flexibility, and enabling rapid iteration based on feedback in every stage of test development.

This poster presentation draws on real-world examples from large-scale English language proficiency assessments to illustrate how Agile practices can contribute to fairness of local tests in three key ways:

1. **Stakeholder Collaboration:** Agile’s emphasis on iterative planning and sprint reviews creates more opportunities to include perspectives from educators, linguists, accessibility experts, and test-takers. This input helps ensure cultural and linguistic relevance and reduces the risk of construct-irrelevant variance.
2. **Responsiveness to Fairness Concerns:** Agile allows teams to respond quickly to concerns that may impact fairness, such as ambiguous item content, access barriers, or potential sources of bias related to cultural or linguistic representation. Items flagged as problematic can be revised or deprioritized within the same sprint cycle, rather than delayed until end-of-cycle reviews.
3. **Accountability and Transparency:** With Agile’s progress tracking tools (e.g., burndown charts, sprint goals), teams can monitor task ownership and timelines to ensure fairness-related issues are prioritized and documented. Agile artifacts, like user stories also make equity-oriented goals more explicit and measurable.

12. Trends, Methods, and Practices in Rater Training: A Scoping Review of Rater Training Research in L2 Performance Testing

Authors: Ryuhei Mizoguchi (Waseda University), Chiho Young-Johnson (Georgia State University), Yumi Fujita (Waseda University), Dylan Burton (Georgia State University)

This scoping review aims to provide a comprehensive overview of L2 rater training research by analyzing (1) publication trends, including contextual distributions, participant characteristics, and test settings, (2) methodological features, such as research designs, data collection methods, and analytical techniques, and (3) rater training procedures, including training methods, modalities, and bias reduction strategies. The review process followed the PRISMA-ScR checklist (Tricco et al., 2018), resulting in 28 studies analyzed between 1994 and 2024.

The findings revealed a wide range of trends and tendencies in rater training research. The studies covered a range of assessment contexts, from low-stakes classroom practices to high-stakes proficiency testing, with a balanced focus on rating speaking and writing. The scope of this research has gradually expanded from a focus on rater expertise to the effectiveness of various training types (e.g., online vs. face-to-face, feedback type). While online training has become more common, few of these studies incorporated discussions during feedback sessions compared to face-to-face formats. Methodologically, sixteen studies employed mixed methods and twelve quantitative, with all analyzing rating data supplemented by questionnaires, interviews, or verbal protocols. In quantitative analyses, more than half used many-facet Rasch measurement, alongside a variety of other statistical approaches, to detect rater effects. The findings of this review will benefit researchers and practitioners in language testing by identifying trends in the literature and highlighting the variety of methods in rater training, thereby offering guidance for the construction of future research projects and quality rater training programs for operational testing.

13. Listening and Reading to Write: Corpus-Based Evidence on Task Type and Proficiency in Portuguese as an Additional Language

Authors: Elisa Marchioro Stumpf (Federal University of Rio Grande do Sul), Juliana Roquele Schoffen (Federal University of Rio Grande do Sul), Isadora Dahmer Hanauer (Federal University of Rio Grande do Sul), Luiza Sarmento Divino (Federal University of Rio Grande do Sul)

Celpe-Bras, the official Brazilian certificate of Portuguese as an Additional Language (PAL), assesses proficiency through integrated tasks that simulate real communicative demands. Despite its relevance for PAL, few empirical studies have examined how candidates at different proficiency levels mobilize linguistic resources from input materials in reading-to-write and listening-to-write tasks (Mendel, 2019; Pileggi, 2015; Vicentini, 2022).

Corpus Linguistics offers a robust framework to investigate these processes by providing observable linguistic evidence related to construct representation and performance differences (Callies & Götz, 2015; Cushing, 2022).

This study compares lexical patterns in two corpora of written responses to Celpe-Bras tasks: one reading-to-write (CorCel2015-2T4 - 2,293 texts) and one listening-to-write (CorCel2016-2T1 - 521 texts), both rated on a 0–5 scale. Texts were compiled under controlled exam conditions (Schoffen et al., 2025) and analyses conducted with Sketch Engine (Kilgariff et al., 2004) included text length, type-token ratios, and n-gram and frequency wordlist extraction.

Results show that in both tasks, higher-level texts are longer and lexically richer, displaying more items distinct from the input material. In the reading-to-write task, lower-level texts contain proportionally longer copied sequences, consistent with previous findings (Plakans, 2008; Gebril & Plakans, 2013). In contrast, the listening-to-write task shows the reverse: higher-level test takers produce longer transcribed segments, suggesting effective comprehension and note-taking strategies (Cubilo & Winke, 2013).

