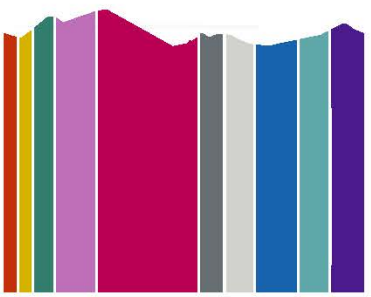
universität innsbruck

ILTA
INTERNATIONAL LANGUAGE TESTING ASSOCIATION



LTRC 2024

45th LANGUAGE TESTING RESEARCH COLLOQUIUM

REFORMING LANGUAGE ASSESSMENT SYSTEMS, REFORMING LANGUAGE ASSESSMENT RESEARCH

JULY 1-5, 2024 INNSBRUCK, AUSTRIA

MICHIGAN LANGUAGE ASSESSMENT

Created by Assessment Experts. Trusted Around the World.

A proud sponsor of Language Testing Research Colloquium 2024, Michigan Language Assessment is dedicated to actively supporting and participating in the language testing field.

Drawing on the expertise of two of the world's leading universities, we are committed to changing lives, advancing global mobility, and supporting diversity, equity and inclusion through our portfolio of secure, trusted English language assessments.

Part of

CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

UNIVERSITY OF MICHIGAN

MICHIGANASSESSMENT.ORG

PLATINUM SPONSORS

BRITISH COUNCIL

duolingo english test

ETS TOEFL®

Pearson | PTE

GOLD SPONSORS

CAMBRIDGE

CAL CENTER FOR APPLIED LINGUISTICS

g.a.s.t. gesellschaft für akademische studienvorbereitung und testentwicklung e.v.

GOETHE INSTITUT
Sprache. Kultur. Deutschland.

IELTS

ITTC GEPT® BESTEP

OXFORD TEST OF ENGLISH

SILVER SPONSORS

FRANCE EDUCATION INTERNATIONAL

MetaMetrics

OET

osd

BRONZE SPONSORS

Lancaster University

LANGUAGE LEARNING AND TESTING FOUNDATION
www.LLTF.net

UNIVERSITY OF MICHIGAN | MICHIGAN LANGUAGE ASSESSMENT

THE UNIVERSITY OF CHICAGO
OFFICE OF LANGUAGE ASSESSMENT

WIDA™
UNIVERSITY OF WISCONSIN-MADISON

universität innsbruck
Fakultät für LehrerInnenbildung

INNS' BRUCK

LAND TIROL

Table of Contents

Message from the ILTA President

On behalf of the International Language Testing Association (ILTA), I would like to extend a warm and heartfelt Herzlich Willkommen to you all – welcome to the 45th Language Testing Research Colloquium in the city of Innsbruck, in the midst of the Tyrolian Alps. Over the coming days, we will be discussing the latest research and approaches to Reforming Language Assessment Systems, and to Reforming Language Assessment Research. I look forward to engaging with you on innovative research methods, reform endeavors and exciting new insights!

But first of all, let me give a huge THANK YOU to the awesome LTRC organizing team: Benjamin Kremmel as conference chair worked wonders with his supportive team consisting of Simone Baumgartinger, Kathrin Eberharter, Viktoria Ebner, Annette Giesinger, Elisa Guggenbichler, Eva Konrad, Doris Moser-Frötscher, Carol Spöttl and Sigrid Hauser. Our thanks also go to our external consultants Nivja de Jong, Luke Harding, Claudia Harsch, and to our LTRC advisory committee and their invaluable input: Beverly Baker, Lorena Llosa, Conor McKeon, Yasuyo Sawaki, Sun-Young Shin, and Elvis Wagner – great to be able to rely on your expertise. A big thank you is also due to our management company Nardone, with Valerie Smith and Laura Haller supporting communications and preparations, competent as always. Furthermore, our thanks and appreciations go to our sponsors British Council, Duolingo English Test, ETS TOEFL®, Pearson PTE, Cambridge University Press & Assessment, g.a.s.t. e.V., Goethe Insitut, IELTS, LTTC GEPT, Centre for Applied Linguistics, Oxford Test of English, MetaMetrics®, OET, France Éducation International, Lancaster University, Language Learning and Testing Foundation, Michigan Language Assessment, The University of Chicago Office of Language Assessment and WIDA – great to have your support! We also extend our warm thanks to our workshop leaders Stefanie A. Wind, Andrea Révész, Joseph Lo Bianco, Mina Patel, Alistair Van Moere and Jing Wei for their commitment and time, and a big thank you to the Center for Applied Linguistics and the British Council for sponsoring two of our workshops. Last but not least, a big thanks goes to you, the presenters and attendees, for coming to Innsbruck and making LTRC possible!

There are exciting events ahead of us. As always, all newcomers to LTRC are warmly invited to our Newcomers' Session on Tuesday, 4.30-5.30pm – a fabulous way into the LTRC family! Many thanks to Meg Malone for leading this again. This is followed by a Welcome Session, which leads us directly into the Opening Symposium at 6pm, a special tribute to our beloved colleague Tim McNamara, who sadly passed away last year. May I invite you to reflect on his scholarship in the realm of fairness and justice, themes that remain ever so important. After this commemorating event, you are invited to the sponsored LTRC Opening Reception at 7pm – a fabulous opportunity to catch up and socialize.

The next three days are full of symposia, paper presentations, works-in-progress, posters and demonstrations, as well as four Special Sessions, a "tradition" that started last year in New York: You can liaise with colleagues in the "Creatively Engaged and Recharged" Session for Mid- and Senior-Career Professionals, in sessions targeting young researchers on "Navigating the job market" and "How to be a (good) reviewer", or in our "Rainbow" session. The SIG meetings take place on Friday 1.30pm, and LTRC offers

many opportunities to network. This year, thanks to a range of generous sponsors, not only will coffee breaks be provided as part of the registration fee, but also your lunches. So you can enjoy lots of opportunities for networking, as well as great local cuisine, on all three conference days! As with last year, our awesome Graduate Student Assembly is offering Mentoring Sessions, an excellent opportunity for emerging researchers to liaise with senior colleagues – many thanks to all who dedicate their time. Wednesday night is dedicated to Networking Dinners, and after that, we dedicate the evening to the memory of our esteemed colleague, friend and former EALTA president Jamie Dunlea, who sadly passed away in March this year. You are invited to commemorate with us in the Galway Bay Irish Pub, in tribute to Jamie's Irish roots. There will also be a book on condolences at the reception desk to leave your memories; we will pass the book on to Jamie's family.

LTRC would not be LTRC without our awards. On Wednesday, we welcome Lynda Taylor giving the Alan Davies Lecture (supported by the British Council) on Experimenting with Uncertainty, Advancing Social Justice: Placing Equity, Diversity, Inclusion and Access Centre Stage. On Thursday, we invite you to attend the Samuel J. Messick Memorial Lecture (supported by Educational Testing Service): Micheline Chalhoub-Deville will share with us her insights on Reimagining validity in accountability testing: Understanding consequences in a social context. On Friday, we congregate for the Cambridge/ILTA Distinguished Achievement Award Lecture (supported by Cambridge University Press & Assessment and ILTA) on Integration and Inclusiveness in Language Assessment – congratulations to Antony John Kunnan for this well-deserved award! All other awards will be celebrated at our LTRC Banquet Friday night.

I would also like to draw your attention to our Annual Business Meeting. All ILTA members will have received the reports by the President and Treasurer, along with information about the bids for LTRC 2027 before the conference. During the ABM, which takes place Thursday 4 July, 12.30-14, we have time to discuss latest initiatives, proposals stemming from the membership survey and from our committees and task forces, as well as any open questions arising from the reports, along with all aspects that need voting. The actual voting will then take place online after the conference, so that the whole ILTA community can participate.

Let me take the chance here to point out the benefits of ILTA to those of you who may not yet be a member – we are a vibrant community dedicated to promoting the improvement of language testing and assessment throughout the world. We offer a range of SIGs, webinars, student support, grants for students and the wider community, resources, and we are dedicated to expanding our outreach and widening access for members from all regions around the globe. We look forward to welcoming you into our community.

A warm welcome once more to the 45th LTRC and to Innsbruck. May we have engaging and inspiring days with networking, new insights, memorials, socializing and recharging!

Claudia Harsch
ILTA President 2024

Welcome from the LTRC Chair

Dear LTRC participants,

On behalf of the entire LTRC 2024 Organizing committee: Welcome to Innsbruck! The Language Testing Research Group Innsbruck (LTRGI) is incredibly excited to have you all here in Innsbruck after our intense year of preparations. My team and I very much hope that you will enjoy the conference we have prepared for you, which ILTA president Claudia Harsch has already detailed in her welcome message.

Being allowed to host LTRC 2024 as the conference chair is an enormous honor. It has been a very special experience for me personally on several levels and I am incredibly grateful for this tiny 'full circle' moment. I attended my very first LTRC in 2014 (Amsterdam). Standing up on that infamous ladder held by John de Jong to receive the IELTS Caroline Clapham Award for the MA thesis I had just completed at Lancaster University, I would have never dreamed of welcoming the ILTA community, which had so warmly received me, to my university in Innsbruck just ten years later. Just like back then, I am filled with overwhelming gratitude for the support of the amazing team that is the Language Testing Research Group Innsbruck, some members of which I was allowed to thank ten years ago and are still incredible sources of mentorship and friendship today. The team have worked their socks off this year to give you all an unforgettable conference experience in these coming days.

Carol has been instrumental in securing the largest sum of sponsorship ever generated for LTRC. Sigrid has been in charge of catering and social program options. Those of you who have been to one of our Innsbruck events before will know why. Those who haven't, will know why after this week. Doris has been the go-to person for all matters pre-conference workshop, EXCEL wizardry, and proofreading. Elisa made sure all the technology and recording is working and has been the main liaison with the ILTA GSA. Viktoria has overseen the conftool platform and has been critical in setting up the app. Eva and Simone have supported with all kinds of email announcement drafts, putting together the abstract booklet and lots of other short-notice tasks. Kathrin, as usual, has been my right hand, managing the program, from abstract review to implementing it all in print and app, and also coordinating, with Linja and Carina, all design processes and decisions. Annette has been the administrative backbone of the entire operation, from venue and hotel room bookings, to ordering all conference goodies, and patiently creating and maintaining the conference website. Everyone chipped in with ideas and support, helping out in all areas, covering for each other as we always do and is the strength of this team. I consider it an incredible privilege to be part of this team, not just when it comes to the organization of LTRC 2024, but every single project we embark on. Thank you for all the hard work, fun and kindness, and for your trust, loyalty, and commitment in following through even the daftest ideas with me.

Luke Harding, my mentor, good friend, and co-conspirator in all of this from the very beginning, deserves a special mention here as well. If it were not for his encouragement and advice, and constant support, I would have never applied for hosting an LTRC some

four years ago. Claudia Harsch and Nivja de Jong have also supported us tremendously with their input throughout the process. Thank you!

I would like to express my gratitude to Laura, Valerie, and Michele at Nardone for their competent support throughout the organization progress, as well as the Lorena Llosa and the LTRC advisory committee, the entire ILTA EB, and the ILTA GSA representatives for the good collaboration.

I am grateful to all our abstract reviewers, the special session leaders and panelists, session chairs, networking dinner table hosts, and SIG convenors for contributing to the exciting program we have ahead of us. I am grateful to our numerous sponsors whose generosity has allowed us to put on the conference we wanted, with lower and thus more accessible fees, and with increased opportunities for mingling and networking through the coffee breaks and on-site lunches.

I want to extend a warm thanks to all the keynote speakers, pre-conference workshop leaders, and opening symposium speakers that have kindly followed our invitation. They have allowed us to put our stamp on the content of LTRC 2024, just the way we wanted.

We chose the theme of “Reforming language assessment systems, reforming language assessment research” partly because of the many exam reform projects our research group has been involved in on a national level, and partly to address the ever-increasing number of changes in language assessment we and many around the world face. Advances in digital technology, a growing but long-overdue awareness of social justice considerations and changes in the very nature of communication itself and the resulting impact on constructs, all these present fresh opportunities but also daunting challenges. At the same time, the Open Science movement, and numerous methodological innovations might allow us to address these developments in novel ways, reforming our research to become more transparent, collaborative, robust, and communicable to stakeholders.

So, I would like to thank all those who responded to our call for proposals on this theme and altogether submitted a staggering 378 abstracts for us to review and create this program now in front of you. The number of high-quality submissions has made the process highly competitive and our decisions very hard. Without these 378 abstracts, we would not have been able to produce such a stimulating program that has attracted more than 340 international delegates to LTRC 2024 Innsbruck. The fact that abstract submissions have gone up by an astonishing 69 percent since Amsterdam 2014 indicates that our field is healthy, prolific, and only gaining in importance and attention.

Finally, whether as a speaker or a participant, thank YOU for coming! Whether it is your first LTRC or you have been to so many LTRCs you have lost count, may you feel as welcomed and inspired by this community as I did ten years ago and have done ever since. I hope you all find LTRC 2024 intellectually and interpersonally stimulating and enjoy your time in Innsbruck.

Benjamin Kremmel
LTRC 2024 Conference Chair

Conference Organization

Chair

Benjamin Kremmel
University of Innsbruck

Organizing Committee

Simone Baumgartinger
University of Innsbruck
Kathrin Eberharter
University of Innsbruck
Viktoria Ebner
University of Innsbruck
Annette Giesinger
University of Innsbruck
Elisa Guggenbichler
University of Innsbruck
Sigrid Hauser
University of Innsbruck
Eva Konrad
University of Innsbruck
Doris Moser-Frötscher
University of Innsbruck
Carol Spöttl
University of Innsbruck

External Consultants

Luke Harding
Lancaster University
Claudia Harsch
University of Bremen
Nivja de Jong
University of Leiden

Student Assistants

Ayu Sandra Dewi
Sarah Egger
Dominik Fekete
Bettina Pircher
Alexander Simml
Nicholas Tagwerker
Hannah Tauscher

Logo

Linja Meller

Cover & Print Design

Carina Gruschi

Abstract Reviewers

Vahid Aryadoust
Beverly Baker
Khaled Barkaoui
Aaron Olaf Batty
Vivien Elizabeth Berry
Rachel Lunde Brooks
Tineke Brunfaut
J. Dylan Burton
Nathan T. Carr
Carol Chapelle
Mark Derek Champman
Ikkyu Choi
Deborah Crusan
Sara Cushing
Jee Wha Dakin
Bart Deygers
Slobodanka Dimova
Kathrin Eberharter
Jason Fan
Timothy Farnsworth
Kellie Frost
Atta Gebril
Ardeshir Geranpayeh
April Ginther
Brent A. Green
Tony Green
Luke Harding
Claudia Harsch
Sahbi Hidri
Ching-Ni Hsieh
Ari Huhta
Yo In'nami
Talia Isaacs
Daniel R. Isbell
Noriko Iwashita
Gerriet Janssen
Yan Jin
Ute Knoch
Rie Koizumi
Antony John Kunnan
Geoffrey T. LaFlair
Yong-Won Lee
Constant Leung
Gad Lim
Lorena Llosa
Sari Luoma
David MacGregor
Susy Macqueen
Margaret Malone
Doris Moser-Frötscher
Fumiyo Nakatsuhara
Heike Neumann
John Norris
Sally O'Hagan
Barry O'Sullivan
Gary John Ockey
Saerhim Oh
Spiros Papageorgiou
Lia Plakans
John A. Read
Carsten Roever
Olena Rossi
Shahrzad Saif
Nick Saville
Yasuyo Sawaki
Jonathan Schmidgall
Brigita Seguis
Sun-Young Shin
Jeffrey Stewart
Ruslan Suvorov
Lynda Brigid Taylor
Veronika Timpe-Laughlin
Dina Tsagari
Carolyn Turner
Alan Urmston
Alistair Van Moere
Erik Voss
Elvis Wagner
Gillian Wigglesworth
Paula Winke
Mikyung Kim Wolf
Jing Xu
Xun Yan
Guoxing Yu
Cecilia Guanfang Zhao
Ying Zheng

ILTA Executive Board and Committee Members

ILTA Executive Board 2024

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

Members at Large

Kathrin Eberharter, University of Innsbruck
HE Lianzhen, Zhejiang University
Gad Lim, Cambridge Boxhill Language Assessment
Carsten Roever, University of Melbourne
Erik Voss, Columbia University (Communications Committee Chair)

ILTA Staff

Valerie Smith
Laura Haller

LTRC 2024 Chair

Benjamin Kremmel, University of Innsbruck

ILTA Nominating Committee

Chair: David Wei Dai, UCL Institute of Education, University College London
Ivy Chen, University of Melbourne
Dylan Burton, University of Illinois Urbana-Champaign
Daniel Isbell, University of Hawai'i at Mānoa

Graduate Student Assembly

Chair: Chengyuan Yu, Kent State University
Vice Chair: Conor McKeon, Georgetown University
Communications Chair: Inyoung Na, Iowa State University

SIG Chairs

Automated Language Assessment (ALASIG)

Xiaoming Xi, Hong Kong Examinations and Assessment Authority
Jing Xu, Cambridge University Press & Assessment

Integrated Assessment (IASIG)

Sharareh (Sharry) Vahed, Purdue University
Xun Yan, University of Illinois Urbana-Champaign
Rebecca Yeager, University of Iowa

Language Assessment Literacy (LALSIG)

Elsa Fernanda Gonzalez, Universidad Autonoma de Tamaulipas
Gladys Quevedo-Camargo, Universidade de Brasilia

Language Assessment in Aviation (LAASIG)

Natalia Andrade, University of Campinas
Ana Lígia Silva, São Paulo State University - UNESP
Angela Garcia, Carleton University

Language Assessment for Young Learners (Young Learners SIG)

Mark Chapman, WIDA
Veronika Timpe-Laughlin, Educational Testing Service
Jeanne Beck, Iowa State University

Test-taker Insights in Language Assessment (TILASIG)

Andy Jiahao LIU, University of Arizona
Ray Jui-Teng Liao, National Taiwan Ocean University

Awards and Grants Committees

ILTA Best Article Award

Chair: Mikyung Kim Wolf, Educational Testing Service
Beverly Baker, University of Ottawa
Aaron Batty, Keio University
Bart Deygers, Ghent University
Kellie Frost, University of Melbourne

Robert Lado Memorial Award (to be awarded at LTRC)

Chair: Gad Lim, Cambridge Boxhill Language Assessment
Fauve De Backer, Ghent University
Ruslan Suvorov, Western University
Jessica Wu, The Language Training and Testing Center (LTTC)

Cambridge / ILTA Distinguished Achievement Award

Chair: Luke Harding, Lancaster University
Evelina Galaczi, Cambridge University Press & Assessment
Yan Jin, Shanghai Jiao Tong University
Charlie Stansfield, Language Learning and Testing Foundation

ILTA / Duolingo Collaboration and Outreach Grant

Chair: Claudia Harsch, University of Bremen
Antony John Kunnan, Carnegie Mellon University
Jirada Wudthayagorn, Chulalongkorn University
Kathrin Eberharter, University of Innsbruck

TOEFL / ILTA Student Travel Grants

Chair: Luke Harding, Lancaster University
Kathrin Eberharter, University of Innsbruck
Carsten Roever, University of Melbourne

SAGE / ILTA Book Award

Chair: Benjamin Kremmel, University of Innsbruck
HE Lianzhen, Zhejiang University
Noriko Iwashita, University of Queensland
Jamie L. Schissel, University of North Carolina at Greensboro
Sun-Young Shin, Indiana University

Award Winners

Cambridge / ILTA Distinguished Achievement Award

Antony John Kunnan, Carnegie Mellon University

TOEFL / ILTA Student Travel Grants

Geisa Davila Perez, Lancaster University
Coral Yiwei Qin, University of Ottawa
Shengkai Yin, The University of Melbourne/Shanghai Jiao Tong University
Monique Yoder, Michigan State University
Chenyang Zhang, The University of Melbourne

ILTA Best Article Award 2022

What can gaze behaviors, neuroimaging data, and test scores tell us about test method effects and cognitive load in listening assessments?

Vahid Aryadoust, National Institute of Education, Nanyang Technological University
Stacy Foo, National Institute of Education, Nanyang Technological University
Li Ying Ng, National Institute of Education, Nanyang Technological University
Published in: *Language Testing 39*(1), https://doi.org/10.1177/02655322211026876

ILTA / Duolingo Collaboration and Outreach Grant

Developing English language proficiency tests for Thai university students: From theories to practices

Dusadee Rangseechatchawan, Chiang Mai Rajabhat University

Promoting sustainable language assessment practices among middle school teachers in India

Rama Mathew, Freelance ELT Consultant

Promoting meaningful assessment in the language classroom in the Dominican Republic

Angel Arias, Carleton University
Jamie L. Schissel, University of North Carolina at Greensboro

SAGE / ILTA Award for Best Monograph on Language Testing

Slobodanka, Dimova, Xun Yan, & April Ginther (2020). *Local Language Testing: Design, Implementation and Development.* Routledge.

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

Micheline Chalhoub-Deville, University of North Carolina at Greensboro (UNCG)

The Davies Lecture Award (Sponsored by British Council)

Lynda Taylor, University of Bedfordshire

Robert Lado Memorial Award

To be announced at the LTRC Banquet

TOEFL® New Scholar Award 2024

Stefan O'Grady, University of St. Andrews

Jacqueline Ross TOEFL Dissertation Award 2024

Dylan Burton, Michigan State University

The role of nonverbal behavior and affect on ratings of second language proficiency

Supervisor: Paula Winke

Caroline Clapham IELTS Masters Award 2023

Shanshan He, The University of Western Ontario

Exploring the Use of Interactive Videos in an L2 Listening Test

Supervisor: Ruslan Suvorov

Duolingo 2023 Doctoral Dissertation Awards

Zhiyuan Deng, University of Maryland at College Park, Supervisor: Bronson Hui
Elisa Guggenbichler, University of Innsbruck, Supervisor: Benjamin Kremmel
Jia Guo, Queen's University, Supervisor: Liying Cheng
Liam Hannah, Ontario Institute for Studies in Education, Supervisor: Eunice Eunhee Jang
Yuanyue Hao, University of Oxford, Supervisor: Robert Woore
Takehiro Iizuka, University of Maryland at College Park, Supervisor: Bronson Hui
Tingting Liu, Nanyang Technological University, Supervisor: Vahid Aryadoust
Fred S. Tsutagawa, Teachers College, Columbia University, Supervisor: James E. Purpura
Xiaozhu Wang, Beijing Language and Culture University, Supervisor: Jimin Wang
Jincheng Wu, University of Macau, Supervisor: Cecilia Guanfang Zhao
Monique Yoder, Michigan State University, Supervisor: Paula M. Winke
Tiancheng Zhang, The University of Auckland, Supervisor: Rosemary Erlam
Shishi Zhang, University College London, Supervisor: Talia Isaacs

TIRF 2023 Doctoral Dissertation Grant Awardees in Language Assessment

Mohd Iqbal Ahamat, Universiti Sains Malaysia, Supervisor: Muhammad Kamarul Kabilan
Marcella Caprario, Northern Arizona University, Supervisor: Naoko Taguchi
Ananda Astrini Muhammad, Iowa State University, Supervisor: Gary Ockey

TOEFL 2024 Grants for Doctoral Research in Language Assessment

Mutleb Alnafisah, Iowa State University, Supervisor: Gary Ockey
Shireen Baghestani, Iowa State University, Supervisor: Carol Chapelle
Suet Sin Cheung, University College London, Supervisor: Andrea Révész
Zhiyuan Deng, University of Maryland at College Park, Supervisor: Bronson Hui
Elisa Guggenbichler, University of Innsbruck, Supervisor: Benjamin Kremmel
Jia Guo, Queen's University, Supervisor: Liying Cheng
Melissa Hunte, University of Toronto, Supervisor: Eunice Eunhee Jang
Takehiro Iizuka, University of Maryland at College Park, Supervisor: Bronson Hui
Margarida Pato, University of Bedfordshire, Supervisor: Fumiyo Nakatsuhara
Coral Yiwei Qin, University of Ottawa, Supervisor: Beverly Baker
Xingcheng Wang, University of Melbourne, Supervisor: Carsten Roever
Jincheng Wu, University of Macau, Supervisor: Cecilia Guanfang Zhao
Shishi Zhang, University College London, Supervisor: Talia Isaacs

Language Testing 2023 Reviewer of the Year Award

Susy Macqueen, Australian National University
Stefanie A. Wind, University of Alabama

Housekeeping and Important Information

We have excellent **drinking water** here in Austria. 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** at any time.

Further information can be found on the conference **website**:
https://www.uibk.ac.at/en/congress/ltrc2024/

The latest program updates can be found on the **Sched** App:
➔ see LTRC 2024

If you want, you can use the **#LTRC2024** for any social media postings.

WiFi access – "**eduroam**"

or

WiFi access – network name "**UIBK**"

Click on the WiFi-symbol in the taskbar, choose network "UIBK" and type in the required information:

User: **c115135**
Password (WPA2-Enterprise): **LTRC#ibk24**

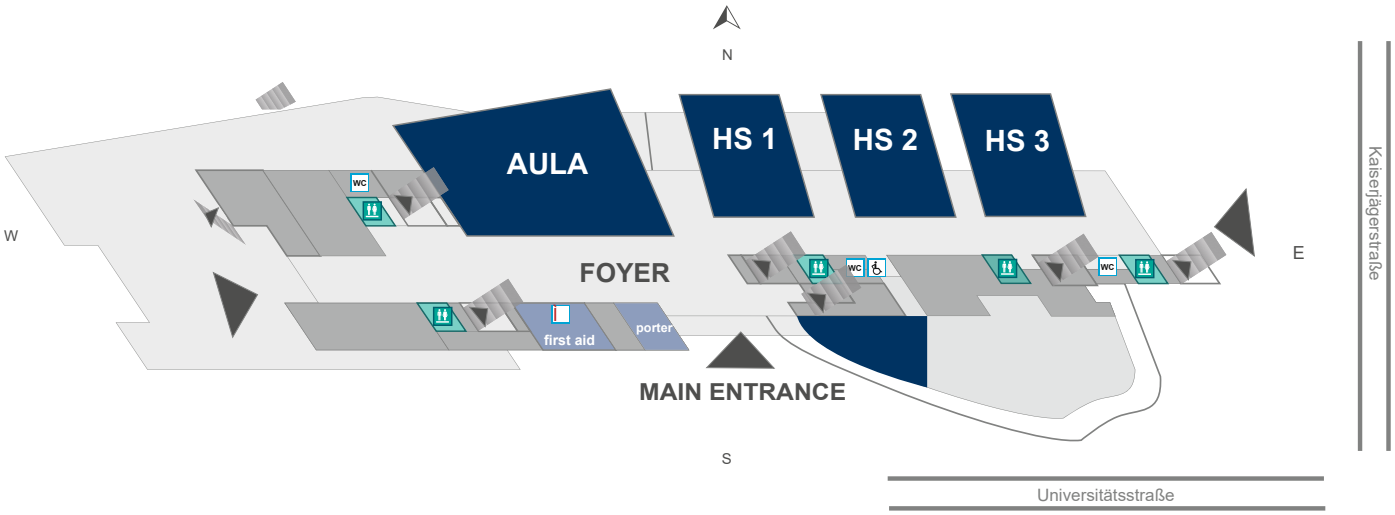
Depending on the settings of your operating system, you might be asked twice to enter the network key (WPA2-Password). Other devices might not prompt you to so at all.

The following rooms offer an induction loop:

- Aula
- Madonnensaal
- Kaiser Leopoldsaal

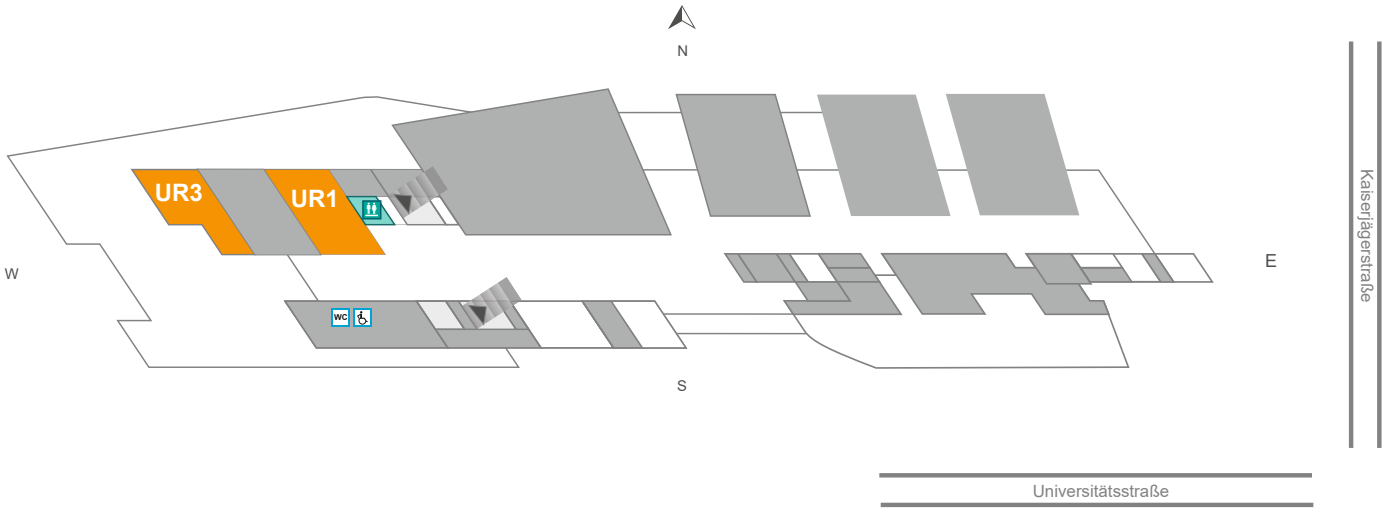
groundfloor

lecture hall nr. 1-3
Aula



basement

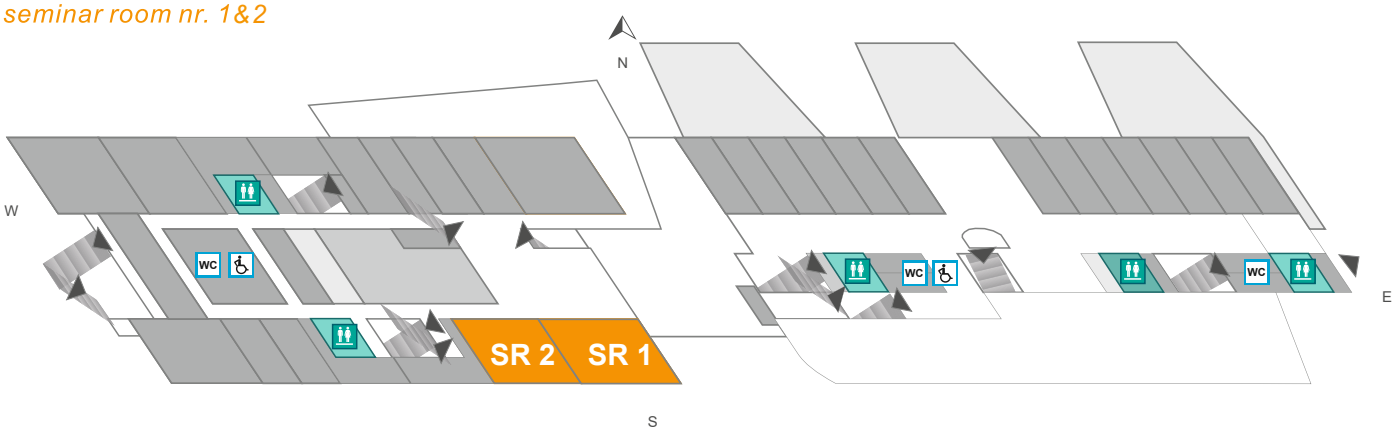
classroom nr. 1&3



Universität Innsbruck
Campus SoWi
Universitätstrasse 15
6020 Innsbruck
T +43 512 507-0

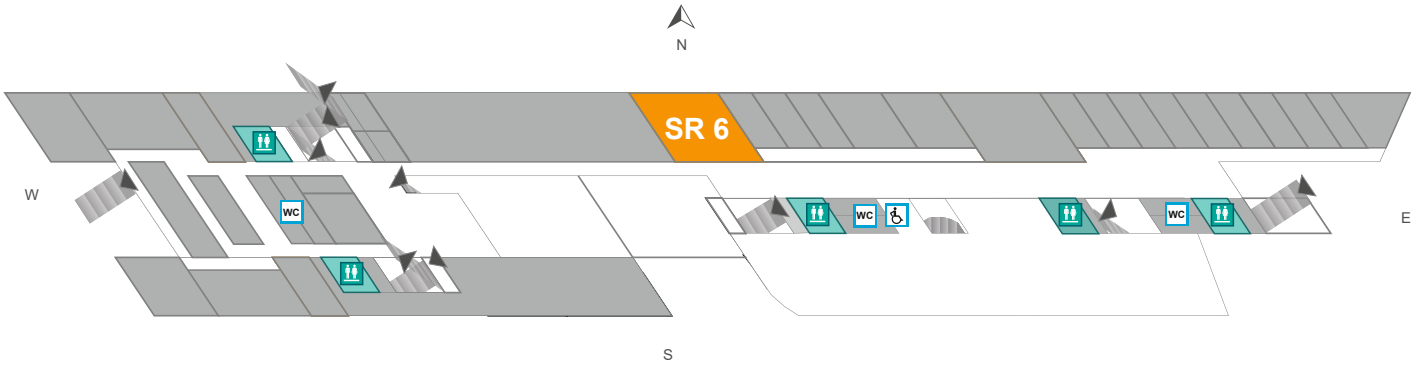
1 st floor

seminar room nr. 1&2



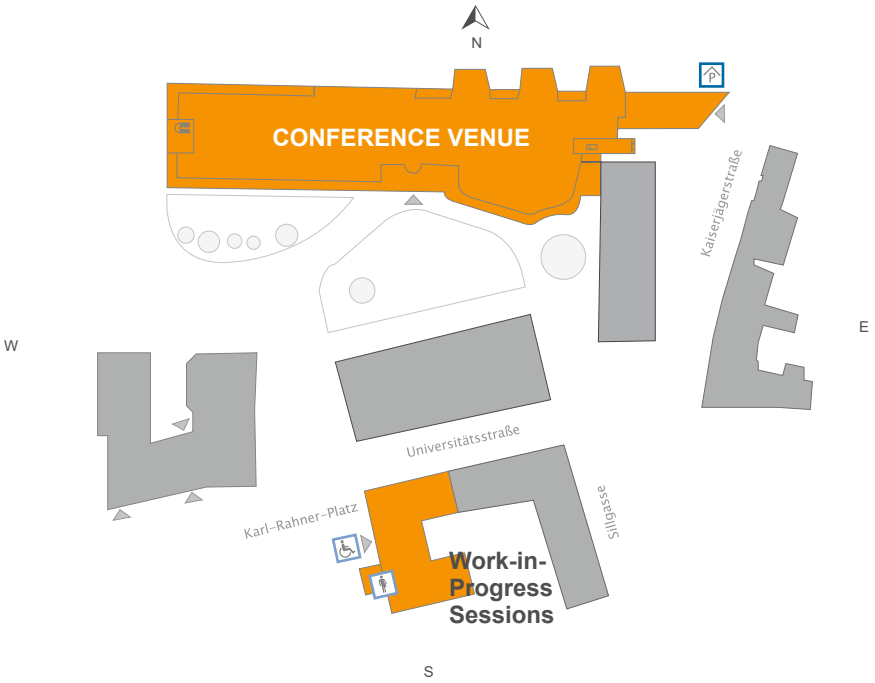
2 nd floor

seminar room nr. 6



Work-in-Progress Sessions

Theologie Building, Karl-Rahner-Platz 2



Universität Innsbruck
Campus SoWi
Universitätstrasse 15
6020 Innsbruck
T +43 512 507-0

Monday, July 1, 2024

8:00 - 9:00	Registration (Location: Foyer)	
	Workshops	
	SR1	SR2
9:00 - 4:00	**Workshop A** (sponsored by Center for Applied Linguistics)**:** **Measurement approaches to exploring survey ratings and rater effects** Stefanie A. Wind	**Workshop B:** **Coding qualitative verbal protocol data for test validation** Andrea Révész

Tuesday, July 2, 2024

8:00 - 9:00	Registration (Location: Foyer)			
	Workshops			
	SR1	UR3	SR2	SR6
9:00 - 4:00	**Workshop A** (sponsored by Center for Applied Linguistics) **Measurement approaches to exploring survey ratings and rater effects** Stefanie A. Wind	**Workshop C** (sponsored by British Council) **Policy Literacy: Exploring Effective Participation for Researchers in Policy Making** Joseph Lo Bianco, Mina Patel	**Workshop D** **Generative AI for content generation and automated scoring: no-code and low-code solutions** Alistair Van Moere, Jing Wei	**ILTA Executive Board Meeting** ***(closed meeting)***
3:00 - 6:00	Registration (Location: Foyer)			
4:30 - 5:30	**Newcomers' Session** (Location: HS1) Meg Malone			
5:45 - 6:00	**Welcome** (Location: Aula)			
6:00 - 7:00	**Opening symposium** (Location: Aula) **Advancing fairness and justice in language testing: Reflecting on Tim McNamara's scholarship** Joseph Lo Bianco, Barbara Seidlhofer, Henry Widdowson, Kellie Frost, Ute Knoch, Susy Macqueen, Jason Fan, Elana Shohamy			
7:00	Welcome Reception (sponsored by Duolingo English Test) (Location: Foyer)			