By offering corpus-based evidence from an under-researched language assessment context, this study contributes to the validation of Celpe-Bras as an integrated-skills examination and advances the discussion of how task type mediates input integration and construct interpretation in proficiency assessment.

14. From Global Standards to Local Policy: Contextualizing GEPT and BESTEP in Taiwan's Language Assessment Landscape

Authors: Jessica Row Whei Wu¹, Rachel Yifen Wu¹, Anita Chun-Wen Lin¹, Anne Liu¹

Organization: 1: The Language Training & Testing Center

This paper explores the development of two major English proficiency assessments in Taiwan—the General English Proficiency Test (GEPT) and the BEST Test of English Proficiency (BESTEP)—as exemplars of how large-scale tests can balance global credibility with local responsiveness. Anchored in the Socio-Cognitive Framework (Weir, 2005) and informed by debates on contextualization (Saville, 2012; Su, Weir & Wu, 2020), the study compares GEPT and BESTEP across key dimensions of purpose, design, validation, technology, and stakeholder engagement.

The analysis reveals how GEPT was shaped by Taiwan's early policy focus on lifelong learning, aiming at broad certification across age groups, while BESTEP was designed in response to the Bilingual 2030 and EMI policy, targeting tertiary-level learners with digital delivery and AI-supported feedback. Despite their differences, both assessments share a commitment to CEFR alignment and long-term validation, demonstrating that global standards and local needs can be integrated rather than opposed.

This comparative study contributes to LTRC 2026's theme of Global to Local Language Assessment by illustrating a model of "transcontextual adaptation," where test design mediates between international frameworks and local reforms. It offers practical and theoretical insights into how assessments can evolve to serve multiple constituencies, functioning not only as measures of proficiency but also as partners in learning and policy implementation.

15. AI literacy as part of student teachers' Language Assessment Literacy - a university teacher students' perspective on the use of generative AI tools for writing assessments in EMI contexts

Authors: Karin Vogt (University of Education Heidelberg), Alexander Mayer (University of Education Heidelberg)

Generative AI is rapidly reshaping academic writing and assessment in English-Medium Instruction contexts. This study positions AI literacy as an integral component of Language Assessment Literacy (LAL) for student teachers and investigates how university teacher students use, perceive, and expect training around generative AI in writing assessments.

The mixed-methods study examines generative AI use in academic writing among university teacher students in English-Medium Instruction (EMI) contexts. Drawing on an integrated LAL framework that integrates AI literacy, the study investigates patterns of disclosed and undisclosed genAI use, perceived affordances and challenges for of genAI use for assessed essays as take-home exams, effects on critical thinking, and expectations for professional development. A convergent design combined a questionnaire (n = 128) with six follow-up semi-structured interviews drawn from the survey sample. Survey data were analysed descriptively and with cross-tabulations, correlations, and Chi-square tests. Interview data underwent deductive and inductive content analysis with subsequent triangulation of findings. Results show that genAI is already integral to students' writing processes; more frequent users demonstrate stronger prompt literacy and report perceived improvements in drafting and language quality. Concurrently, respondents reveal a pronounced deficit in formal AI literacy, experience ethical uncertainty around disclosure and academic integrity, and report ambivalent effects on critical thinking. Participants express a clear demand for structured, curriculum-embedded AI literacy within teacher education. The study advances theory by operationalising AI literacy within LAL and offers practical implications for regulatory policies and assessment-aligned AI literacy curricula as part of LAL professional development for student teachers.

16. Local practices for developing language assessment literacy (LAL) in Brazil: a comparison between LAL courses in three public universities

Authors: Isadora Teixeira Moraes (Universidade Federal de Minas Gerais), Estogildo Gledson Batista (Universidade Estadual de Maringá), Renan William Silva de Deus (Universidade Estadual de Londrina)

Language assessment courses are often scarce in English teacher education programs in Brazil and elsewhere, which highlights the importance of language assessment literacy (LAL) worldwide (Gan & Lam, 2022; Giraldo & Yan, 2025). Therefore, this study investigated, through a questionnaire, pre-service teachers' (PSTs) perceptions of tasks carried out in three LAL courses at three public universities in Brazil. Across institutions, the most highly valued tasks involved integrating theory with practice, especially the development of assessment tools and rubrics. In contrast, peer- and self-assessment were consistently ranked lowest, as PSTs felt unprepared or uncomfortable conducting them. This result echoes previous research conducted in Brazil (Moraes et al., 2024). Divergent findings also emerged: in Institution C, a poster task on assessing the four skills was highly valued but was ranked lowest in Institution A. Additionally, in Institution C, PSTs developed an end-of-the-term paper, which was the task seen as the most conducive to learning. This aligns with findings in Brazil and abroad (Giraldo & Murcia, 2019; Moraes et al., 2024). Considering Fulcher's (2020) call for pedagogies of LAL, our results seem to suggest the need for a glocal approach (Kumaravadivelu, 2008) to such initiatives. This implies that local contexts and needs should be prioritized (Scarino, 2013) while maintaining global trends in mind. By comparing three initiatives, this research offers one of the few cross-institutional perspectives on LAL in Brazil. We hope this research contributes to broader debates on how language assessment is conceptualized and practiced worldwide.