Wednesday, July 3, 2024

8:00 - 8:30	Registration (Location: Foyer)				
8:30 - 8:50	***Welcome and Opening Remarks*** (Location: Aula)				
9:00 - 10:30	**Parallel Session 1**				
	Aula *Chair: Claudia Harsch*	HS1 *Chair: John Pill*	HS2 *Chair: Susy Macqueen*	HS3 *Chair: Elvis Wagner*	UR3 *Chair: Sara Cushing*
9:00 - 9:30	**Multimodal EAP assessment reconceptualised** Sathena Chan, Nahal Khabbazbashi, Tony Clark	**The sound of one hand clapping: what monologues can tell us about interactional competence** Carsten Roever, Naoki Ikeda	**Reducing language barriers and improving diversity and inclusion in participant recruitment to randomized trials: A role for language assessment** Talia Isaacs, Andrea Vaughan, Eva Burnett, Zsofia Demjen, Marie-Anne Durand, Kate Gillies, Kamlesh Khunti, Jamie Murdoch, Nuru Noor, Leila Rooshenas, Frances Shiely, Harpreet Sood, Fiona Stevenson, Matt Sydes, Shaun Treweek, Katie Biggs	**An eye-tracking study of response processes on C-test items in the Duolingo English Test** Ruslan Suvorov	**Clarifying Links Between Actionable Feedforward and Remediation in Diagnostic Language Assessment: Insights from Medical and Dynamic Assessment** Yong-Won Lee
9:30 - 10:00	**Multimodality: a new construct in writing assessment** Duygu Candarli	**Human- versus artificial-intelligence-based role-play tasks for the assessment of interactional competence: An applied conversation analytic study** Masaki Eguchi, Kotaro Takizawa, Fuma Kurata, Mao Saeki, Yoichi Matsuyama	**Exploring the language and communication demands of early childhood and school teachers in Australia: Implications for language assessment for teacher registration** Xiaoxiao Kong	**Young EFL learners' cognitive processes of taking digitalized picture-based causal explanation speaking tasks: Linking eye gaze with speech production** Wenjun (Elyse) Ding, Guoxing Yu	**Implementing Formative Assessment in the Chinese University EFL Classroom: Understanding Students' Perceptions** Qiaozhen Yan, Xiangdong Gu

Wednesday, July 3, 2024 (continued)

	Parallel Session 1 (continued)				
	Aula *Chair: Claudia Harsch*	HS1 *Chair: John Pill*	HS2 *Chair: Susy Macqueen*	HS3 *Chair: Elvis Wagner*	UR3 *Chair: Sara Cushing*
10:00 - 10:30	**Processing of multimodal input – Towards a more comprehensive definition of integrated writing assessment** Sonja Zimmermann	**Exploring the Potential of Conversational AI for Assessing Second Language Oral Proficiency** Yasin Karatay, Jing Xu	**Assessing the language proficiency of internationally-graduated professionals: The intended vs. actual interpretations** Shahrzad Saif	**Young EFL Students' Writing Performance: Patterns by CEFR Levels and Task Types** Mikyung Kim Wolf, Michael Suhan	**Reforming teacher education to enhance language assessment literacy: New insights from pre-service teachers' reflections** Armin Berger, Helen Heaney
10:30 - 11:00	Coffee Break (sponsored by MetaMetrics)				
11:00 - 12:10	**Alan Davies Lecture (Sponsored by British Council)** (Location: Aula) **Experimenting with Uncertainty, Advancing Social Justice: Placing Equity, Diversity, Inclusion and Access Centre Stage** Lynda Taylor				
12:10 - 12:15	Group picture (Location: Aula)				
12:15 - 1:30	Networking Lunch (sponsored by Pearson)				
1:30 - 3:00	**Works-in-progress**				
	Kaiser-Leopold-Saal (Theologie building) *Chair: Kathrin Eberharter*		Madonnensaal (Theologie building) *Chair: Carol Spöttl*		SR VI (Theologie building) *Chair: Eva Konrad*
	1. AI for dynamic and diagnostic assessment: Automatic task design and mediation to support development of L2 English reading and writing Ari Huhta, Dmitri Leontjev, Roman Yangarber, Matthew E. Poehner		**9. Exploring the impact of test mode on test takers' turn management in paired discussion tasks** Yaqian Zhang, Yan Jin		**17. Validating Prompts and Rubrics in an Office-Hour Role-Play Task – a mixed method approach to local test reformation** Stephen Daniel Looney, Haoshan (Sally) Ren
	2. Diagnosing Chinese EFL Learners' Speaking Proficiency: A Machine Learning-Based Cognitive Diagnostic Modeling Approach Shuting Zhang, Lianzhen He		**10. Academic language socialization: Transforming research findings into a self-assessment/diagnostic tool for students and teachers** Heike Neumann, Saskia Van Viegen, Sandra Zappa-Hollman		**18. ChatGPT versus human raters in integrated writing assessment: Comparing rating performance across test taker levels and rating criteria** Haeyun Jin

	3. Process and Product in Diagnostic Assessment of Writing: What Do Experts See? Michelle Czajkowski	**11. Exploring English writing proficiency among 15-year-old students in Sweden** Eva Olsson, Linda Borger, Sofie Johansson	**19. Diagnosing L2 English Academic Reading Ability in the CEFR Context: A CDA Approach** Tugba Elif Toprak Yildiz, Claudia Harsch
	4. What Inferences can we Draw from Scores on Paired Discussion Tasks Delivered Through Spoken Dialog Systems? A Study on Construct-Relevant and -Irrelevant Factors Nazlinur Gokturk, Evgeny Chukharev	**12. AI-Supported Automated Scoring of Constructed Response Tasks for Second-Language Academic Reading Proficiency Assessment** Marcello Gecchele, Ahmet Dursun	**20. Writing assessment literacy and the factors shaping its development: the case of pre-service and in-service English and French second language secondary school teachers in Quebec** Amira Ben Hmida
	5. The Role of L1 in L2 Models Adopted to Assess L2 Learners' Writing Quality Ping-Yu Huang	**13. Developing a scenario-based test to assess the language assessment knowledge of EFL teachers in Chile** Salomé Villa Larenas	
	6. A Mixed-Methods Investigation into Raters' Perceptions and Challenges about Rating Prosodic Features Meng-Hsun Lee	**14. There are C-Tests and C-Tests: Digitalised Formats and Reduced Times - Changed Constructs?** Anastasia Drackert, Anna Timukova, Franziska Möller	
	7. Building a corpus of academic writing in EMI contexts: Exploring applications for language assessment Dana Gablasova, Luke Harding, Raffaella Bottini, Haoshan (Sally) Ren, Vaclav Brezina	**15. Indigenous Assessment Criteria in a Test of English for Tourism Students: Adopting Pill's (2016) Approach** Gina Elizabeth Ward	
	8. Accommodations in listening assessment: Exploring the effect of self-paced listening on test scores and anxiety of learners with differing L1 literacy skills Elisa Guggenbichler	**16. Exploring test takers' experiences with instructions in reading-into-writing tasks** Lies Strobbe, Goedele Vandommele, Sterre Turling	
3:00 - 3:30	Coffee Break (sponsored by Cambridge University Press & Assessment)		

Wednesday, July 3, 2024 (continued)

<table>
<tr><td>3:30 - 5:30</td><td colspan="4">Parallel Session 2</td></tr>
<tr><td></td><td>Aula</td><td>HS1</td><td>HS2
Chair: Ari Huhta</td><td>HS3
Chair: Tony Clark</td></tr>
<tr><td>3:30 - 4:00</td><td rowspan="4">Symposium: Cross-continental perspectives on language policies and practices for immigration and citizenship

Antony John Kunnan (Chair & Discussant), Cecilie Carlsen, Lorenzo Rocca, Kellie Frost, Coral Yiwei Qin, Eunice Eunhee Jang, Maryam Wagner, Jeanne Sinclair, Melissa Hunte</td><td rowspan="4">Symposium: Applying diagnostic assessment in AI-assisted language learning

Lianzhen He (Chair), Xiaoming Xi (Discussant), Shangchao Min, Hongwen Cai, Xunyi Pan, Wenzhi Chen, Liqing Qiao, Min Wang, Huiyang Shen, Zihui Zhang</td><td>Cooccurrence of Disfluency Features of L2 Speech across Proficiency Levels in Controlled and Spontaneous Tasks

Yulin Pan</td><td>Innovating constructs and assessments: The development and investigation of multimodal viewing-to-write tasks

Tineke Brunfaut, Judit Kormos</td></tr>
<tr><td>4:00 - 4:30</td><td>How reliable were human raters when assessing second language English prosody? A Bayesian meta-analysis

Yuanyue Hao</td><td>Investigating cognitive strategy use in an intertextual reading-into-writing Summary task through online think-aloud interviews

Nathaniel Ingram Owen, Haiyan Xu, Oliver Bigland</td></tr>
<tr><td>4:30 - 5:00</td><td>The validation and usability of an L2 Chinese prosody rating scale in three speaking task types

Sichang Gao, Mingwei Pan</td><td>Source use patterns in integrated writing tasks: The role of discourse synthesis quality and linguistic features

Atta Gebril</td></tr>
<tr><td>5:00 - 5:30</td><td>Engagement, emotional valence, and attention: Investigating the impact of facial behavior on speaking test scores

J. Dylan Burton</td><td>Sequence analysis of log data: an application example from a study of integrated writing

Ximena Delgado-Osorio, Valeriia Koval, Johannes Hartig, Claudia Harsch</td></tr>
<tr><td>5:30 - 7:00</td><td colspan="4">Special Sessions</td></tr>
<tr><td></td><td colspan="2">HS1</td><td>HS2</td><td>HS3</td></tr>
<tr><td>5:30 - 6:30</td><td colspan="2">Creatively Engaged and Recharged: A Session for Mid- and Senior-Career Professionals

Micheline Chalhoub-Deville, Mikyung Kim Wolf</td><td>Navigating the job market

Ute Knoch, Antony John Kunnan, Barry O’Sullivan, Paula Winke, Alistair Van Moere</td><td>How to be a (good) reviewer

Talia Isaacs, Elvis Wagner, Daniel R. Isbell</td></tr>
<tr><td>6:30 - 7:00</td><td colspan="2">Rainbow Connections (Location: SR 6)

Niles Zhao</td><td colspan="2"></td></tr>
<tr><td>7:00</td><td colspan="4">Networking dinners, meet your host in front of the main conference venue if you are signed up</td></tr>
<tr><td>8:30</td><td colspan="4">‘One for the road‘ - an evening at the Irish Pub to celebrate Jamie Dunlea, The Galway Bay Irish Pub (downstairs), Kaiserjägerstraße 4</td></tr>
</table>

Thursday, July 4, 2024

08:00 - 08:30	Registration (Foyer)				
8:30 - 10:30	**Parallel Session 3**				
	Aula	HS1 *Chair: Bart Deygers*	HS2 *Chair: Tineke Brunfaut*	HS3 *Chair: Mikyung Kim Wolf*	UR3 *Chair: Sari Luoma*
8:30 - 9:00	**Symposium: Open Science in Language Testing: Bridging Academic and Industry Perspectives** J. Dylan Burton (Chair/ Discussant), Paula Winke, Jason Fan, Jin Yan, Jieun Kim, Daniel R. Isbell, Spiros Papageorgiou, Karen Dunn, Geoffrey T. LaFlair	**Language testers as policymakers** Laura Schildt	**Construct relevant or irrelevant? The impact of background noise on listening comprehension** Xun Yan, Yan Tang	**Building an argument for test score interpretation and use for a fully automated online assessment of L2 spoken interaction** Yasuyo Sawaki, Yuya Arai, Masaki Eguchi, Shungo Suzuki, Yoichi Matsuyama	**Delayed measures of speaking proficiency: Questioning assumptions** Anastasia Ulicheva, Sumita Ishaque, Rose Clesham
9:00 - 9:30		**A Theory of Action in Working for Social Justice** Cecilie Hamnes Carlsen, Lorenzo Rocca, Nick Saville, Graham Seed	**Equality, Diversity, and Inclusion in Practice: Candidate Reactions to Global English Accents in a Listening Test** Gemma Bellhouse	**Automated scoring and validity: Expanding evidence through explainability** Sarah R. Hughes	**Analyzing Argumentative Skills in Foreign Language Learners: Integrated Task Assessments and Rhetorical Moves Analysis** Jorge Luis Beltran Zuniga
9:30 - 10:00		**Language and knowledge of society tests for citizenship: implications for vulnerable migrant groups** Marieke Vanbuel, Edit Bugge	**What makes listening comprehension difficult?: A feature-based machine learning approach to understanding item difficulty** Huiying Cai, Ping-Lin Chuang, Yulin Pan, Mingyue Huo, Xun Yan	**Evaluating score accuracy for an automated scoring system in a high-stakes writing test** Trevor Breakspear, Edmund Jones, Shilin Gao, Trevor Benjamin, Jing Xu	**Evaluating General Language Proficiency Speaking Test Assessment Criteria: Evidence From Non-Language Specialists** Curtis Gautschi
10:00 - 10:30		**The sufficiency question: untangling relevance, representativeness and sufficiency** Ute Knoch, Susy Macqueen	**The road to understanding in lecture listening: how students integrate auditory and textual information** Nicola Latimer, Daniel Lam, Chihiro Inoue, Sathena Chan	**Exploring two novel applications of Generative AI in Automated Essay Scoring** Jing Wei, Alistair Van Moere, Steve Lattanzio	**Conceptualizing and operationalizing the construct of critical thinking in EAP speaking: The development and validation of a rating scale** Shengkai Yin
10:30 - 11:00	Coffee Break (sponsored by g.a.s.t.)				

Thursday, July 4, 2024 (continued)

11:00 - 12:10	**Samuel J. Messick Memorial Lecture (Sponsored by Educational Testing Service)** (Location: Aula) **Reimagining validity in accountability testing: Understanding consequences in a social context** Micheline Chalhoub-Deville
12:10 - 2:00	Networking Lunch (sponsored by The Language Training and Testing Center (LTTC))
12:30 - 2:00	**ILTA Annual Business Meeting** (Location: HS 3)
2:00 - 3:30	**Posters**

A Digital Mapping of High Leverage Communicative Practices in School-Age Content-Area Contexts Lynn Shafer Willner

A Multifaceted Investigation on the Assessment of French Language Competence of K-12 Teachers in Canada Samira ElAtia, Komla Essiomle, Elissa Corsi, Pierre Rousseau, Danielle Dallaire

Augmented Assessment: Shaping EFL Speaking Assessment with Mobile AR Technology Jung-Hee Byun

ChatGPT in the Classroom: Pre-Service English Language Teachers' Perspectives on AI Integration in Language Assessment Training
Asli Lidice Gokturk-Saglam

Computerized Dynamic Reading Assessment as an Enhancer of Reading Development of Students with Lower Proficiency Chansak Siengyen, Punchalee Wasanasomsithi

Developing a new writing rubric as part of an exam reform project Mark Derek Chapman, Tanya Bitterman, Heather Elliott

Developing an efficient EAP placement test using integrated tasks to assess receptive and productive skills Rebecca Yeager, Alfonso Martinez

Evolving Modalities: Exploring Changes in Language Assessment Practices in Higher Education Michelle Reyes Raquel, Simon David Boynton, Wim Vergult, Grace Chang, Anne Hu

Examining the Writing Style of ChatGPT using AI-Generated Text Detection Peter Kim

Exploring Language Assessment Literacy: What do Taiwanese CLIL teachers need to learn and relearn? Yu-Ting Kao

Implementing a Learning-Oriented Academic Reading and Writing Assessment Model at a Tertiary Level in Thailand Punchalee Wasanasomsithi

Language testing and assessment academic production in Latin America: a bibliometric analysis Gladys Quevedo

Language testing and language policy change: A case study from Ukraine Karen Jeanette Dunn, Jamie Dunlea, Zhanna Sevastianova, Irina Umbetaliyeva, Martin Murphy

Measuring verbal and non-verbal features of L2 learners' spoken interaction: Rethinking automated speaking assessment Anna von Zansen

	Scoring validity of an AI-powered essay-scoring system for a task-based writing test Yoshihito Sugita **Test-Taker Insights in Language Assessment Literacy: The Road Less Travelled** Andy Jiahao Liu **The Process and Impact of Streamlining a Placement Test: Factor Analysis and Rasch Modeling in Practice** Jieun Kim, Maggie McGehee **Unveiling learners' perspectives during speaking disfluencies: Building learners' disfluency profiles across various proficiency levels in OPI assessment** Yu (Joyce) Wu, Qiaona Yu **Using ChatGPT as a tool for automated writing evaluation: impact on syntactic and lexical complexity** Bart Deygers, Liisa Buelens, Laura Schildt, Marieke Vanbuel **Virtual Administration of an Oral English Proficiency Test: Procedures, Challenges and Student Perceptions** Sharareh Taghizadeh Vahed				
3:30 - 4:00	Coffee Break (sponsored by Goethe-Institut)				
4:00 - 6:00	Parallel Session 4				
	Aula	HS1 *Chair: Fumiyo Nakatsuhara*	HS2 *Chair: Lianzhen He*	HS3 *Chair: David Wei Dai*	UR3 *Chair: J. Dylan Burton*
4:00 - 4:30	**Symposium: Reforming the Diagnosis of L2 Abilities: The Complementary Contributions of Dynamic and Diagnostic Language Assessment Frameworks** Dmitri Leontjev, Matthew E. Poehner (Chairs), Claudia Harsch (Discussant), Jie Zhang, Tianyu Qin, Lu Yu, Magdalini Liontou, Ari Huhta, Luke Harding, Tineke Brunfaut, Benjamin Kremmel	**A collaborative approach to examining BESTEP's impact on tertiary EAP in Taiwan** Jessica R. W. Wu, Heng-Tsung Danny Huang, Shao-Ting Alan Hung, Anita Chun-Wen Lin, Joyce Shao Chin, Ali Shuhsuan Ke	**Aligning Proficiency Level Descriptors with Audiences and Uses: Enhancing Equitable Communication in a K-12 Language Assessment System** Lynn Shafer Willner, Margo Gottlieb	**Analyzing the Variances in Two Test Administration Modes: Time for a change in the assessment paradigm?** Linda Nepivodova, Simona Kalova	**Exploring a new method for multi-lingual alignment of language frameworks: Developing a Global Scale for Multiple Languages Using Comparative Judgement** Ying Zheng, Booth David