17. The Language Assessment Accessibility Resource Centre (LaARC): Supporting Inclusive English Assessment in Latin America

Authors: Juliana Reichert Assunção Tonelli (Londrina State University), Gladys Quevedo Camargo (Universidade de Brasília)

This poster aims at presenting the Language Assessment Accessibility Resource Centre (LaARC). It consists of an online repository created with the support of the British Council (Assessment Research Grant 2023) (Quevedo-Camargo & Tonelli, in print) to assist English language teachers who work with both young and older learners with special educational needs and disabilities (SEND) in Latin America. Intended to promote equity, social inclusion, and the rights of all learners to quality education, inclusive education policies inspired by UNESCO (2020) have expanded the presence of SEND students in mainstream classrooms. Although essential, such policies present substantial pedagogical, infrastructural, and systemic challenges which call for improved language assessment literacy. Based on studies by Kormos & Kontra (2008), Kormos & Smith (2012), Kormos & Taylor (2021), Nijakowska, Tsagari & Guz (2023), Tsagari & Spanoudis (2013), Tsagari & Sperling (2017), among others, a mixed-methods study investigated the needs of 75 teachers from five Latin American countries revealed key challenges: limited resources, insufficient training, and difficulty adapting assessments. While teachers employed creative strategies, systemic barriers hindered equitable practices. In direct response to these findings, the LaARC repository was designed as a dynamic, multilingual, and teacher-centered free online resource hub. It contains 8 tabs: Home, Resource library, Professional learning, Research & Evidence, Community Hub, Accessibility & Inclusion, About the project, and Share.

The repository provides continuously updated, contextually relevant resources to help Latin American educators design accessible language assessments, thereby fostering greater inclusion in the English language classroom.

18. Automated Topic Detection in Spoken Language Assessment: Comparing Human-Labeled Training and Prompt-Based Category Prediction

Authors: Young-A Son (Language Testing International), Erik Voss (Teachers College, Columbia University), Kim Sallee (Language Testing International), Meg Malone (ACTFL)

Topic detection plays an essential role in automated language assessment because responses must be evaluated not only for linguistic quality but also for their topical relevance to the prompt. The ACTFL Oral Proficiency Interview–computer (OPIC) is a large-scale assessment used globally and also implemented within specific local contexts such as South Korea for business purposes. Automated scoring systems, which include topic detection models, must therefore account for the communicative realities of these contexts to ensure reliable evaluation. As language assessments increasingly use AI technology to support human rating, test takers may attempt to game automated scoring systems by reciting memorized or generic responses (Knill & Gales, 2024; Malinin et al., 2017); this approach can, artificially inflate ratings (Wang et al., 2019) and therefore threaten the test’s validity argument. Automated identification of off-topic responses is thus essential for determining ratings on spontaneous speech tasks.

This study explored the development of an off-topic detection model by comparing human-labeled training and prompt-based category prediction. Two models were developed using two large language models (LLMs) trained on a dataset of 1,000 human-labeled OPIC responses from 2020 to 2021. One model was trained on response transcriptions to identify topical alignment based on human-labeled data. The second model was trained to predict the topic category from the same response transcriptions based on the prompt category. Findings suggest that direct supervision with human-labeled data yields higher model performance than prompt-based category prediction.

19. Comparing Input Prompt Visuals Using Virtual Reality in L2 English Speaking Task

Authors: Yilin Zhang (Teachers College, Columbia University), Erik Voss (Teachers College, Columbia University)

Audio-visual media have long been adopted to simulate more authentic input in language assessment (Wagner, 2010). However, understanding how novel visual modalities can influence the assessment of second language speaking performance remains unexplored (Ockey et al., 2017). This study examines the potential of Virtual Reality (VR) as a novel input modality to enhance the authenticity of speaking assessment tasks. Employing the Meta Quest 2 Head-Mounted Display (HMD), participants ($n = 28$) who speak English as a second language engaged with speaking tasks presented in three randomized visual input modalities: still image, 2D video, and VR. Following a mixed method, within-subject design, participants’ spoken discourse was analyzed to assess the differences in their performance across modalities. Statistical analyses revealed that speaking performance varied across conditions. Although participants in the still image condition demonstrated higher syntactic and lexical complexity compared to those in the 2D video and VR conditions, variations were observed in lexical complexity between the 2D Video and VR conditions. Lexical bundle analysis showed that the VR condition produced the highest number of lexical bundles ($n = 96$). Functional analysis further indicated that attitudinal/modality stance bundles and place reference bundles were most frequent in the VR condition, suggesting that VR may have offered a more immersive environment that elicited language use similar to what participants would produce when physically situated in the test scenes. This study highlights that visual input does affect language performance should further research be needed on the use of VR to provide immersive and authentic input.