Thursday, July 4, 2024 (continued)

	Parallel Session 4 (continued)				
	Aula	HS1 *Chair: Fumiyo Nakatsuhara*	HS2 *Chair: Lianzhen He*	HS3 *Chair: David Wei Dai*	UR3 *Chair: J. Dylan Burton*
4:30 - 5:00	**Symposium** (continued): **Reforming the Diagnosis of L2 Abilities: The Complementary Contributions of Dynamic and Diagnostic Language Assessment Frameworks** Dmitri Leontjev, Matthew E. Poehner (Chairs), Claudia Harsch (Discussant), Jie Zhang, Tianyu Qin, Lu Yu, Magdalini Liontou, Ari Huhta, Luke Harding, Tineke Brunfaut, Benjamin Kremmel	**Where the Lines are Drawn: A Survey of English Proficiency Test Use in Admissions among U.S. Research-Intensive Universities** Nicholas Coney, Daniel R. Isbell	**Investigating score reporting systems and practices: Content and genre analyses of parent versions of standardized language test score reports** Monique Yoder	**Assessment method reform: Examining the comparability of linguistic features of communication elicited in virtual and physical settings** Slobodanka Dimova	**Using AI to enhance JEDI: multilingual constructs to reform monolingual tests** Graham Seed
5:00 - 5:30		**Supporting Higher Education institutions through language assessment reform: Evaluating the impact of change on admissions tests** Tony Clark, Emma Bruce, Karen Ottewell	**Intersecting Voices: A Sociocultural Exploration of Test-takers' and their Parents' Experiences and Perceptions in English Tests of Young Learners** Jia Guo, Liying Cheng	**How does extended time affect dyslexic test-takers with different item types in an online English test?: An exploratory study** Chihiro Inoue, Lynda Taylor	**The role of policy actors' agency in test impact: Assessment of languages other than English in China's senior secondary education** Chenyang Zhang
5:30 - 6:00		**An investigation of the alignment of national language teaching policy with the advanced-level secondary school leaving examination in foreign languages in Hungary** Katalin Piniel, Gyula Tankó, Zsuzsanna Andréka	**Shedding Light on the Test-Taking Experiences of Francophone African Learners of English in High-Stakes English Proficiency Testing** Kadidja Koné, Paula Winke	**A literature review on the ordering of test components** Ramsey Lee Cardwell, Ben Naismith	**Comparing reading item difficulty: Does A1 equal A1?** Katharina Karges
6:00 – 7:00	*Language Assessment Quarterly* anniversary celebration (closed meeting, sponsored by Taylor & Francis), Hotel Bar Grauer Bär, Universitätsstraße 5-7				
6:30	Guided Walking Tour of Innsbruck, meet in front of the main conference venue if you are signed up				
7:30	*Language Testing* Editorial Board Dinner (closed meeting, sponsored by SAGE), Gasthaus Weißes Rössl, Kiebachgasse 8				
8:00	Graduate Student Assembly Biergarten Social, Bierstindl, Klostergasse 6				

Friday, July 5, 2024

08:00 - 08:30	Registration (Foyer)				
8:30 - 10:30	**Parallel Session 5**				
	Aula	HS1 *Chair: Jing Xu*	HS2 *Chair: Beverly Baker*	HS3 *Chair: Doris Moser-Frötscher*	UR3 *Chair: Kathrin Eberharter*
8:30 - 9:00	**Symposium: Locating competence, exploring constructs: Taking forward Tim McNamara's work in performance assessment** John Pill, Lynda Taylor (Chairs), Lynda Taylor (Discussant), William Agius, Susy Macqueen, Geisa Dávila Pérez, David Wei Dai	**Human-Centered AI for Test Development** Alina A. von Davier, Andrew Runge, Yigal Attali, Yena Park, Geoffrey T. LaFlair, Jacqueline Church	**Speaking of reform: introducing large-scale speaking assessment into a lower-secondary school system** Johanna Motteram, Jamie Dunlea, Barry O'Sullivan, Fumiyo Nakatsuhara, Akihiro Matsuura, Robin Skipsey	**Reforming sign language assessment: setting up a longitudinal learner corpus of rated elicited imitation performances to develop an AI-driven sign language assessment system** Franz Holzknecht, Tobias Haug, Alessia Battisti, Katja Tissi, Sandra Sidler-Miserez, Sarah Ebling	
9:00 - 9:30		**Humans vs. LLMs: How good are LLMs in generating input texts for reading tasks on B2/C1 levels of the CEFR?** Anastasia Drackert, Andrea Horbach, Anja Peters	**Nuanced approach to the English Language Examination Reform in Japan** Noriko Iwashita, Megan Yucel	**Automatic CEFR classification of written learner texts using Natural Language Processing** Torsten Zesch, Jeanette Bewersdorff, Josef Ruppenhofer	**An Online Diagnostic Assessment System for English Language Teaching and Learning at Schools, Colleges, and Universities** Yan Jin, Zunmin Wu, Liping Liu
9:30 - 10:00		**Can GPT write good items? Comparing item characteristics of human-written and GPT-4-written items** Yena Park, Jacqueline Church, Yigal Attali	**Developing an evaluation framework for proficiency testing for education and employment in Taiwan** Richard Spiby, Emma Bruce	**Fairness of TCF Writing using human raters and a hybrid automated rating model: from construct validity to psychometrics, to an argument-based approach** Vincent Folny, Rodrigo Souza Wilkens, Rémi Cardon, Thomas François	**Probing attribute structures in testlet-based listening assessment: An application of cognitive diagnostic models** Lidi Xiong, Lianzhen He
10:00 - 10:30		**Cloning Tasks with GPT Models for Automated Difficulty Estimation** Sylwia Macinska, Andrew Mullooly, Luca Benedetto, Hannah Bouteba, Mark Elliott	**Revising the ILTA Code of Ethics, and the impact of ethical consensus in the global language testing community** Bart Deygers, Meg Malone	**Use of a technology-assisted rating tool for assessing integrated English academic writing ability** Haeyun Jin	**Integrated Diagnostic Grammar Assessment: A Systemic Functional Linguistics Approach** Roz Hirch

Friday, July 5, 2024 (continued)

10:30 - 11:00	Coffee Break (sponsored by France Éducation International)				
11:00 - 12:00	**Parallel Session 6**				
	Aula *Chair: Heike Neumann*	HS1 *Chair: Daniel R. Isbell*	HS2 *Chair: Yan Jin*	HS3 *Chair: Talia Isaacs*	UR3 *Chair: Yasuyo Sawaki*
11:00 - 11:30	**"Context-limited" or "boundary-crossing"? The essential contribution of case study research in language assessment** Beverly Baker, Lynda Taylor	**Investigation of Differential Item Functioning Analyses Due to Multiple Manifest Grouping Variables: Rasch Perspective** Sanshiroh Ogawa, Hong Jiao	**Communicating ELP Assessment Changes to K-12 Educators** Ahyoung Alicia Kim, Lorena Alarcon, Jason Kemp, Fabiana MacMillan	**Exploring the moderating role of assistance in assessing speaking ability for argumentation** Jorge Luis Beltran Zuniga	**The Effects of Linguistic Features and Genre of Test Prompt as Predictors of College Writing Placement for L2 Students** Weejeong Jeong
11:30 - 12:00	**"Father brings books; son writes; mother worries; daughter volunteers." Gender representations in Chinese Gaokao English (2014-2023)** Xiaoqin Huang, Xiangdong Gu, Yong Wang	**The differential impact of COVID-19 on EL proficiency: unpacking language domains** Narek Sahakyan	**Lost in translation? Reporting the results of a CEFR linking study to educators** David MacGregor, Katie Schultz, Mark Chapman, H. Gary Cook	**You may say this better: Consequential validity evidence for diagnostic speaking assessment on lexical use** Shungo Suzuki, Hiroaki Takatsu, Ryuki Matsuura, Mao Saeki, Yuya Arai, Yoichi Matsuyama	**Comparative judgement as a foreign language assessment tool: an overview of the Crowdsourcing Language Assessment Project** Peter Thwaites, Magali Paquot
12:00 - 1:30	Networking Lunch (Sponsored by Oxford University Press)				
12.30 - 1:30	LAQ Editorial Board Meeting (closed meeting, location: SR 2)				
1:30 - 2:30	**SIG Sessions**				
	Aula	HS1	HS2	HS3	
	Automated Language Assessment (ALASIG) Jing Xu, Xiaoming Xi	**Integrated Assessment (IASIG) & Language Assessment Literacy (LALSIG)** Rebecca Yaeger, Xun Yan, Sharry Vahed, Elsa Fernanda Gonzalez, Gladys Quevedo-Camargo	**Test-taker Insights in Language Assessment (TILASIG)** Andy Jiahao Liu, Ray Jui-Teng Liao	**Language Assessment for Young Learners (YLSIG)** Mark Chapman, Veronika Timpe-Laughlin, Jeanne Beck	
2:30 - 3:30	**Cambridge/ILTA Distinguished Achievement Award Lecture (Sponsored by Cambridge University Press & Assessment/ILTA)** (Location: Aula) **Integration and Inclusiveness in Language Assessment** Antony John Kunnan				
3:30 - 4:00	***Closing*** (Location: Aula)				
6:00	LTRC Banquet and Awards Ceremony (Sponsored by British Council), Villa Blanka, Weiherburggasse 31 (drinks and music as of 5:30 pm)				

notes:

notes:

notes:

notes:

Plenary Sessions

Alan Davies Lecture

Sponsored by British Council

Experimenting with Uncertainty, Advancing Social Justice: Placing Equity, Diversity, Inclusion and Access Centre Stage

Prof. Lynda Taylor, University of Bedfordshire

Wednesday, July 3, 2024, 11:00 to 12:10 Location: Aula

Language testing has long been a locus for investigating the social context, consequences and power of assessment. During the 1990s, Professor Alan Davies (1931-2015) was among the first to focus our attention on the complex ethical dimensions of how we behave as language testing specialists. Davies drew on moral philosophy as a rich seam for mining core principles to inform and guide good professional conduct (1990, 1997). He was instrumental in helping to inspire and shape the ILTA Code of Ethics in 2000, and the ILTA Guidelines for Practice in 2007 which seek to instantiate ethical principles in terms of actual behaviours and practices. Both the Code of Ethics and the Guidelines for Practice have been revised and are kept under review in response to needs and changes within the professional field.

Since 2000 the field of language assessment has seen growing interest in matters of fairness, justice, ethics and social responsibility, leading to increased concern for equity, access and inclusion for all test takers, especially those from underserved communities. One such community includes those with specific language assessment requirements due to life circumstances, or to a disability or condition, whether temporary or permanent. Other communities are those who speak less commonly taught or spoken languages, especially marginalised languages. Principles of fair access and equitable treatment are now well-established in education and society, and the rights of minorities are increasingly enshrined in legislation. Despite this, however, advancing social justice through sound policy and good practice remains a challenge for the language testing profession in relation to knowing how best to address the aspirations and needs of test takers with special requirements.

In a Special Issue of Language Testing focusing on accommodations for test takers with disabilities, Taylor & Banerjee (2023a) highlighted the sensitive balance that language test providers have to maintain: between their professional commitment to test standardization, reliability and validity demands on the one hand, and a commitment to advancing equity of access and inclusion for all test takers, regardless of their circumstances. This is especially true where the latter may require a departure from standardized test procedures or risks compromising or undermining established test validity claims. Research findings to inform practical decisions about suitable

accommodations (e.g. use of extended time or digital aids) can be scarce, due to challenges encountered with empirical research in this area: population cohorts can be hard to identify/reach and sample sizes can be too small for quantitative analysis.

Given the LTRC 2024 theme of 'Reforming language assessment systems – reforming language assessment research', the conference offers a welcome opportunity to highlight positive advances in language testing ethics over the past two decades, but also to examine how we might need to refresh and reframe our thinking and our practice so as to advance social justice for the benefit not just of specific communities but society as a whole (Taylor & Banerjee 2023b). In my presentation I will reflect on some specific JEDI-related challenges that I believe we need to confront, e.g. concerning construct definition, stakeholder engagement, human rights in a digital world. My aim will be to explore how we address and resolve such challenges in an ethically principled and evidence-based way. Through this 2024 Davies Lecture, I also wish to pay tribute to Professor Alan Davies as a pioneer and leading light in the movement to advance ethical practice and social justice in language testing; someone who was willing to experiment with uncertainty and to live creatively with the tension between speculation and empiricism.

Lynda Taylor is Visiting Professor at the Centre for Research in English Language Learning and Assessment (CRELLA) at the University of Bedfordshire, UK. She has worked for many years in the field of language testing and assessment, particularly with IELTS and the full range of Cambridge English qualifications. Her research interests include speaking and writing assessment, test takers with special needs and language assessment literacy.

She was formerly Assistant Research Director with Cambridge Assessment English and has advised on test development and validation projects around the world. She has given presentations and workshops internationally, published extensively in academic journals, and authored or edited many of the volumes in CUP's Studies in Language Testing (SiLT) series. In 2022 she was awarded Fellowship of the UK Academy of Social Sciences and she is currently serving a second term as President of the UK Association for Language Testing and Assessment (UKALTA) (2023-2025).

Samuel J. Messick Memorial Lecture

Sponsored by Educational Testing Service

Reimagining Validity in Accountability Testing: Understanding Consequences in a Social Context

Prof. Micheline Chalhoub-Deville, University of North Carolina at Greensboro

Thursday, July 4, 2024, 11:00 to 12:10 Location: Aula

For the past 25 years, my research and professional endeavors have revolved around accountability testing systems and their implications for language testing. This involvement has resulted in publications addressing topics such as the nature of policy-mandated testing and validity conceptualization (e.g., Chalhoub-Deville, 2009a, 2009b, 2016, 2020; Chalhoub-Deville & O'Sullivan, 2020). In my presentation, I examine our current thinking and practices in this field and advocate for the adoption of alternate, socially-mediated theories to guide future efforts.

Accountability testing is closely tied to the rise of a pervasive Audit Culture, which became formalized in the U.S. through the enactment of No Child Left Behind. This accountability movement, which extends globally has been referred to as the Global Education Reform Movement (GERM). GERM systems necessitate a reevaluation of established validity frameworks and methodologies to encompass aggregate scores and socio-educational contexts. With accountability testing, educators and educational institutions are held responsible for student performance. Our validation approaches, however, primarily target student scores (e.g., Yumsek, 2023), overlooking aggregate scores and the broader contexts where crucial interpretations and decisions are made. Consequently, accountability testing underscores the need for adjustments in validation models to incorporate aggregate scores and to consider the broader educational and societal contexts of testing (Chalhoub-Deville, 2016, 2020; Hoeve, 2022).

Sources such as the Standards for educational and psychological testing (AERA, APA, & NCME, 2014) and the published literature tend to zero in on fairness, which deals broadly with issues of accommodations, differential item functioning, and universal design. While these efforts serve a critical role in improving practices and score inferences, the broader implications of accountability testing, i.e., socio-educational consequences tend to remain outside the purview of test publishers' research agendas. Research engagement with the broader socio-educational dimensions of fairness and consequences needs to be conceptualized as a shared responsibility among key stakeholder groups, including test publishers. In conclusion, this paper advocates for a shift towards conceptualizing GERM and related accountability testing within a broader socio-educational framework and suggests different models to help achieve that goal.

Micheline Chalhoub-Deville holds a Bachelor's degree from the Lebanese American University and Master's and Ph.D. degrees from The Ohio State University. She currently serves as a Professor of Educational Research Methodology at the University of North Carolina at Greensboro (UNCG) where she teaches courses on language testing, validity, and research methodology. Prior to UNCG, she worked at the University of Minnesota and the University of Iowa. Her professional roles have also included positions such as Distinguished Visiting Professor at the American University in Cairo, Visiting Professor at the Lebanese American University, and UNCG Interim Associate Provost for Undergraduate Education.

Her contributions to the field include publications, presentations, and consultations on topics like computer adaptive tests, K-12 academic English language assessment, admissions language exams, and validation. She has over 70 publications, including books, articles, and reports, has delivered more than 150 talks and workshops. Additionally, she has played key roles in securing and leading research and development programs, with a total funding exceeding $4 million. Her scholarship has been recognized through awards such as the ILTA Best Article Award, the Educational Testing Service—TOEFL Outstanding Young Scholar Award, the UNCG School of Education Outstanding Senior Scholar Award, and the national Center for Applied Linguistics Charles A. Ferguson Award for Outstanding Scholarship.

Professor Chalhoub-Deville has served as President of the International Language Testing Association (ILTA). She is Founder and first President of the Mid-West Association of Language Testers (MwALT) and is a founding member of the British Council Assessment Advisory Board--APTIS, the Duolingo English Test (DET) Technical Advisory Board, and English3 Assessment Board. She is a former Chair of the TOEFL Committee of Examiners as well as a member of the TOEFL Policy Board. She has participated in editorial and governing boards, such as Language Assessment Quarterly, Language Testing, and the Center for Applied Linguistics. She has co-founded and directed the Coalition for Diversity in Language and Culture, the SOE Access & Equity Committee, and a research group focused on the testing and evaluation in educational accountability systems. She has been invited to serve on university accreditation teams in various countries and to participate in a United Nations Educational, Scientific, and Cultural Organization (UNESCO) First Experts' meeting.

Cambridge / ILTA Distinguished Achievement Award

Sponsored by Cambridge University Press & Assessment / ILTA

Integration and Inclusiveness in Language Assessment

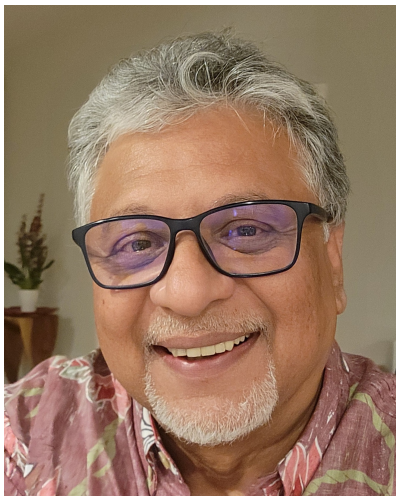
Dr Antony John Kunnan, Carnegie Mellon University

Friday, July 5, 2024, 2:30 to 3:30 Location: Aula

In considering how to make language assessment more integrated and inclusive, I believe two areas need our attention. They are (a) understanding that language assessment is part of language teaching and learning and (b) including language learners/test takers from the Global South in assessment policies and practices.