20. Integrating Collocations into Predicting L2 Oral Proficiency: Evidence from the TEM8 Speaking Test in China

Authors: Jinjie Wu (Shanghai International Studies University), Xini Liao (Shanghai International Studies University)

Collocations have been widely recognised as key to accurate and fluent speech (e.g., Tavakoli, 2025), yet their precise roles, including their fixedness and semantic accuracy in L2 speaking proficiency, remain unclear. The current study thus investigated the role of collocational profiles along with accuracy and fluency in predicting L2 oral proficiency. The collocations in this study were operationalised using two lexical combinations: adjective + noun and verb + noun.

The study employed a quantitative research design. 180 spontaneous speech samples from the TEM8 speaking test, a national, standardised achievement test in China, were collected from intermediate to advanced L2 speakers later dichotomised by their overall speaking scores. After being individually assessed by two experienced raters, the collected samples were processed for collocation extraction and measuring temporal fluency (i.e., speed, breakdown, and repair). Collocational profiles were operationalised by log-transformed raw frequency and mutual information (MI), and semantic accuracy were measured by an ordinal scale from 1-3 (Xu, 2015). Kruskal-Wallis test, Spearman's correlation and linear multiple regression were used for further analysis.

The results revealed that in contrast to frequency, MI scores differed significantly across proficiency groups. Additionally, speaking scores were moderately associated with speed and breakdown fluency. Regression analysis suggested collocational profiles, accuracy and fluency accounted for 45% of the variance in overall proficiency scores with fluency and collocational profiles as the most significant predictors. This study highlights the importance of collocations in predicting oral proficiency, with direct implications for integrating collocational criterion into rubric design and automated scoring.

21. Bridging global standards and local needs: a conceptual reflection from the Italian scenario-based assessment case study

Author: Giulia Peri (University for Foreigners of Siena), Sabrina Machetti (University for Foreigners of Siena)

Globally adopted assessment frameworks and large-scale standardized language testing systems often struggle to accommodate the diversity of learners and the contexts in which they communicate. High-stakes proficiency assessments have traditionally been designed to measure independent language skills; however, this approach can be limiting when capturing how L2 learners use language to perform complex, integrated tasks in authentic communicative settings (Purpura, 2021). Contemporary language use requires the ability to integrate knowledge, solve problems, and address the inseparability of content and language (Chalkiadaki, 2018; Purpura, 2021) in situated contexts. Balancing the need for consistency across contexts with the recognition of diversity thus becomes a central challenge in large-scale language assessment. Within this global-to-local tension, the CILS Centre of the University for Foreigners of Siena, operating within the CLIQ Association, bases its exam construct on the Common European Framework of Reference for Languages (CEFR; CoE, 2001, 2020). While the CEFR ensures shared standards of validity and fairness, it also reflects a static conception of language that neglects learner individuality, plurilingual repertoires, and contexts (Machetti, 2022). The notion of glocalization (Tsui, 2020) offers a productive perspective for rethinking assessment, emphasizing the interrelation between global frameworks and local pedagogical, cultural, and affective dimensions.

This poster draws on the conceptual design of the Italian Scenario-Based Assessment (Peri, 2025), engineered to measure situated Italian language proficiency online. Preliminary data from its first administration will be discussed, providing insights into how LOLA and SBA principles can inform large-scale assessment while remaining responsive to local contexts and learner diversity.

23. Flexibility Without Compromise: Adapting the Delphi Method to Language Assessment Studies

Authors: Birgitte Grande (Norwegian Defence University College), Clayton David Leishman (The Unconventional)

Despite widespread adoption across disciplines, the Delphi method continues to face significant methodological criticism. Concerns include clear standardization in definitions and processes, questions about reliability and validity across the numerous variants that have emerged, and the need for consensus on what constitutes rigorous application. The method's greatest strength, its flexibility, has thus led to methodological confusion and disputes over scientific rigor. These challenges raise the question of how much the Delphi method can be altered without compromising reliability and validity. This poster complements earlier Delphi studies targeting educational contexts, from the perspective of language assessment, guided by two research questions:

1. What methodological safeguards prevent Delphi adaptations from compromising scientific rigor in language assessment research?
2. How can the Delphi approach balance flexibility with its framework when addressing the topics of rater subjectivity and evolving language constructs?