In regard to (a), in the last decade, a survey of the research literature shows that the focus of the field has moved from fairness and validation of assessments to an understanding of the role of language assessment integrated within language teaching and learning. This underreported field is Learning-Oriented Language Assessment (LOLA) based on substantial research (for general definitions see Black and Wiliam, 1998; Carless, 2007, 2015; Gebril, 2021; Jones and Saville, 2016; Pellegrino et al., 2001; Purpura, 2004, 2014. 2021; Saville, 2021; Turner and Purpura, 2016). And yet, the frameworks and principles from LOLA have not entered contemporary language assessments systematically. For example, following Purpura (2021), very few of the performance moderators (instructional, socio-cognitive, affective, social-interactional, and technological factors) have been conceptualized and operationalized in local, national and international assessments. This primary goal of this reforming activity would be to integrate assessments with learning so that assessments could contribute to building language learners' capability.

In regard to (b), for the last seven decades, international assessments have provided assessments to migrants from the Global South (GS) who are forced to take language tests in in order to assist institutions in the Global North (GN) with school and university admission, employment, residency and citizenship. Illustrating this point from the perspective of international English language assessments, providers have desultorily considered issues related to English as Lingua Franca (ELF) (for general definitions and discussions of ELF, see Brown, 2014; Canagarajah, 2006; Jenkins and Leung, 2014, 2024; Ockey and Hirsch, 2020, and Harding, 2022). Inspirational works recently include the development of an ELF construct (Harding and McNamara, 2018) and a demonstration test (Ockey and Hirsch (2020)). Following the latter study, issues that need to be addressed include rhetorical sensitivity, context sensitivity, international communication competence, grammatical and lexical appropriacy, and discourse sensitivity. Such a plan would ensure a more inclusive and representative language assessment, derived with dialogue with stakeholders such as students, faculty, community members as well as higher education institutions in the GS. This reforming activity would help GN institutions understand the language capabilities and needs of GS language learners/test takers better.



Dr Antony Kunnan has had a long and distinguished career in language testing and assessment, covering many parts of the globe. This was reflected in nominations we received from a diverse range of contexts. His accomplishments across the five criteria on which the award is judged were clearly outstanding. Some particular highlights include that Dr Antony Kunnan was the founding editor of Language Assessment Quarterly; the founding President of the Asian Association for Language Assessment; the editor of the four-volume Companion to Language Assessment; a former ILTA President, Vice President and Treasurer; and a prominent thinker in conceptualising test fairness. He is a well-known and widely-respected member of our academic community, but beyond that – in the words of one of our committee members – Dr Antony Kunnan has been a "change maker", opening up new avenues for publication, engagement and thought in our field.

Dr Antony Kunnan graduated from the University of California Los Angeles (UCLA) in 1991, and his trajectory was already marked with distinction when he won the Jacqueline Ross TOEFL Dissertation Award for his research. This study was subsequently published in the Cambridge University Press Studies in Language Testing series as Test-taker characteristics and test performance: A structural modeling approach. Since that time, Dr Antony Kunnan has held academic posts/professorships at California State University, Los Angeles, the University of Hong Kong, the American University in Armenia, Nanyang Technological University, Tunghai University, Guangdong University of Foreign Studies, and the University of Macau, as well as visiting posts at the University of California, Los Angeles, and Chulalongkorn University. He recently took up a position as Principal Assessment Scientist at Duolingo and a Senior Research Fellowship at Carnegie Mellon University. He has published 11 books and over 80 articles or book chapters. His initial work on validity and structural equation modelling has led to a more recent focus on ethics, fairness and policy. During his career, Dr Antony Kunnan has given over 125 invited talks and workshops across 36 countries, demonstrating his considerable international reputation and impact.

Pre-Conference Workshops

Workshop A: Measurement Approaches to Exploring Survey Ratings and Rater Effects

Stefanie A. Wind, University of Alabama

Day 1: Monday, July 1, 2024, 9am to 4pm Location: SR1
Day 2: Tuesday, July 2, 2024, 9am to 4pm

Learning Objectives

After completing day one of this workshop, participants will be able to:

- Describe the major characteristics and theoretical motivations for polytomous measurement models that can be applied to survey ratings, including Rasch models, Mokken scaling models, and polytomous non-Rasch IRT models.
- Make informed decisions about which modeling approaches may be appropriate for various survey research purposes and contexts.
- Use pre-written code to estimate and extract key item- and person-related results from measurement models for survey ratings.
- Interpret results from measurement model analyses to make informed decisions about item functioning, including rating scale functioning for individual items.

After completing day two of this workshop, participants will be able to:

- Describe major types of rater effects from a measurement modeling perspective.
- Make informed decisions about which modeling approaches may be appropriate for evaluating ratings in various language assessment contexts
- Use pre-written code to estimate and extract measurement model results related to various rater effects
- Interpret results from rater analyses based on measurement models to make informed decisions about rating quality.

Stefanie A. Wind is an Associate Professor of Educational Measurement at the University of Alabama. She received her PhD in Educational Measurement from Emory University. Her primary research interests include the exploration of methodological issues in the field of educational measurement, with emphases on methods related to rater-mediated assessments, rating scales, latent trait models, and nonparametric item response theory. Her publications appear in methodological journals in the field of educational measurement as well as applied journals. She has authored and co-authored several books, including Exploring Rating Scale Functioning for Survey Research, Rasch Measurement Theory Analysis in R, and Invariant Measurement with Raters and Rating Scales. She has received awards for her research, including the Alicia Cascallar early career scholar award from the National Council on Measurement in Education and the Georg William Rasch Early Career Scholar award from the American Educational Research Association.

Workshop B: Coding qualitative verbal protocol data for test validation

Andrea Révész, University College London

Monday, July 1, 2024, 9am to 4pm Location: SR2

Increasingly, analysing and coding qualitative data generated through introspective methods (e.g. think-alouds, stimulated recall) constitutes a key component of test validation projects, reflecting the growing recognition of the benefits of mixed-methods approaches to investigating test-takers' cognitive processes. The aim of this workshop is to make participants familiar with the major steps involved in coding qualitative verbal protocol data, that is, organizing and classifying raw data into categories for the purpose of further analysis and interpretation. We will primarily focus on theory-driven coding methods, but we will also consider qualitative coding that emerges bottom-up from the data.
We will begin the workshop with discussing the concepts of validity and reliability in relation to coding qualitative verbal protocol data. Then, we will review the various steps involved in coding and consider strategies that can help increase the validity and reliability of the coding process at each stage.

We will focus on the following specific steps:

1. Selecting verbal protocol data for coding, that is, how many and which part of the data to code.
2. Preparing data for coding, for example, planning the level of detail to include in the transcription process.
3. Deciding whether to adopt or adapt an existing coding scheme or develop one's own.
4. Selecting and training coders to ensure accurate and consistent coding.
5. Checking the reliability of coding, calculating coder reliability, and dealing with disagreements between coders.
6. Reporting coding procedures.

When considering each of these steps, workshop participants will have the opportunity to apply the concepts covered to sample coding situations and datasets, taken from a variety of L2 assessment projects. Participants will also gain hands-on experience with coding verbal protocol data collected in studies investigating L2 users' cognitive processes during test performance. In particular, workshop activities will include coding test-takers' think-aloud/stimulated recall comments describing/recalling their thoughts during listening, speaking, reading, and writing assessments as well as integrated testing tasks involving a combination of skills and/or modalities. We will also look at ways of triangulating various data sources at the coding stage to obtain a fuller and more complete picture of the cognitive activities in which L2 users are involved during test performance. The workshop is intended for graduate students and researchers who are interested in validation research but have little experience coding qualitative verbal protocol data.

Andrea Révész is a Professor of Second Language Acquisition at the IoE, UCL's Faculty of Education and Society. Her main research interests lie at the interfaces of second language acquisition, instruction, and assessment, with particular emphases on the roles of task, input, interaction, and individual differences in SLA. Currently, she is also working on projects investigating the neurocognitive processes underlying second language speaking and writing performance. She is co-winner of the 2017 TBLT Best Research Article Award and co-recipient of the 2018 TESOL Award for Distinguished Research. Currently, she serves as associate editor of Studies in Second Language Acquisition, co-editor of the John Benjamins Task-based Language Teaching series, and past president of the International Association for Task-based Language Teaching (TBLT).

Workshop C: Policy literacy: Exploring effective participation for researchers in policy making

Joseph Lo Bianco, Professor Emeritus at the University of Melbourne
Mina Patel, British Council

Tuesday, July 2, 2024, 9am to 4pm Location: UR3

Abstract

- What does it mean to become policy literate?
- What are the benefits of developing policy literacy?
- What are the disciplines, and key readings, that inform us on how to 'read' policy?
- How can we develop skills (knowledge, awareness, capability) to participate in and help direct policy settings?
- What is the relationship between knowledge or expertise and power or authority?
- What are effective ways to link research to decision-making authority and decision-making processes?
- What is a 'problem' as understood in policy making circles?
- How can we plan and conduct research or project work in such a way as to enhance its traction in policy?

The sequence of activities undertaken as part of this Policy Literacy workshop will delve into both theoretical and applied literature to pose the above questions. We will explore the nuances and character of the link between research, or scholarship, and policy. We will investigate 'propitious moments' in the policy process where it is most beneficial to have an impact on policy settings. We will look at the special role of 'program evaluation' in shaping future policy. We will identify the various players (roles associated with influence, knowledge, or power) in the processes of policy making. We will look to implementation as a critical point where policy designs sometimes/often falter and distinguish some of its features.

The aim of this is to stimulate us to think about our own work in relation to the question of policy; who makes policy, how is it made (conceived, designed, implemented and evaluated).

There will be a specific focus on language (teaching, learning and assessment) and its special role in policy, and some inherent challenges that are special if not unique to making language an object of policy.

A key assumption in this workshop is to explore the nature of language problems.

- Who determines what is a language problem?
- Why do some language problems get elevated to public policy attention and not others?
- What is the process of agenda setting that is most amenable to input from researchers?
- What is a policy window? How can researchers most effectively open such windows?

Are language problems objective, or easily discernible and agreed realities which we need to merely discover and then resolve? Or are problems more fluid, capable of being understood and 'represented' in different ways? What is the nature of language problems and how is this question relevant to having research and expertise impact on policy processes?

While the core component of the entire sequence is English, its present and predicted position in the world, policy examples, readings and activities will range beyond English into the wider communicative context. Therefore, we will address issues of language ecology, the Dominant Language Constellation, language repertoire, testing and assessment, etc.

The workshop will involve input from the presenters and participation through exercises and discussion. There will be some pre-workshop reading and an exercise and a post-workshop exercise and reflection.

Dr Joseph (Joe) Lo Bianco is professor emeritus in language and literacy at the Faculty of Education, University of Melbourne, where until 2020 he was Chair Professor. He is a language policy specialist combining academic research and hands-on policy engagement. His theoretical and analytical studies of language problems and policy solutions have been conducted in many parts of the world especially in South and Southeast Asia, Oceania, North America, Europe and in some African countries (Ethiopia, South Africa and Tunisia). Over three decades he has led multi-country language problem solving teams in Southeast Asia and advised the EU/Council of Europe and UN on multi-country language planning.

He has been commissioned on this work by UNICEF, UNESCO, national governments and conducted assignments under World Bank funding. He has collaborated with the British Council on many occasions.

For several years he ran senior policy official and politician training in Bangkok for Asian officials, including the participation at ministerial level, attached to the UNESCO/UNICEF Mother Tongue Based Multilingual Education process of the region.

Mina Patel is Head of Research – Future of English at the British Council. Her background is in English language teaching and training. She has worked in the UK, Greece, Thailand, Sri Lanka and Malaysia as a teacher, trainer, materials developer and project manager for ELT (English Language Teaching) projects and has extensive experience working with ministries of education in East Asia. Mina has presented at numerous national and international conferences on ELT-related topics. Her academic interests lie in the areas of language assessment literacy, teacher education and development and qualitative research methodology. She is currently a PhD student at the *Centre for Research in English Language Learning and Assessment (CRELLA)* at the University of Bedfordshire, UK. Most recently, Mina has co-authored Future of English: Global Perspectives.

Workshop D: Generative AI for content generation and automated scoring: no-code and low-code solutions

Alistair Van Moere, MetaMetrics Inc and University of North Carolina at Chapel Hill
Jing Wei, MetaMetrics Inc

Tuesday, July 2, 2024, 9am to 4pm Location: SR2

With the advancement of generative AI technology, such as GPT, test development organizations and language testing researchers have started to explore the potential of leveraging generative AI to automate aspects of test development. One example application is the automated generation of test contents, such as reading comprehension passages at target CEFR levels, multiple-choice questions, short-answer questions, and graphics. A second application is using prompt-engineering so that the large language model directly generates scores and feedback for essays. This bypasses the usual automated essay scoring approach wherein expert human raters provide ratings for a sample of essays and models are developed to predict human scores.

The purpose of this workshop is to provide an introduction to the large language models (LLMs) that are available for automating content generation and automated scoring of writing. We will introduce the best practices for writing prompts in ChatGPT, refining prompts iteratively to get the intended outcomes, and building scalable production solutions through API interactions. This means that we will (a) use freely available online interfaces such as ChatGPT and Claude.ai to try out prompts, and (b) use Python code, which we will provide, so that we can instruct large language models to complete tasks at scale, e.g., for hundreds of essays, rather than one at a time.

We will focus on applying Generative AI technology for the following tasks:

- Generate test content and test items
- Score student essays
- Provide qualitative feedback on student essays

Participants will be asked to apply different prompt engineering approaches (e.g. few-shot and zero-shot learning) to:

- Write reading comprehension passages at different CEFR levels
- Generate multiple-choice, cloze, and short-answer questions for reading comprehension passages
- Generate graphics as a visual support for reading comprehension passages
- Assign scores to student essays that are aligned to writing rubrics
- Discuss the best approaches to prompt engineering
- Critique each other's prompts
- Compare the pros and cons of different LLMs
- Discuss the ethical implications of using LLM in test development

This workshop will be suitable for participants who have played with ChatGPT but who want to take their usage to a more professional level, and who have little or no experience of using Python or APIs. In advance of the workshop, participants will be expected to set up an account and payment method for OpenAI's API access (instructions will be provided).

Alistair Van Moere drives innovation in educational AI and assessments, and manages the Lexile Framework which reaches 35 million students every year. Before joining MetaMetrics, Alistair was President at Pearson, where he managed artificial intelligence scoring services for tens of millions for students in speaking and writing programs. He has worked as a teacher, university lecturer, assessment developer and ed-tech executive, in the US, UK, Japan, and Thailand. He has an MA in Language Teaching, Ph.D. in Language Testing, and an MBA, and as authored over 20 research publications on assessment technologies.

Jing Wei is responsible for shaping MetaMetrics' AI capabilities and impact, providing thought leadership on using AI to drive business values, and integrating innovative solutions into MetaMetrics' existing and future products. Prior to joining MetaMetrics, Jing served as a research scientist at the Center for Applied Linguistics, leading the development and validation of a portfolio of high-stake digital assessments used by millions of students every year. Jing brings 15+ years of experience in test development, measurement and statistics, machine learning, and product development. She holds a bachelor's degree in English from Shanghai International Studies University, a M.Phil. in Second Language Education from University of Cambridge, and a Ph.D. in Language Testing from New York University.

Special Sessions

Creatively Engaged and Recharged: A Session for Mid- and Senior-Career Professionals

Organizers: Micheline Chalhoub-Deville, University of North Carolina, Greensboro
Mikyung Kim Wolf, Educational Testing Service (ETS)

Wednesday, July 3, 5:30 pm to 6:30 pm Location: HS1

Following the exceedingly positive feedback received from attendees at the 2023 LTRC, the organizing committee for LTRC 2024 has invited us to host this session once again. We warmly welcome mid- and senior-career professionals, including those who participated last year, to join us for a session designed to reconnect and reinvigorate all attendees.

This special session aimed at mid- and senior-career professionals in the language testing field is designed to be a dynamic and reenergizing experience. The goal is to create an environment that sparks creative thinking and generates fresh ideas. The session will include a panel of experts who will share their experiences and perspectives about their professional journeys and the language testing field. It will also feature interactive activities and stimulating discussions. Participants will have the opportunity to network with peers and learn from each other, enhancing their professional development.

Panel

Liying Cheng, City University of Macau, SAR China
Atta Gebril, The American University in Cairo, Egypt
Benjamin Kremmel, University of Innsbruck, Austria
Lynda Taylor, University of Bedfordshire, UK

Navigating the Job Market

Organizers: Ute Knoch, University of Melbourne
Antony John Kunnan, Carnegie Mellon University
Barry O'Sullivan, British Council
Paula Winke, Michigan State University
Alistair Van Moere, MetaMetrics Inc.

Wednesday, July 3, 5:30 pm to 6:30 pm Location: HS2

This session is tailored for graduate students, recent graduates and emerging scholars seeking guidance in entering the job market or new professional opportunities. The speakers will answer your questions and report on their professional journeys to provide insights from both the academic and industry job market. They will talk about their experiences in both mentoring and preparing individuals for the job market, as well as hiring individuals in their various roles and what is important from an employer's perspective. They will explore effective strategies for job hunting, and give tips on

networking, strengthening one's profile and CV, honing your application, or standing out in job interviews. The session also delves into the evolving landscape of language assessment (research) careers, shedding light on emerging opportunities and industry expectations. Engaging discussions and real-world insights shared by experienced professionals aim to equip attendees with the practical tools and knowledge needed to navigate the transition from academia to a successful career in language assessment.

How to Be a (Good) Reviewer

Organizers: Talia Isaacs, University College London, Co-Editor Language Testing
Elvis Wagner, Temple University, Co-Editor Language Assessment Quarterly
Daniel R. Isbell, University of Hawai‘i at Mānoa, Associate Editor Language Learning

Wednesday, July 3, 5:30 pm to 6:30 pm Location: HS3

The first part of this interactive session will center on how to write a useful review and challenges in the peer review enterprise from editors', reviewers', and authors' perspectives. We'll address the importance of and strategies for writing insightful, constructive, and professional peer reviews with implementable feedback for authors. We will draw on Committee for Publication Ethics (COPE) guidelines on ethical reviewer practices pre-, during, and post-publication. We will also cover the use of generative AI tools in the peer review process and ethical implications (e.g., should not be used to craft reviews). Next, we will discuss different types of peer review (e.g., single-/double-blind, open, etc.) and panel attendees' perspectives on where they think journals in our field should be headed. Finally, we will discuss initiatives to better induct and support new peer reviewers in our field, including but not limited to introducing cross-journal guidelines for reviewers across language testing journals and potentially introducing an LTRC abstract mentoring scheme, where invited junior reviewers are paired with more senior reviewers and jointly evaluate abstracts for future LTRCs.

Rainbow Connections

Convenor: Niles Zhao

Wednesday, July 3, 6:30 pm to 7:00 pm Location: SR6

This special networking session is for conference attendees who self-identify as outside cis heteronormativity and/or LGBTQIA+ individuals. Here in Innsbruck, a city with its own unique cultural diversity, we aim to create a safe and welcoming environment inspired by the inclusive spirit of the LGBTQIA+ community. Our goal is to facilitate informal connections, offering a space where we can meet, connect, and support each other. Our only agenda is to foster engagement among our community members and provide a networking space.

Please note that this session is intended for those who self-identify as LGBTQIA+.

Opening Symposium

Advancing Fairness and Justice in Language Testing: Reflecting on Tim McNamara's Scholarship

Chair: Luke Harding

Tuesday, July 2, 6:00pm to 7:00pm Location: Aula

Presentations of the Symposium

Knowledge and Power: Shaping Policy Action from Research

Joseph Lo Bianco (University of Melbourne)

Over many decades of close work with Tim (Lo Bianco, 2019) I had innumerable discussions with him about how research could impact more closely on public policy determinations. Some of these discussions were transactional, such as how a specific research project he was engaged in, or a group of individuals whose language problems he was researching, could be supported in decision making within bureaucracies. Others looked at the intellectual presuppositions of the dynamic in which scholarly knowledge and political power come to interact with each other. Over time we came to delineate limits and problems associated with the aspiration of linking scholarly activity with practical action and identified important justifications for distance as well as engagement. I will describe these interactions as a 'policy conversations', and how they formed a kind of preparation for the culture of engagement between knowledge and power around dynamics of tensions and difficulties in epistemic orientation, purpose, timeframes and emotional investment between scholars and officials.

The Tension Between Conformity and Creativity in English-as-a-Lingua-Franca Communication

Barbara Seidlhofer (University of Vienna), Henry Widdowson (University of Vienna)

The range of Tim McNamara's research encompassed issues relating both to conformity and creativity. His work on language testing necessarily has to do with the measurement of competence, with what normalized conventions of usage it is possible or appropriate for learners to conform to as a measurement of competence in a language. But he also wrote insightfully about the necessary non-conformist uses of language in the creative realization of individual identity. Such uses, which of course would normally be negatively assessed in tests of (English) language competence, make clear that language users have the intrinsic capability of exploiting language as a communicative resource beyond the confines of communal convention. Such a capability is also what is abundantly evident in the use of English as a lingua franca. In our conversations with Tim, in correspondence and particularly when he was guest professor in Vienna, we argued that this implied a crucial distinction between justice as an institutionalized measure of conformity to convention, and fairness as a recognition of capability. We further discussed whether and to what extent the institutional requirement for reliable competence measurement was inherently unfair in that it denied recognition to how 'English' is capably and resourcefully used as a global means of lingua franca communication.

Interrogating the Social and Political Values Underlying Language Testing Practices

Kellie Frost (University of Melbourne), Ute Knoch (University of Melbourne), Susy Macqueen (Australian National University), Jason Fan (University of Melbourne)

A consistent theme of Tim's scholarship was his recognition and interrogation of the social and political values underlying language testing practices, particularly the way these values are manifest in test constructs and supported by the technical, measurement-related qualities of testing instruments. This theme encompasses an understanding of a complex and contested relationship between measurement fairness (absence of bias), and the fairness-as-justice questions associated with test constructs and uses. In this presentation, we begin by reporting on two recent projects inspired by this theme of Tim's work. The first focuses on a revised version of the Australian citizenship test, and considers the fairness implications of test design and use in reference to concerns raised about an earlier version of the test by McNamara and Ryan (2011). The second features a national project in China, aimed to investigate the practices of English test providers and survey test stakeholders' perceptions, views, and attitudes towards English language testing. We conclude by reflecting on Tim's theorizing of fairness and justice in language testing, particularly his efforts to bring a poststructuralist perspective to bear on questions of social values, constructs, and consequences, the mutually constituting, entangled nature of measurement-ideology this perspective implies, and its implications in light of recent calls for a renewed criticality in language testing (e.g. Randall et al., 2024) and in applied linguistics more generally (e.g. Kubota, 2022; Pennycook, 2022).

Power and Justice in Language Testing Embedded in 'The Meaning of Life'

Elana Shohamy (Tel Aviv University)

The issue of the power of tests - the uses of tests and their impact on learning, teaching and future lives of test takers - was always very central in the continuous conversations that Tim and I had over the course of his life. It was also an integral part of a longer list of items on our agendas, the last of which we never really reached entitled 'the meaning of life'. Yet, in fact, every conversation with Tim was about 'the meaning of life', as each conversation generated deep ideas and implications to a meaningful life. When I first met Tim, he was mostly involved in psychometric topics. It was a surprise for him that I was interested in test uses, misuses and people who are victims of tests. After all, we testers may be responsible for it. In this paper I will track the ideas, thoughts, research, writings and impact of Tim McNamara on multiple dimensions of language tests' uses, power, ethicality and justice. From a memorable encounter with a testing victim in a bar in the Netherlands, to our mutual work on language tests for citizenship, the influence of Derrida's writings on Tim's work, to broader issues addressed in his book, *Language and Subjectivity* (2019). Tim's work was/is meaningful and critical for test takers and society at large; it will continue to be inspirational for all of us in the years to come.

Symposium 1

Cross-continental perspectives on language policies and practices for immigration and citizenship

Chair(s): Antony John Kunnan, Carnegie Mellon University

Discussant(s): Antony John Kunnan, Carnegie Mellon University

Wednesday, July 3, 2024, 3:30 to 5:30 Location: Aula

The 20th and 21st centuries have seen unprecedented numbers of people forcibly displayed (UNHCR; https://www.unhcr.org/us/global-trends), something which have led several countries in the Global North to introduce measures to restrict migrants' access to entry, residency and citizenship. While in some countries policies and practices regarding immigration and citizenship have favored inclusion, tolerance, and diversity, others have introduced policies leading to exclusion, discrimination, and racism. Language requirements have been central to these policies, as part of an agenda of inclusion and integration or as part of a (hidden) agenda of gatekeeping and exclusion. When different countries are compared, a large degree of divergence is revealed regarding the level of language proficiency required for the same context, ranging from no requirements or only a very basic level in some countries (A1 on the CEFR-scale) achievable for the large majority of migrants to an academic level (B2) excluding all but highly educated migrants, in others (Rocca et al., 2020).

Fifteen years ago, Shohamy and McNamara (2009) in the first special issue on this topic in Language Assessment Quarterly, posed questions that language policy and assessment researchers were asking at that time: Does language present a valid and fair criterion or an excuse for a form of ethnic cleansing and expulsion of unwanted immigrants? Is it realistic to require immigrants to acquire a new language at a later point in their lives? How do these tests relate to immigrants who are not literate in their first language? What level of proficiency is needed for proper functioning? These questions were addressed by the first wave of researchers examining top-down matters: tests and test fairness and validation (examples, Blackledge, 2009; Schüpbach, 2009, Kunnan, 2009a, 2009b; Gysen et al., 2009). Many of these questions are still relevant but a wider range of issues with a broader set of instruments have emerged. These studies have focused on low-literate migrants (Carlsen and Rocca, 2021); testing language and testing culture (Slade and Möllering, 2010), transition from fairness to social justice (Frost, 2019), migrants' linguistic trials and negotiations (Khan, 2019), linguistic integration of adult migrants (Rocca et al., 2020), and countering the nationalist narrative for a subaltern-immigrant perspective (Kunnan, 2021). In short, the current thinking has moved from top-down to bottom-up studies; from examining test fairness and validation to one of critical language assessment and social justice.

The five studies that will be presented in the symposium will advance our knowledge of this topic from the critical language testing and the social justice perspective directly

learning from immigrants. The authors interrogate various narratives: (1) on the value of language requirements for immigrants and/or citizens, the local community and the country; (2) on the concept of language attainment and immigrant integration; (3) nationalist narratives vs. immigrant-subaltern voices; and (4) language needs in the less skilled and professional workplace. The authors are from three continents providing cross-continental perspectives on language assessment policies and practices. Research findings come from document and survey questionnaire analyses, nationalist narrative data, and listening to first-person immigrant voices.

Presentations of the Symposium

Language testing for residence and citizenship in Europe – a cross-national study

Cecilie Hamnes Carlsen (Høgskulen på Vestlandet), Lorenzo Rocca (Independent)

Most European countries today set formal language and knowledge of society (KoS) requirements for adult migrants applying for permanent residence and citizenship. This apparent convergence of requirement policy however covers a considerable divergence in terms of the relative strictness of the requirements set in different countries (Rocca et al., 2020). While the implementation of these requirements is typically justified by policy makers to enhance migrants' motivation for language and KoS learning, thereby facilitating integration, there is limited empirical evidence regarding the impact of this policy on different migrant groups and in countries with varying degrees of policy strictness (Strik et al., 2010; van Oers, 2014). Notably, for low-literate adult migrants (LESLLA learners) who have been deprived the fundamental right to schooling in childhood, such requirements may represent an unsurmountable barrier and a significant impediment to equal rights and opportunities (Kurvers & van de Craats, 2007; Minuz, 2017).

This study, situated within the framework of critical language testing (Shohamy, 2001; McNamara & Roever, 2006) and Messick's validity framework emphasizing the social consequences of tests (Messick, 1989), investigates the perceived effects of language and KoS requirements on LESLLA learners in 20 European countries. The severity of language policies across these countries is gauged using the Language Policy Index for Migrants (LAPIM) scale (Carlsen & Rocca, 2023). Employing a mixed-methods approach, the study gathers LESLLA teachers' perspectives on the impact of these requirements on their learners through an electronic survey (n=1079) conducted in 20 European countries as well as through interviews with teachers in 8 countries (n=40). The analysis of teachers' opinions is contextualized with the policy strictness in their respective countries. This research aligns with an ethically conscious language assessment paradigm, aiming for equity and social justice in accordance with the human rights values endorsed by the Council of Europe.

An investigation of language practices in the workplace: Problematising English language testing for skilled migration in Australia

Kellie Frost (University of Melbourne)

Skilled migrants with English as an additional language (EAL) face poorer than average employment outcomes in Australia, despite possessing high level technical skills in areas of high demand. These outcomes have been widely attributed to migrants' lack of work-ready English skills, serving to justify the inclusion of high English test scores in skilled migration selection processes. However, few studies have investigated actual communication practices in linguistically diverse workplaces, and these have focused mainly on interactions between multilingual workers (Piller & Lising, 2014). As a result, little is known about the actual communicative demands migrants face in workplaces where the majority of their co-workers speak English as their only language. Insights into these demands are urgently needed to evaluate the appropriateness of existing language testing practices for skilled migration purposes.

This paper reports on a project investigating both (i) migrant and local workers' perceptions of workplace communication challenges and (ii) the features of spoken interactions between skilled migrants with EAL and 'local' employees with English as their only language in two workplaces in regional Australia. Migrants and local workers were first interviewed at each workplace, and then four migrant participants, recorded their spoken interactions with local workers over several days across a period of four weeks, generating a total of 10 hours of audio recordings. These were transcribed according to conversation analysis conventions. Results showed that skilled migrant workers were able to engage a wider repertoire of interactional strategies and behaviours than local workers to negotiate shared understandings, resolve misunderstandings, and build rapport. Findings undermine a deficit view of migrant English, raise questions about the appropriateness of English requirements for skilled migration, and highlight a need to support the development of intercultural competencies among monolingual English speakers in Australian workplaces.

Language attainment for immigrant integration in the U.S.: Countering the official nationalist narrative

Antony John Kunnan (Carnegie Mellon University)

It has generally been asserted that immigrant integration (in the receiving country) is largely dependent on success in language learning. In the U.S., About 85% of immigrants learn English through formal instruction at school or informally at their place of work. Success depends on many variables coming together in an optimal manner and for a sustained period, especially for limited English proficiency immigrants. Immigrants who want to become U.S. citizens need to pass the English and History and Civics components.

This paper reports the official nationalist narrative on language attainment with data from the Department of Homeland Security regarding the lack of attainment of speaking English "at less than very well." According to Freeman and Tendler (2012, 2021), the lack of attainment of speaking English at "less than very well" by immigration and L1 status show that noncitizens (62.2%) report much higher rate of "less than very well" than foreign-born not naturalized (52.3%) and naturalized citizens (38.9%).

However, results from the storytelling project with immigrant narratives conducted in 2022 tells a different story – the subaltern story. A total of 52 immigrant-participants (not U.S. citizens) with low-level proficiency (A1 or A2 level) and in low-skilled workplace contexts were interviewed by 12 ESL teachers in a storytelling project in southern California. Questionnaire and storytelling data were collected on personal, educational, cognitive, linguistic and economic variables. The emerging story from the questionnaire survey and narratives is that these immigrants succeed in carrying out daily tasks with translanguaging or code mixing and code switching techniques (English-Spanish, English-Vietnamese, English-Mandarin) in four service contexts (post offices, grocery stores, restaurants, gas stations) and in four workplace contexts (gardening, construction, painting, nursing assistance). These findings counter the official data but also question the need to have the English component of the U.S. Naturalization test.

The role of language assessment for immigration purposes: The case of Chinese immigrants in Canada

Coral Yiwei Qin (University of Ottawa)

Language assessment has been widely used by governments as a selection tool for controlling migration (Frost & McNamara, 2018; Fulcher & Davidson, 2009; Van Avermaet, 2009), functioning not only as a measurement tool for language proficiency, but also as a mechanism to regulate the flow of desirable immigrants (Blackledge, 2009; Dickie, 2016; Frost, 2019; Shohamy, 2001; 2006; 2009). Despite numerous studies highlighting inconsistencies between test purpose and use in immigration policy, Canada, with the world's highest immigration rates per population (Statista Search Department, 2023), remains underexplored. Additionally, the substantial Chinese immigrant population in Canada, exceeding 1.7 million (Statistics Canada, 2023), adds a crucial perspective to the global discourse on immigration and language requirements. In prioritizing immigrant voices, this study, guided by policy archaeology (Scheurich, 1994) and model of investment (Darvin and Norton, 2015), aims to understand the social and political roles of English language assessments in Canadian immigration policy and their impact on Chinese immigrants.

Data collection involved 108 documents, 11 focus groups with 52 participants, and follow-up interviews with 10 participants. The innovative co-analysis protocol, where interview participants served as co-researchers, enabled drawing connections between policy statements and real-life experiences. Study results unveiled unexpected perspectives among Chinese immigrants. Despite challenges in reaching the required language proficiency level, participants generally supported using language assessment to select immigrants with higher proficiency. Moreover, despite acknowledging ideological differences, a significant percentage believed integration was their responsibility. The findings, interpreted through the model of investment, provided potential explanations rooted in ideology, identity, and capital.

The implications of the findings are valuable for language assessment and immigration researchers, offering insights into social cohesion and nation-building discussions in Canada and beyond. The inclusion of immigrants' perspectives contributes significantly to understanding the broader impact of language testing on individuals and society at large.

Crossing frontiers and fulfilling standards: The role of language proficiency in integrating internationally educated nurses in Canadian healthcare

Eunice Eunhee Jang (University of Toronto), Maryam Wagner (McGill University), Jeanne Sinclair (Memorial University of Newfoundland), Melissa Hunte (University of Toronto)

Healthcare, a globally mobile profession, draws professionals beyond their training countries. In Canada, Internationally Educated Nurses (IENs) form about 9% of the nursing workforce (Canadian Institute for Health Information, 2017, 2021), with their employment rate in nursing outpacing other occupations. Despite this, Canada faces a nursing shortage, with an estimated deficit of 60,000 nurses by 2022, excluding nurse outmigration intensified by the COVID-19 pandemic (Tomblin et al., 2012). Canadian immigration policy has responded by creating dedicated immigration streams for healthcare workers, particularly IENs. Yet, paradoxically, 58% of degree-qualified IENs are overqualified for their roles (Cornelissen, 2021). This paradox arises from two main challenges. First, IENs confront a complex, costly process to register as regulated nurses. Second, global variations in nursing training and experience mean some IENs are immediately competent, while others require further education. This situation is compounded by the competing priorities of Canadian nursing regulators. They must balance addressing the nursing shortage with ensuring public safety, necessitating IENs to prove language proficiency through standardized testing. This is critical as competent nursing in Canada demands advanced communication skills across diverse patient populations and settings. However, a universally accepted nursing communication skill set is yet to be clearly defined, making it difficult to maintain current and relevant language proficiency standards.

Our presentation will focus on immigrants from the highly-skilled profession of nursing and to explore how test-based language proficiency requirements for IENs are being challenged, negotiated, and adapted in migrant-rich countries like Canada. We aim to shed light on the complexities of integrating IENs into the Canadian healthcare system, balancing the urgent need for skilled nurses with the imperative of maintaining high standards of patient care and communication.

Symposium 2

Applying diagnostic assessment in AI-assisted language learning

Chair(s): Lianzhen He, Zhejiang University

Discussant(s): Xiaoming Xi, Hong Kong Examinations and Assessment Authority

Wednesday, July 3, 2024, 3:30 to 5:30 Location: HS1

As technology continues to revolutionize education, AI-driven tools have emerged as formidable companions in language assessment and learning. Despite a growing interest in AI-assisted language testing within the language assessment community (Aryadoust, 2023; Ockey, 2023) and AI-assisted language learning within the second language acquisition (SLA) community (Zhang et al., 2023), there is a lack of research demonstrating the effective integration of AI technology with assessments to facilitate language learning and teaching. Taking diagnostic assessment as an example, this symposium shows the potential of leveraging AI to empower diagnostic practices in a university and aligning diagnostic assessment with teaching and learning to enhance the quality of English as a Foreign Language (EFL) education at tertiary level.

Diagnostic assessment, characterized by its focus on identifying learners' strengths and weaknesses, serves as a cornerstone for tailoring instructional content to individual needs, thereby fostering a more targeted and effective learning experience. Through the in-depth analysis of learners' strengths and weaknesses, diagnostic assessments enable AI algorithms to personalize instructional content and adapt teaching strategies dynamically. This personalized approach transcends the one-size-fits-all model, fostering a more efficient and engaging learning environment.

In this symposium, we will report on the use of a standards-based intelligent English learning and teaching platform, aiming to provide customized language education solutions for all students at a university. The system starts with a standards-based diagnostic assessment, administered to incoming undergraduates, graduates, and Ph.D. students to this university. This assessment yields individualized feedback encompassing various aspects of language proficiency, including students' ability in four domains (i.e., listening, reading, writing, speaking) and corresponding levels, subskill mastery statuses, and suggestions for remedial actions. Following this, standards-based subskill-oriented online courses are offered to those who need to improve one or more of their language subskills, together with standards-graded materials and AI-facilitated spoken dialogue systems for daily practice. The learning outcomes are periodically assessed through diagnostic assessments administered at three-month intervals. Based on performance in these assessments, the assigned courses and learning resources for each student are automatically adjusted.