Informed by our recent study on oral proficiency level descriptors and subjective ratings, utilising a modified Delphi method and consulting an expert panel (n= 14), our poster offers safeguarding steps that can prevent the method from turning into a questionable research practice, thereby contributing to a growing body of evidence that the Delphi method presents a viable option for research within the fields of language assessment and applied linguistics.

24. The development of an audio-visual listening test for the Canadian Armed Forces Military Training and Cooperation Programme

Authors: Nancy Powers (Department of National Defence), May Tan (Department of National Defence)

This poster presents the development of a new audio-video (AV) listening test for students studying English under the Military Training and Cooperation Programme at the Canadian Forces Language School. At the end of their course, students are evaluated on their listening, speaking, reading and writing skills.

This test development started because listening scores from the current test are significantly lower compared to scores in other skills, and do not reflect students' listening comprehension abilities as demonstrated during their speaking test and interactions with teachers.

The new computer-delivered test contains both audio and video items. The items were developed according to the STANAG 6001 language descriptors and authentic material was used as much as possible. Military content was also included. Students can control their AV player and make adjustments when they are listening.

The Assessment Use Argument claims underpinning this test development are that:

- 1) the test will have beneficial consequences for students because having more control over the listening process decreases anxiety.
- 2) the test score interpretations will be meaningful because the listening passages reflect the Target Language Use domain better.
- 3) it will be more relevant to military students because of the use of STANAG 6001, a common standard used by the NATO military community, in item development, and the inclusion of military content
- 4) the new AV listening test will be more equitable, especially to students who are at the lower levels of the scale.

Student comments and results from trialing tend to bear out these claims.

25. Bridging Cognitive Evidence and Test Validity: A Scoping Review of Biometric Research in L2 Listening Assessment

Authors: Shanshan He (University of Western Ontario)

Assessing second language (L2) listening is inherently complex because it requires real-time processing of multimodal input while being shaped by both global assessment frameworks and local contextual factors. Understanding how test tasks cognitively engage listening stimuli is, therefore, central to ensuring the cognitive validity of listening assessments. To address this issue, this scoping review synthesises research that has employed biometric methods to explore the mental processes underlying L2 listening assessment.

The review follows established guidelines for qualitative synthesis (Chong & Plonsky, 2021) and incorporates the Synthesis Methods and Reporting Tool (Chong, 2025) and the PRISMA-ScR framework (Tricco et al., 2018) to enhance transparency and replicability. Studies were included if they involved L2 learners, employed biometric measures, and were published in peer-reviewed journals or doctoral dissertations. A systematic search conducted in March 2025 across six databases identified 46 eligible studies. A coding scheme, developed from the review questions, was piloted, refined, and applied across the dataset, with a second coder verifying the reliability of 10–20% of the results.

Preliminary findings suggest that most studies have compared L1 and L2 listening processes, with eye tracking and fMRI being the most frequently employed methods. However, fMRI studies face ecological and methodological challenges, and few have connected biometric data to test validation or fairness concerns. This scoping review identifies key methodological and conceptual gaps in biometric listening research. It demonstrates how biometric evidence can inform the design of valid, context-sensitive listening assessments that bridge global frameworks with local practices.

26. Bridging Global and Local: Collaborative Reflections on Developing Language Assessment Literacy in Canadian TESOL Teacher Education

Authors: Shanshan He (University of Western Ontario), Chuan Liu (University of Western Ontario)

Language assessment literacy (LAL) is increasingly recognized as a critical component of Canadian TESOL teacher education, equipping preservice teachers to design and implement classroom-based assessments. Despite this emphasis, preservice teachers often struggle to bridge theoretical knowledge with practical application and to adapt assessment practices to their own teaching contexts. While LAL courses remain the primary avenue for developing these skills, limited research has explored the collaborative processes that shape LAL curriculum development.

This study responds to calls for research in diverse educational contexts by examining how two LAL educators adapted and co-taught a graduate-level language assessment course, originally designed by Dr. Ruslan Suvorov, in a Canadian TESOL program. Employing collaborative autoethnography (Chang et al., 2016), we analyzed curriculum artefacts (syllabi, slides), narrative accounts of teaching experiences, and transcripts of collaborative planning and reflection sessions. Data analysis proceeded through individual reflection, collaborative dialogue, and iterative co-construction of themes.

Three key findings emerged. First, collaboration served as reciprocal professional learning, as both instructors deepened their pedagogical knowledge through dialogue and reflection. Second, shared commitments guided how the curriculum was adapted and enacted, with readings and tasks deliberately connecting preservice teachers' home contexts to Canadian and international assessment standards. Third, divergent pedagogical approaches shaped course delivery, with one instructor emphasizing teacher-led discussions and the other prioritizing student-centered activities.