We assemble four groups of researchers from distinct disciplinary backgrounds to explore the advantages and obstacles associated with the application of diagnostic

assessment in AI-assisted language learning. We will first discuss two issues in the context of diagnostic listening and reading assessment: the advantage of diagnostic-by-design approach over retrofitting approach in providing accurate and reliable diagnostic information beyond a total score (Paper 1) and the implementation of an integrated approach to generate individualized qualitative feedback, finding a balance between specificity and generalizability for remedial teaching and learning (Paper 2). We will then turn to issues concerning writing and speaking assessment: how interventions based on the individualized test feedback effectively enhance learners' writing ability development (Paper 3) and the impact of the output of AI-assisted dialogue systems on L2 learners' oral test performance (Paper 4). Finally, a discussant will establish connections among the contributions and address the primary challenges associated with the utilization of AI in the digital era of language assessment and learning.

Presentations of the Symposium

Developing and validating an EFL diagnostic reading assessment: a case for the diagnostic-by-design approach

Shangchao Min (Zhejiang University), Hongwen Cai (Guangdong University of Foreign Studies)

Language tests that are designed and built for diagnostic purposes are believed to provide more diagnostic information than tests designed to measure general proficiency (Min & Cai, 2023). However, little cognitive diagnostic assessment (CDA) research has been conducted on purpose-built diagnostic language tests. This study is a CDA attempt to explore the diagnostic potential of an EFL reading test that is designed for diagnostic purposes, comprising distinct clusters of items to represent multiple dimensions. The data used in the study were 10,175 candidates' item-level response data of the reading subtest of a large-scale EFL diagnostic test for tertiary-level students. The test is built on a model of cognitive processing and limited to five cognitive operations that are targeted in the curriculum, namely, understanding words and expressions, understanding literal meaning, reconstructing meaning, understanding the organizational structure of the text, and summarizing paragraphs. CDA analyses were completed with the GDINA R package, version 2.7.8. The results showed that 1) the five attributes could be distinguished from each other, with attribute correlations between .422 and .693; 2) the mastery profiles were distributed in a balanced manner, with a flat profile rate of .558, lower than the values reported in most previous studies; 3) the classification reliability was generally satisfactory, with test-level accuracy of .763, and pattern-level accuracy over .700 for all patterns that agreed with intuition. The findings support the use of EFL reading tests built in conformity with the diagnostic-by-design approach to provide added information beyond the total score for remedial learning and instruction, particularly for test takers in intermediate statuses between all nonmastery and all mastery statuses.

Developing individualized feedback for an EFL diagnostic listening assessment: Combining standard setting and cognitive diagnostic assessment approaches

Lianzhen He (Zhejiang University)

Although combining standard setting and cognitive diagnostic assessment (CDA) approaches presents a plausible way to provide individualized descriptions about each test taker's ability to process certain levels of written and oral texts (Green, 2018; Powers et al., 2017) and their cognitive strengths and weaknesses while processing (Jang et al., 2015; Kim, 2015), to our knowledge, there has been little effort in combining the two approaches to provide individualized feedback for test takers in order to facilitate remedial learning and instruction in language assessment research and practice. This is probably owing to the fact that little is known about the relationship between performance-level classification based on standard setting and mastery/non-mastery classification of subskills based on CDA. In this study, we present the development of individualized feedback for a large-scale EFL diagnostic listening assessment by combining standard setting and CDA approaches. We used the performance data from 10,175 students' item-level responses to an EFL diagnostic listening test for CDA analyses. Fourteen experienced panelists served on the standard setting panel, who linked the diagnostic listening test to the national standards via the modified Angoff method. The results showed that proficiency classifications and subskill mastery classifications were generally of acceptable reliability, and the two kinds of classifications were in alignment with each other at individual and group levels. We then generated individualized qualitative feedback based on standard setting and CDA results, and conducted semi-structured interviews to 18 students to elicit their perceptions of the individualized diagnostic feedback. 12 out of them voiced great appreciation of the standards-based feedback in terms of the level designation and subskill mastery classification. The current study, by illustrating the feasibility of combining standard setting and CDA approaches to produce individualized feedback, contributes to the enhancement of score reporting for language assessment.

The contribution of individualized test feedback and intervention on EFL writing development: A five-month investigation

Xunyi Pan (Zhejiang University), Wenzhi Chen (Zhejiang University), Liqing Qiao (Zhejiang University)

In writing assessment, analytic profiles of learners' performance provided in diagnostic tests are considered to outperform holistic scales for its potential in capturing the multi-faceted nature of writing (Alderson et al., 2015; Knoch, 2011). However, most cognitive diagnostic assessment (CDA) research on writing focuses on the development or validation of rating scales for diagnostic tests (Knoch, 2011; Zhang, 2018). The process of integrating individualized descriptions on strengths and weaknesses identified in diagnostic writing tests in subsequent instructing and learning practices remain understudied (Kong & Pan, 2023). The present study attempts to explore the impact of providing diagnostic feedback coupled with follow-up targeted training on learners' EFL writing development. Participants took an in-house diagnostic test aligned with national standards and received analytic test feedback with standards-based descriptors on an AI-assisted English learning and teaching platform.

Teaching intervention was designed based on their subskill mastery statuses and administered over a five-month period on the platform. Online learners' reflective journals were completed on a regular basis and collected for analysis, including responses to a set of open-ended questions on their perceptions of difficulties, goals and strategies in writing. We analyzed their writing performance in terms of global, clausal and phrasal complexity using holistic and fine-grained indices. The results showed that diagnostic feedback generally matched learners' perceptions of their overall writing competence, and feedback in combination with targeted follow-up intervention facilitated learners' writing performance at multiple levels. The study indicates the need to provide refined feedback for better self-regulation in learning, and to understand the interface between feedback and treatment in the day-to-day practices of teaching and learning.

Exploring the impact of the output of AI-assisted dialogue systems on L2 learners' oral test performance: An interactive alignment perspective

Min Wang (Zhejiang University), Huiyang Shen (Shanxi Normal University), Zihui Zhang (Zhejiang University)

In the era of AI, the integration of AI-assisted dialogue systems has emerged as a novel approach in L2 speaking tests. Existing research primarily focuses on test-takers' language performance when interacting with these systems (Ockey & Neiriz, 2021). Little has been known about the psychological processes in AI assisted tests. Even less is known about how machine language output affects test-takers' socio-affective factors and task performance. These factors, however, critically contribute to the validity and reliability of AI-assisted tests.

Interactive alignment, the cognitive process wherein mental representations converge during conversation, has been proved to be a pivotal psychological mechanism underpinning human communication and has been extensively validated across various dialogue contexts (Pickering & Garrod, 2004, 2021). Hence, it serves as an apt paradigm to investigate the psychological processes in different interactional contexts. Research on human-machine dialogue reveals that individuals align with machine partners on multiple linguistic levels, thereby adapting their own language behavior (e.g., Branigan et al., 2011; Cowan et al., 2015; Linnemann & Jucks, 2016). Concurrently, the alignment between machines and humans shapes users' socio-cognitive and emotional experiences (Shen & Wang, 2021, 2023).

Building upon insights from this body of research, this presentation adopts interactive alignment as a conceptual framework to explore the potential impact of machine language output features (e.g., language complexity, voice anthropomorphism, alignment with humans) in dialogue system-based oral proficiency testing. The investigation encompasses both the impact on test-takers' socio-affective factors, such as perceptions of task difficulty and cognitive load, perceived authenticity, and naturalness of the interlocutor, and that on test takers' language performance. The presentation concludes by proposing specific research directions and design considerations for empirical studies in this domain.

Symposium 3

Open Science in Language Testing: Bridging Academic and Industry Perspectives

Chair: J. Dylan Burton, University of Illinois Urbana-Champaign

Discussant(s): J. Dylan Burton, University of Illinois Urbana-Champaign

Thursday, July 4, 2024, 8:30 to 10:30 Location: Aula

Open Science practices—including data/material/code sharing, preregistrations, and open access publishing—emphasize accessibility, transparency, reproducibility, and replicability, which are all critical for enhancing the credibility and robustness of scientific findings. In applied linguistics, scholarship in Open Science has gained substantial traction over the past two decades, with publications, databases, and conferences all supporting the initiative (Liu et al., 2023). The benefits of adopting Open Science practices are numerous, including a reduction in publication bias (Nosek et al., 2018) and a higher replicability of novel research findings (Protzko et al., 2023). For language testers, however, adopting Open Science presents a number of challenges due to its disciplinary duality in both academic and industry spaces. The application of Open Science may not always be feasible due to factors such as proprietary data, sensitive information, and intellectual property rights. Before the field adopts a "one size fits all" stance towards Open Science practices, it is necessary to highlight the different contextual needs from a variety of stakeholders in the field. This symposium aims to address this gap by inviting language testing researchers from a range of academic and industry backgrounds to discuss the specific challenges and opportunities of Open Science for language assessment.

The symposium will consist of a two-hour program, featuring six speakers, divided equally between academia and industry voices. Following a 10-minute introduction to the symposium, each speaker will have 15 minutes to present their perspectives. At the conclusion of the symposium, there will be a 20-minute slot to open the floor for discussion and interaction with the audience. The academic speakers will address the significance of Open Science for academic research. They will shed light on the ways in which accessibility, transparency, reproducibility, and replicability can enhance the quality and credibility of language testing research. These speakers will discuss practical issues surrounding adopting Open Science practices, sociopolitical considerations in East Asian contexts, and challenges and opportunities for graduate student training. The industry speakers will tackle the limitations and contextual constraints that industry may encounter when sharing resources and test data. They will discuss the intricacies of data privacy, proprietary concerns, and the protection of sensitive information that pose unique challenges to the language testing industry. The perspectives will highlight the needs of large scale, global proficiency testing agencies, insight from grant-funded work in sensitive geopolitical contexts, and views on Open Science in the world of big data.

The purpose of the symposium is to provide a broad understanding of the potential for Open Science to reform research practices in language testing. The symposium will provide a roadmap for aligning the principles of Open Science with the unique needs and constraints of testers in academic and industry settings. By fostering a dynamic dialogue between academia and industry, the symposium will contribute to a more open, inclusive, and robust future for language testing research, addressing the rapidly evolving landscape of Open Science and its application in language testing.

Presentations of the Symposium

Open Science: A step-by-step guide for language testers

Paula Winke (Michigan State University)

I review current Open Science practices in the field of language testing, and explain why Open Science practices are needed. According to the United Nations Educational, Scientific, and Cultural Organization (UNESCO, 2023), Open Science "combines various movements, practices and actions that aim to make all fields of scientific research accessible to everyone for the benefit not only of scientists but also society as a whole." In sum, Open Science can foster deliberate methodologies, resulting in higher caliber research. I discuss the practice of Open Science for individual language testers as a series of four steps. Step 1: Use persistent identifiers, such as an ORCID. Step 2: Earn the four Open Science Badges when you do research (Preregistration, Open Materials, Open Data, or Open Code). Step 3: Write transparently. Step 4: Publish openly. I argue that Open Science will put language testing's focus on quality, not quantity, which, as others have stressed, is a worthy consequence of Open Science (Marsden & Morgan-Short, 2023). I explain the field of language testing has a role to play in promoting Open Science, such as working toward accessible corpuses of testing data, improving peer review through more transparent processes, and moving hidden labor on testing research out from behind the scenes. As language testers, our goal is to scientifically and socially guide the design of language assessments and the uses of scores from them. Open science will help us do this as part of a humanistic drive to make language education and assessment more accessible, comprehensible, and fair.

Open science in language assessment in China: Prospects and problems

Jason Fan (University of Melbourne), Jin Yan (Shanghai Jiaotong University)

Open Science (OS) practices are widely acknowledged for their numerous benefits for science and its stakeholders (Banks, et al., 2019). However, their adoption has encountered numerous challenges and constraints, particularly in developing countries (e.g., Jin et al., 2023; Zhang et al., 2023). Despite the recent upsurge of interest in OS in the field of language assessment (LA) (Winke & Burton, 2023), empirical research on this topic is slow to catch up. Situated in the context of LA research in China, this study explored the current status of OS in LA research in China, as well as the challenges and constraints of engaging with OS movements and practices, from the perspective of LA researchers.

This study consists of two stages. During the first stage, we analysed the information on the websites of the leading LA journals in China, aiming to explore current OS policies

and practices. During the second stage, we conducted focus groups (n = 8) and one-on-one interviews (n = 12) with LA researchers, to explore their understanding of OS, their values on and perceptions of the current OS movements in LA, as well as the challenges and constraints they face in engaging with OS practices.

Our findings reveal that: (a) there was little evidence of OS practices with academic journals on LA in China; (b) LA researchers had limited OS literacy but expressed willingness and interest to engage with OS practices; and (c) major challenges identified included lack of awareness and motivation, insufficient funding and/or institutional support, and socio-cultural factors. Our findings underscore OS as a multifaceted, value-laden concept, and its practices are embedded in and interact in complex ways with personal, social, political and cultural contexts. We conclude by articulating the significance of fostering an OS culture, strengthening the collaboration among stakeholders, and enhancing OS literacy.

Graduate Student Insights: Exploring Open Science Challenges and Opportunities

Jieun Kim (University of Hawai'I at Manoa), Daniel R. Isbell (University of Hawai'i at Manoa)

Open Science (OS) has emerged as a transformative movement in research, promising enhanced transparency, accessibility, and reproducibility. However, this movement is still in its early stage in language testing, leaving graduate students to face unique challenges in implementing OS practices. During this talk, we will discuss challenges faced by graduate students and potential learning opportunities related to OS.

Graduate students in language testing face several unique challenges, particularly limited access to proprietary materials and data. Unlike other areas of applied linguistics that actively use shared repositories, such practices are rare in language testing, with few studies based on open materials/data and virtually no preregistration and replication studies. Particularly, students in language testing, lacking industry and academic connections, must often rely on publicly available resources or grapple with restricted access to essential materials/data.

Financial constraints present another obstacle specific to language testing research. To use official test materials and score data, more senior researchers often obtain large grants from test developers or government sources. Graduate students are ineligible for those grants, and research involving language exams, including practice versions, can be costly in terms of participant compensation. Additionally, students may find open access publishing costs prohibitive, given their limited institutional support and personal finances compared to post-PhD researchers. Consequently, they face limited options for publication and reproducing OS practices.

Despite these challenges, OS practices offer hope for addressing some of the difficulties faced by graduate students. For instance, preprints and postprints provide affordable options for open access publishing. Journals can also play an important role by increasing general expectations for transparency. Moreover, we believe that more senior scholars can promote OS through their own work and the courses they teach. We will invite attendees to envision how OS could be integrated into graduate training for the next generation of language testers.

Challenges for the language assessment industry in the context of Open Science: How can they be addressed?

Spiros Papageorgiou (Educational Testing Service)

Providers of language assessments support Open Science in many ways, some of which are the publication of freely accessible technical reports and the provision for public use data sets. However, support for Open Science also comes with challenges and limitations for the language assessment industry, which relate to protecting the privacy of individuals and institutions, as well as proprietary information and capabilities. I will first present examples of the adoption of Open Science by the language assessment industry. I will then discuss some of the inevitable challenges language assessment professionals face as they continue to adopt Open Science, specifically when it comes to preregistering research protocols and grants proposals, publishing anonymized test data and responses to speaking tasks, and publishing openly. I will conclude by arguing that the Open Science movement can truly advance if the perspective of language assessment providers is fully understood by the academia.

Getting the balance right: open science in diverse cultural and political assessment contexts

Karen Dunn (British Council)

The British Council provides an interesting case study on the potential benefits, both to test developers themselves and to the field, of open science while highlighting some of the challenges. The Council's work in assessment spans a diverse range of contexts across grant-funded education reform projects to income-generation activity through the development and delivery of assessment "products". Work in all areas is carried out under our charitable mission, with the aim of contributing to a "greater good." But these areas have different demands, with income-generation areas overlapping with the "industry" referred to in the symposium proposal. Grant-funded areas have special requirements, particularly when operating in collaboration with international aid and development agencies. Data protection and sharing requirements will vary from country to country, meaning a one-size-fits-all approach won't work, and needs to be negotiated on a case-by-case basis, taking into consideration both cultural and local legal considerations. Projects closely connected to government education reform will require the timing and process of making information, research results, and data public to be negotiated carefully with all stakeholders. While these situations are by no means unique to the Council, the presentation will highlight some of the rich (and thus challenging) contexts in which language testing takes place. The talk will also argue that we should not view "academia" and "industry" as a dichotomy. We would be better thinking about the challenges and opportunities of open science for those working in language assessment as being more like a Ven diagram. The range of "actors" is diverse, with organizations, for-profit and charitable, universities, and national testing agencies being just some. These groups will have some unique considerations, but often overlap in relation the potential for sharing research and data with the field, particularly as they often work in collaboration on complex assessment projects.

Big Data and Open Science: Balancing Collaboration, Privacy, and Protection

Geoffrey T. LaFlair (Duolingo)

The recent shift to digital-first assessments presents a transformative opportunity for assessment research, offering unprecedented access to large datasets of a diverse nature, which can drive innovation in language assessment. Similarly, recent shifts in applied linguistics research towards practices that align with open science principles offer access to resources that can help move the field forward at a faster pace and facilitate collaboration, learning, and understanding of language assessment research. These two movements have potentially large positive implications for language test development and research, and it might seem like they are a natural fit for each other. However, from the perspective of an industry whose primary customers are test takers, sharing research artifacts is a cautious activity that must prioritize the protection of test takers as well as intellectual property and proprietary information.

Protecting test takers' personal information is paramount in this new landscape. As bigger and more diverse types of data are collected and shared, the risk of re-identification of test takers through the use of other large publicly available datasets increases. However, methods of anonymization that decrease the likelihood of re-identification (e.g., generalization, data swapping, or data perturbation) might run contrary to principles of the open science movement. Intellectual property, including proprietary information about test items, scoring processes, and prompt engineering, also requires thorough protection. The open science movement, while promoting transparency and collaboration, could inadvertently place proprietary information at risk. Therefore, it is crucial to establish clear guidelines for data sharing that respect and protect intellectual property rights. Doing so ensures that organizations feel more protected and open to supporting open science practices. While big data and open science hold great promise for advancing language assessment, careful attention must be paid to protecting all stakeholders and maintaining the integrity of the assessment process.