The study highlights how collaborative curriculum-making can strengthen preservice teachers' LAL by integrating global frameworks with local needs, while also enriching instructors' professional development. It offers concrete strategies for TESOL programs seeking to foster LAL in contextually responsive ways.

27. The Development and Validation of an Innovative, Fully-automated Interactive Speaking Task and its Implications for Contextualized Assessments

Authors: Yena Park (Duolingo), Yigal Attali (Duolingo), Xiaowan Zhang (Duolingo), Jacqueline Church (Duolingo), Kai-Ling Lo (Duolingo), Andrew Runge (Duolingo), Geoff LaFlair (Duolingo)

This paper presents the development and validation of an innovative Interactive Speaking task that assesses spoken communication through a fully automated system powered by generative AI (genAI). This task bridges the authenticity and interactivity of direct assessments with the scalability and efficiency of standardized semi-direct assessments (Clark, 1979; Qian, 2009). Leveraging recent advancements in genAI, this task engages the test taker in interview-style, topic-driven conversations by adaptively administering questions based on real-time evaluations of task completion.

We conducted three empirical studies with 3,959 test takers worldwide to investigate the validity of this task. Study I examined the quality of AI-powered question generation, task-completion evaluation, and follow-up question selection against expert judgments. Study II examined the task's psychometric properties, including reliability and concurrent validity. Study III used multidimensional analysis to compare the linguistic features of responses elicited by the Interactive Speaking task versus an indirect monologic task.

Findings demonstrated strong alignment between AI and human experts in question generation and task-completion evaluation, with some divergence in follow-up question selection. Psychometric analyses indicated that the AI-generated task scores were highly reliable and predictive of external speaking criteria. Additionally, responses from the Interactive Speaking task exhibited higher prevalence of features connected to interactivity than monologic responses.

This work highlights the potential of genAI in shaping future speaking assessments. The ability of genAI to generate and adaptively administer questions at scale may serve as a critical first step toward contextual localization within a standardized testing framework.

28. Writing and Identity Development through Bilingual Writing Portfolios in a Dual Language Bilingual Program in Texas

Author: GoMee Park (University of Texas Rio Grande Valley), Jennifer Burden (Graciela Garcia Elementary)

The persistent achievement gap between English learners (ELs) and their monolingual peers is exacerbated by assessments rooted in monoglossic ideologies that privilege English-only performance. Under accountability pressures, schools often “teach to the test,” narrowing instruction and weakening bilingual program goals (Menken, 2006). As Messick (1996) argued, learning a language and preparing for its test should be seamlessly connected, yet for bilingual programs this alignment remains unresolved.

This participatory action research (Efron & Ravid, 2020) was conducted at a dual language bilingual education (DLBE) elementary school in the U.S.–Mexico borderlands, where 72.3% of students are ELs. A school principal and a university researcher collaborated as co-researchers to document the implementation of a schoolwide bilingual writing portfolio designed to support students' writing development and their performance on the writing domain of the state-mandated English proficiency test. Guided by translanguaging theory (García, 2009), Learning-Oriented Assessment (Carless et al., 2006), and Critical Language Testing (Shohamy, 2001), the project sought to position students as co-constructors of learning while challenging the dominance of English-only assessment practices.

Students in grades 3–5 engaged in daily 20-minute writing tasks, first with simple writing prompts and later with accompanied content-based readings to improve academic writing. Students chose to write in English and/or Spanish, while teachers provided informal feedback using the TELPAS rubric. At the end of the year, 157 students and 15 teachers completed surveys. Preliminary findings from both teacher and student surveys indicate that the portfolio enhanced writing development and supported bilingual identity of these students.

29. Investigating writing behaviours, neuro-cognitive processes, and their relationships to test scores in two integrated writing formats

Authors: Andrea Revesz¹, Hyeonjeong Jeong³, Sathena Chan², Haining Cui³, Atsushi Doi³, Dominique Jiang¹, Sewon Kim¹, Nicola Latimer², Motoaki Sugiura³

Organizations: 1: University College London, 2: University of Bedfordshire, 3: Tohoku University

While the impact of proficiency on writing outcomes is well documented, its influence on writing behaviours and neurocognitive processes remains underexplored. This project addresses this gap by examining how writing behaviours and associated neurocognitive processes vary across proficiency levels in two integrated writing formats (a reading-listening-writing task and a shorter reading-writing discussion task). It aims to gather validity argument evidence for using these task types to assess contemporary academic language skills.

Sixty L2 users (B1–C1 CEFR levels) complete two reading-listening-writing and two reading-writing discussion tasks from TOEFL iBT. One version of each is performed in an fMRI scanner to capture brain activity; whereas the other is completed under more natural conditions with eye-tracking to record test-takers' gaze behaviours. Task order and test conditions are counterbalanced. Writing behaviours are logged via keystroke-logging software, and a subset of test-takers also participate in stimulated recall after their last performance. We analyse the keystroke data for fluency, pausing, and revision behaviours; score responses using TOEFL iBT criteria; and examine fMRI and eye-tracking data for brain activation patterns and gaze behaviours. Stimulated recall data is coded using models of writing, reading, and listening (Kellogg, 1996; Khalifa & Weir, 2009; Field, 2013). Mixed-effects regressions assess the relationships between proficiency and process indices.

Preliminary findings suggest higher-proficiency test-takers write faster, revise more, show more global reading behaviours, engage in more planning, and display greater conceptualisation-related brain activation. We will discuss how these insights may inform formative feedback provision and development of scoring engines through process-based metrics.

30. The Impact of Exposure to Global Englishes Accents on Chinese Children's Learning of EFL (English as A Foreign Language)

Author: Zhuohan Chen (University of Oxford)

An ongoing PhD project explores how exposure to diverse English accents influences Mandarin-speaking children's L2 word recognition, speech production, and Global Englishes (GE) awareness. Grounded in the Representation Quality Hypothesis, the project builds on evidence that acoustic variability strengthens lexical representations and promotes generalisation. While High-Variability Phonetic Training (HVPT) has established adult gains in L2 phoneme perception and production, research on accent variability, i.e., structured differences shaped by talkers' regional or social identity, remains limited, especially for children. Bridging HVPT and GE perspectives, this project investigates whether exposure to GE varieties supports comprehension and intelligibility without harming linguistic development.

Study 1 (systematic review with meta-analysis) synthesised 17 experimental and 13 cross-sectional studies ($n = 5,392$, 18 countries). Results showed no detrimental effects: talker variability significantly improved phoneme identification ($d = 0.61$, $p < .01$), while accent/talker variability modestly enhanced word recognition ($d = 0.34$, $p = 0.054$), with tentative advantages for learners from linguistically distant L1s.

Study 2 (cross-sectional) tested 69 Mandarin-speaking children ($Y1 = 38$; $Y4 = 31$) using AI-generated, authenticity-checked audio in four accents (RP, GA, Mandarin, Indian English) and a Berkeley Puppet Interview. Y1 children showed uniform recognition but slower processing of Indian English; Y4 children showed reduced accuracy and more negative attitudes toward Mandarin-accented English.

Study 3 (ongoing RCT) examines whether exposure to varied versus single accents via AI-generated GE videos affects Year 4 children's word recognition, production, and GE awareness, with pre-, post-, and delayed-post-tests planned. Collectively, this project provides empirical evidence to inform GE-oriented pedagogy and challenge native-speakerism.



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

Lancaster University



Offered by our expert team:

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

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

Enquiries: socialscienceteaching@lancaster.ac.uk

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



Linguistics at
Lancaster University is
ranked **2nd in the world**
– QS World Subject
Rankings 2026

LTTC® The Language Training & Testing Center

Powering Research in Language Assessment

Real assessments. Real data. Research that matters.

The Language Training & Testing Center (LTTC) is a non-profit institute founded in 1951 and registered with Taiwan's Ministry of Education. For 75 years, LTTC has led the development of rigorous, locally grounded and internationally aligned language assessments—advancing quality, fairness, and innovation in language education.

LTTC's flagship tests—BESTEP, FLPT, GEPT, GEPT Kids—are widely used across educational and policy contexts and provide high-quality operational data for empirical research in language assessment and learning.

LTTC offers Teaching and Research Grants and Joint-Funded Research Grants supporting research in language assessment, linguistics, and language education. Researchers may apply for funding, access to test materials and datasets, or resource access only.



Scan the QR code to learn more about grant opportunities.



Scan the QR code to view selected research outcomes and publications from previous LTTC-funded projects.



BESTEP
BEST Test of English
Proficiency

FLPT
Foreign Language
Proficiency Test

GEPT
LTTC 全民英檢

GEPT Kids
小學英檢





British Council English Language Research

British Council English Language Research experts drive innovation through impactful research in product development and language 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 research and our programmes, collaborating with partners and fostering emerging talent in the field.

English Language Research |
British Council



Research and insight |
British Council





WIDA Summer Research Internships

WIDA offers summer research internships in language assessment to graduate students. Interns will participate in WIDA assessment research projects and collaborate with WIDA researchers on projects that address academic language development in the K-12 context. Research interns have co-presented their work with WIDA researchers at conferences such as LTRC, MwALT, ECOLT, and NCME.

Eligibility

- Full-time enrollment in a doctoral program related to language assessment
- Completion of a minimum of two years of coursework toward a doctoral degree, prior to beginning the internship

For more information, visit
wida.wisc.edu/about/careers/internship.



wida.wisc.edu

WIDA is housed within the Wisconsin Center for Education Research at the University of Wisconsin-Madison. © 2025 The Board of Regents of the University of Wisconsin System, on behalf of WIDA



› Discover our LANGUAGECERT Research & Validation Series

LANGUAGECERT exams are backed by rigorous research, continuous quality assurance, and global benchmarking.

Dive into our open-access research volumes to discover how we ensure validity, reliability, and fairness in language testing.



Download
your copies at
languagecert.org



Trusted by governments, universities, and employers worldwide, LANGUAGECERT assessments open doors for learners and professionals across the globe.



Online
Events

Network
Opportunities

Affordable
Conference

Conference
Proceedings



The Canadian Association of Language Assessment (CALA) was officially established in 2009. CALA is a professional and academic association dealing with language assessment in Canada.

L'Association canadienne pour l'évaluation des langues (ACÉL) a été officiellement créée en 2009. L'ACÉL est une association professionnelle et scientifique qui œuvre dans le domaine de l'évaluation linguistique au Canada.

Join Us!
Rejoignez-Nous!

<https://www.cala-acel.org/>



Call for papers - June 29
L'appel à communications - 29 juin

Conference/conférence
Halifax, NS
September/septembre 25-26, 2026

événements
en ligne

opportunités de
réseautage

conférence
abordable

actes du
colloque



TOEFL® iBT is Entering a New Era



Recent Publications

Davis, L., & Norris, J. M. (2023). *A comparison of two TOEFL® writing tasks* (ETS Research Memorandum No. RM-23-06). ETS. <https://www.ets.org/Media/Research/pdf/RM-23-06.pdf>

Manna, V. F., Li, S., Papageorgiou, S., & Gu, L. (2025). *TOEFL iBT® technical manual* (TOEFL Research Report No. RR-106). ETS. <https://doi.org/10.64634/eje8f497>

Papageorgiou, S., Li, S., Gu, L., & Barton, K. (in press). *A banded score scale for the TOEFL iBT test: Rationale, development, and score linking*. (ETS Research Report Series). ETS.

Hsieh, C.-N., & Ohta, R. (in press). *Examining L2 writing processes in the TOEFL iBT Write for an Academic Discussion task using eye-tracking, keystroke logging, and stimulated recall*. ETS Research Report Series. ETS.

Isbell, D. R., & Crowther, D. (in press). Investigating the real-world relevance of an academic English speaking test: Extrapolating subjective evaluations and linguistic performance characteristics. *Language Testing*.

For more...



ETS Research Reports (2025-)
rr.ets.org



ETS Research Reports (through 2024)
onlinelibrary.wiley.com/journal/23308516

Supporting and Recognizing Research

- TOEFL research grants support established researchers and graduate students
- TOEFL awards recognize excellence in research for both students and professionals

Learn more here!



ets.org/toefl/grants



Visit ets.org/toefl/research

Copyright © 2026 by ETS. TOEFL and TOEFL iBT are registered trademarks of ETS in the United States and other countries. The Eight-point logo is a trademark of ETS.



A Total Testing Solution For a Variety of Purposes



Fully Customized

Tailored to each program
Tests are designed based on the curriculum and the program expectations.



All Language Skills

Covering all language skills & sub-skills
Each test has multiple parallel versions that are randomly assigned to test takers.



Meaningful Reports

Detailed analysis of performance
Results are reported in CEFR or CLB.
A variety of reports can be accessed through a dashboard and scores can be seamlessly integrated with existing systems.



Built-in AI-Proctor

Comes standard with all tests
The AI-proctor ensures test integrity by flagging every significant occurrence during the test.



Dedicated Dashboard

Secure online access to test results and proctoring reports are provided through a user-friendly dashboard.



Accessibility Tools

Built-in accessibility tools provide accommodation to the individual needs of each student.



Cost Efficient

Online registration, AI-Powered proctoring, accessibility tools, and a customizable dashboard come with all tests for one all-inclusive price.



precisesolutions.com



info@precisesolutions.com



ILTA

International
Language Testing
Association



GHENT

Belgium



GHENT UNIVERSITY



LTRC 2027

LANGUAGE ASSESSMENT
IN INTERDISCIPLINARY
RESEARCH

JUNE 26-30, 2027



duolingo english test

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.



Built on the latest language assessment science

- ✓ Accessible by design, supporting test takers wherever they are for just \$70
- ✓ Built on rigorous research and industry- leading security
- ✓ Integrates the latest assessment science and AI for accurate results
- ✓ Accepted by over 6,000 programs worldwide



englishtest.duolingo.com