Symposium 4

Reforming the Diagnosis of L2 Abilities: The Complementary Contributions of Dynamic and Diagnostic Language Assessment Frameworks

Chair(s): Dmitri Leontjev, University of Jyväskylä; Matthew Edward Poehner, The Pennsylvania State University

Discussant(s): Claudia Harsch, Universität Bremen

Thursday, July 4, 2024, 4:00 to 6:00 Location: Aula

This symposium serves as a forum for a discussion focusing on integrating dynamic and diagnostic language assessment. Both dynamic (henceforth DA) and diagnostic (DiagA) assessment frameworks started to gain the attention of second/foreign language (L2) researchers in the early 2000s, not least because of the growing concerns over perceived negative consequences of testing on teaching and learning (Shohamy, 2000). To date, DA and DiagA have both generated considerable bodies of research (for overviews, see, respectively, Author et al., 2021 and Authors, 2023). While the two approaches emerge from different theoretical traditions and differ methodologically, they share a goal of diagnosing learner abilities and using the information obtained to support and guide subsequent teaching and learning. Briefly, DA is informed by Vygotskian Sociocultural Theory and regards learner independent performance of assessment tasks as indicative of abilities that have already fully developed; understanding abilities that have begun to develop but have not yet fully formed requires the provision of mediation (e.g., leading questions, prompts, models, feedback) during the assessment. Mediation thus represents a form of teaching that is central to a diagnosis of development, and the outcome of the procedure points to those abilities that can best be targeted by future teaching. DiagA, in turn, focuses on identifying learner strengths and particularly weaknesses in learner L2 knowledge as they currently appear, emphasising careful modelling of constructs informed by SLA research. DiagA also underscores the importance of actionable feedback to teachers and learners. As we argued elsewhere (Authors, 2023), both approaches might be enriched through an in-depth investigation of their commensurabilities and intersections, a topic L2 researchers have yet to fully explore. In our view, doing so may also reform language assessment research and practice, perhaps yielding a new framework that builds on the strengths of both approaches.

Following a ten-minute conceptual introduction by the organisers, the symposium includes four presentations (fifteen minutes with five minutes for questions each) reporting original studies by researchers working in different assessment contexts in Australia, Finland, the U.K., and the U.S. Two of the papers concern DA, one involving oral proficiency testing of L2 Chinese and the other the assessment of English academic writing. The other two papers focus on DiagA, one investigating DiagA via self-assessment in two different educational contexts and the other reconceptualising an established online diagnostic assessment system in multiple languages. Each paper

offers reflections on how the selected assessment approach (DA or DiagA) contributes to the diagnosis of learner L2 development as well as how the research might engage with principles or methods from the other assessment approach. Finally, the discussant will have fifteen minutes for observations and remarks preparatory to inviting the audience to think together during the remaining fifteen minutes about the continued development of both approaches, focusing on such topics as how constructs might be operationalised in a dynamic-diagnostic assessment framework and possible designs of dynamic-diagnostic procedures.

Presentations of the Symposium

OPI-DA: A dynamic approach to diagnosing learner L2 oral proficiency

Matthew E. Poehner (The Pennsylvania State University), Jie Zhang (University of Oklahoma), Tianyu Qin (University of North Georgia)

We report initial data from a larger project that explores a new diagnostic language procedure that integrates Dynamic Assessment (DA) principles (e.g., Poehner & Wang, 2020) into the American Council on the Teaching of Foreign Languages (ACTFL) Oral Proficiency Interview (OPI). Widely used for purposes such as meeting language program requirements and credentialing language teachers, the OPI yields a sample of learner language production that is ranked according to ACTFL's eleven-level proficiency scale that captures increasingly complex communicative functions. Our project is motivated by Vygotsky's (1998) argument that a diagnosis of development must include insight into underlying difficulties as well as practical recommendations for improvement. The primary research question guiding the study is, how might profiles of learner L2 oral proficiency be elaborated through the use of mediation, including prompts, leading questions, models, and feedback, targeting areas of difficulty that emerge during the OPI? The project thus follows the principles of diagnostic language assessment in organizing assessment procedures according to a carefully defined construct, in this case operationalized according to the ACTFL proficiency scale. At the same time, the scale serves as a basis for orienting mediation during a DA administration of the OPI for the purpose of (a) determining how far learners might 'stretch' their language proficiency during cooperative interaction and (b) identifying the amount or quality of support they require to do so. Following an introduction of the OPI-DA procedure, we provide an overview of the project, currently underway at a large U. S. university with students of L2 Chinese in a Chinese Language Flagship Program. We illustrate the approach with examples of transcribed interactions from the interviews and the diagnostic insights they offer into learner language proficiency. Implications of assessment outcomes for learner ongoing language study in the program are discussed.

Learning to write in the ZPD: Dynamic Assessment of L2 English Learners' Academic Writing Development

Lu Yu (University of Melbourne), Dmitri Leontjev (University of Jyväskylä)

This presentation reports data drawn from a research project that investigates the effectiveness of dynamic assessment (DA) and a DA-informed enrichment program in promoting L2 English learners' academic writing development. Grounded in Vygotskian

Sociocultural Theory (SCT) and its concept of the Zone of Proximal Development (ZPD) (Vygotsky, 1978), DA introduces mediation (e.g., prompts, hints, suggestions, modeling, etc.) into the assessment procedures. By observing learner responsiveness to mediation, DA not only diagnoses an individual's emerging abilities but also triggers the development of the abilities being assessed (Poehner, 2008, 2018). Thirteen ESL students were recruited from a public U.S. university. The study included three stages: Time 1 assessments, a five-week, one-on-one writing instructional intervention program, and Time 2 assessments. The same Sandwich format of DA (Sternberg & Grigorenko, 2002) was implemented at Time 1 and Time 2 assessments, where the learners composed an essay independently in response to a reading-writing integrated task, then engaged in an interactionist DA session with a mediator while jointly reviewing the essay with reference to an analytic rubric, and finally revised the essay independently. During the writing instructional intervention, the learners were randomly assigned into an enrichment program group (n = 6), where the instruction was ZPD-attuned and informed by DA diagnoses (Feuerstein et al., 2010), and a non-enrichment program group (n = 7) that received standard, generic writing instruction. The enrichment group evidenced greater gains over time than the standard group in terms of rubric ratings. This presentation focuses on analyzing the interactions that occurred during the two DA sessions with one focal enrichment participant and his developmental trajectory. Discussion addresses (1) additional diagnostic inferences obtained through DA procedures, which were not disclosed in learner independent writing performance, and (2) the potential of integrating dynamic and diagnostic assessment frameworks in promoting learner writing development.

Paving the way for dynamic assessment through diagnostic tests and self-assessment lists: a comparison between Finnish and Chinese university students

Magdalini Liontou (University of Jyväskylä), Ari Huhta (University of Jyväskylä)

Typically, diagnostic assessment (DiagA) targets specific skills to produce a profile of strengths and weaknesses for learners to guide further teaching and learning. However, DiagA can also focus on broader levels (e.g., speaking, reading), often as the first step. In this study, a general language test (Oxford Placement Test, OPT) and an extensive 174-statement self-assessment battery covering reading, writing, listening, spoken production and spoken interaction were administered to 60 Finnish and 80 Chinese first-year university students as part of a compulsory English language course in Finland and China. The statements came from the CEFR and the European Language Portfolio and covered CEFR levels A2–C1.

Self-assessment data were analysed with the Rasch software Winsteps to calibrate the statements and to identify any misfitting statements, as misfit may indicate difficulties in interpreting particular statements. Comparing the calibrated positions (rankings) of the statements with their original CEFR levels allowed us to identify unexpectedly difficult (indicating weaknesses) or easy (indicating strengths) statements for one or both student groups. The difficulty of most statements in our study corresponded to their original difficulty, for example, the original A2 statements were mostly calibrated as the easiest ones. However, some unexpected calibrations, typically only for one group, suggest context-specific strengths and weaknesses in students' proficiency.

Self-assessments, thus, can provide diagnostic information about areas of weakness that can be addressed using dynamic assessment in the classroom. We will discuss how such diagnostic information received in learner self-assessment can serve as a basis for dynamic assessment interactions with learners. We will also propose how mediating learners' self-assessment processes can enhance diagnostic information about learners' strengths and weaknesses (see Poehner, 2012).

Is there a role for mediation in a theory-driven approach to diagnostic language assessment? Exploring the feasibility of a dynamic approach in Dialang 2.0

Luke Harding (Lancaster University), Tineke Brunfaut (Lancaster University), Benjamin Kremmel (University of Innsbruck), Ari Huhta (University of Jyväskylä)

Developed in the 1990s, DIALANG remains a unique example of a computer-based diagnostic language assessment system (Alderson, 2005). DIALANG is still hosted on an open-access platform (https://dialangweb.lancaster.ac.uk), providing purpose-built diagnostic tests in 14 languages (as well as 18 instructional and feedback languages) targeting reading, listening, writing, vocabulary, and grammar. DIALANG continues to be a popular tool for learners and teachers; however it faces two threats. First, DIALANG is no longer funded and therefore has no sustained technical support. Second, the field of diagnostic assessment has made numerous advances in the past two decades, and DIALANG no longer represents current thinking in theorising diagnostic language assessment (Alderson et al., 2015; Huhta et al., 2024). For that reason, prior to the pandemic, a team of researchers from several universities around the world started work on conceptualising DIALANG 2.0: an online system designed to represent a radically different approach to diagnostic language assessment from the original format.

A key feature of DIALANG 2.0 is that it aims to model an ideal diagnostic procedure including four stages: (1) observation, (2) initial assessment, (3) hypothesis checking, and (4) decision making. As the operational blueprint for DIALANG 2.0 has continued to take shape, collaborations with scholars from the dynamic assessment tradition have prompted a key question: can (and should) elements of dynamic assessment be integrated within DIALANG 2.0?

In this presentation, we will first outline the most recent conceptualisation of DIALANG 2.0. Second, we will focus specifically on the potential to embed mediation in the diagnostic process – a promising current direction – with an example provided for a hypothetical test of pragmatic understanding. Finally, we will address the feasibility of embedding mediation given the considerable practical challenges involved in developing DIALANG 2.0.

Symposium 5

Locating competence, exploring constructs: Taking forward Tim McNamara's work in performance assessment

Chair(s): John Pill, Lancaster University; Lynda Taylor, University of Bedfordshire

Discussant(s): Lynda Taylor, University of Bedfordshire

Friday, July 5, 2024, 8:30 to 10:30 Location: Aula

A key element of Professor Tim McNamara's legacy in our field is his pioneering work in second language performance testing and specific-purpose language assessment. His seminal monograph Measuring Second Language Performance was published in 1996. It provided a critical examination of the practice and theory behind performance-based assessment in the context of second language learning, together with a comprehensive introduction to Rasch Measurement, which enables investigation of the many facets involved in performance tests (e.g., tasks, criteria, raters). In 1999, Sally Jacoby and Tim McNamara co-published a paper in the journal English for Specific Purposes exploring the complex inter-relationship of linguistic and professional workplace skills, and the way in which understanding of these can inform the assessment enterprise, particularly the design of test tasks and assessment criteria. 2024 marks the 25th anniversary of the publication of that paper – entitled 'Locating Competence'. As LTRC 2024 celebrates the enormous contribution Tim made through his life, both personally and professionally, this symposium is intended as a tribute to his legacy in one of many areas. It will enable us to reflect on progress made over the past two decades responding to his challenge to investigate and understand the complex nature of 'competence', and to consider again how this work can inform the activities of test design, development and validation in language assessment.

The symposium comprises a 10-minute introduction followed by four presentations (20 minutes each, including Q&A) focusing on research in different contexts in which considerations of language ability, workplace/situation-related knowledge and skills, and interactional and intercultural competence all come into play with potential implications for assessment policy and practice. The contexts under scrutiny include the aviation industry, patients' literacy in healthcare settings, communicative language tests seen through an English as a Lingua Franca (ELF) lens, and the perspectives of linguistic laypersons. The presentations therefore complement one another and allow consideration of a range of issues:

- what competence means in different contexts
- what matters and what is only marginally relevant
- where boundaries of 'language' are located
- how (test) performance provides evidence of competence and how we can measure it
- how we foreground what matters in an assessment to reflect the real world more faithfully
- how changing a construct might refocus or reweight assessment criteria to accommodate the required competence

The research presented reflects a variety of methodologies, including assessor interviews, model/framework creation, focus groups with domain insiders, conversation analysis, membership categorization analysis and corpus-assisted analysis.

The closing discussion (10-15 minutes) will draw together the main threads emerging from the four presentations and reflect on them in light of the insights into performance assessment shared in Jacoby and McNamara's 1999 paper. Consideration will also be given to how research in this area might be taken forward. The final 10-15 minutes are for audience questions.

The symposium input and discussion address the overarching conference theme of 'reform', touching upon how language assessment policy and systems need to be reformed as assessment constructs evolve in response to ongoing research into specific contexts of language use.

Presentations of the Symposium

Locating competence in the air traffic controller profession

William Agius (Zurich University of Applied Sciences & Lancaster University)

The linguistically informed rating scale currently mandated for use in the assessment of air traffic controllers' English language proficiency focusses on six areas of language use: pronunciation, vocabulary, structure, fluency, comprehension, and interaction. However, several authors (Kim & Elder, 2015; Knoch, 2014; Prinzo & Thompson, 2009) have questioned the validity of the scale. This paper reports on a study to identify indigenous features of language performance (Jacoby & McNamara, 1999) that air traffic controllers and pilots value in their peers, and that they consider pertinent as measures of the extent of an air traffic controller's workplace language socialisation.

This research combines qualitative and quantitative methods to achieve a broad perspective of the research context, verifying data through triangulation. In the first phase of the qualitative part of the research, air traffic controllers and pilots described the air traffic controller profession and the role of communication in its execution. In the second phase, new participants listened to recordings of test takers in a commercial test of English for air traffic controllers and subsequently commented on their perception of each test taker's operational readiness. Participants in these focus groups were domain insiders and linguistic laypersons. Their contributions informed an indigenous perspective of the air traffic controller profession. In the quantitative part of the research, the features of language performance derived from analysis of the focus group data were evaluated by assessors for a test of English for air traffic controllers through a questionnaire.

The results showed domain insiders' primary orientation to interactional and communicative competence. They paid close attention to test takers' operational and procedural knowledge and, for the most part, ignored linguistic correctness. Based on these results, a preliminary set of five new assessment criteria is proposed: handling of interaction, achievement of objectives, repair management, vocabulary, and pronunciation.

Locating patients' competence: The dynamic assessment of health literacy

Susy Macqueen (Australian National University)

Newly diagnosed patients enter a professional world in which a medical diagnosis initiates learning about their medical condition, their own bodies and identities, and institutional structures and processes. This learning occurs under heightened affective circumstances which can make it difficult to take on information and make decisions. In health research, health literacy tests aim to assess the extent to which patients (and care-givers) understand health information. Recent definitions have expanded the health literacy construct to include the ability to seek out, interpret, negotiate and act on health information (Berkman et al., 2010).

Unlike most Language for Specific Purpose (LSP) assessments, which focus on the language trajectories of subject matter experts, health literacy assessments are layperson assessments, that is, assessments used when it becomes necessary to formally gauge the understanding of non-specialists who have cause to engage with specialist content (Knoch & Macqueen, 2020). While there are several widely used health literacy assessments in English, these focus primarily on the comprehension of written health information. Yet most discussions of health literacy indicate a more expansive construct than those operationalised in common tests, for example, incorporating contextual considerations and the confidence to act on health knowledge (Nutbeam, 2008). Furthermore, patients' experiences of clinical contexts are multimodal, with much information about their conditions arising in spoken interactions.

This paper offers a reconceptualization of health literacy assessment by examining commonly used health literacy assessments and comparing these with the spoken interactions of patients undergoing treatment for heart failure. It sets out principles for on-the-ground dynamic assessment of health literacy in clinician-patient interactions, as patients develop understandings of their conditions and as clinicians work towards empowering patients to self-manage.

Locating features of English as a Lingua Franca (ELF) communication in the conversation task of the Graded Examinations in Spoken English (GESE)

Geisa Dávila Pérez (Lancaster University)

Accommodation strategies (i.e., translanguaging, paraphrase, repetition) contribute to effective communication in ELF settings (Cogo, 2020). However, while accommodation is central to conceptualisations of 'Lingua Franca competence' (Canagarajah, 2007; Harding & McNamara, 2018), accommodation strategies and related phenomena (e.g., repair) remain negatively viewed, as evidence of candidates' low proficiency and/or a source of interlocutor variation in the Language Proficiency Interview (LPI; Lam et al., 2023; Ross, 2017). Since unscripted LPIs may be seen as ELF encounters, research is needed on how accommodation strategies manifest in such exams and on examiners' perceptions of these features.

I address these issues by investigating examiner and candidate use of accommodation strategies in the conversation task in the GESE, a communication-oriented speaking test administered by Trinity College London. The study followed an exploratory sequential

design. In Phase 1, I conducted a corpus pragmatics study to explore the frequency and dispersion patterns of accommodation strategies, and their apparent functions across proficiency levels and speaker roles. Data were selected from 2,053 conversation task performances in the Trinity Lancaster Corpus, comprising a sub-corpus of 1,674,680 tokens. In Phase 2, 12 GESE examiners were interviewed about accommodation strategies, and the data were analysed thematically.

Phase 1 findings show that, while there is minor variation across proficiency levels, patterns of accommodation strategies used by examiners and candidates are both similar and different. This is supported by Phase 2 results. For instance, while examiner and candidate both genuinely clarify meaning, the examiner also 'stages' clarification as a way to elicit assessable talk. Based on these findings, I contend that (i) the use of accommodation strategies in LPIs – an assessment context – may be similar to but not necessarily the same as in 'real-world' ELF communication; and (ii) candidates and examiners regulate their strategy use mindful of the task construct and purpose.

Are non-language specialists' constructs usable for assessment? On the complementary use of sequential-categorial analysis

David Wei Dai (UCL Institute of Education, University College London)

Since Hymes's seminal work on communicative competence (CoCo), there have been a few iterations of CoCo as a construct in language assessment. To date, few constructs have captured the volitive and affective dimensions of Hymes's CoCo, largely due to the complexity of assessing such aspects of interpersonal interaction. In order to ascertain what a fuller CoCo construct entails, this methodology paper combines the criteria from non-language specialists (NLSs) and sequential-categorial analysis.

To develop a CoCo construct bottom-up, I first collected 22 test-takers' sample performances on a nine-item roleplay speaking test. I then asked 22 NLSs to listen to these test-taker performances and elicited their perspectives on what they felt constituted successful communication. After analysing their interview transcripts and written comments in NVivo, I noted that NLSs, who had no formal linguistic training, oriented more to the emotional, logical, moral and categorial (social role) aspects of communication, which concurred with the CoCo construct as envisioned by Hymes. The accounts of these components of CoCo were, however, quite vague and impressionistic, as the NLSs did not possess the metalinguistic repertoire to describe what they had perceived.

To further specify these elusive components of CoCo, I went back to test-taker discourse and conducted sequential-categorial analysis, which combines Conversation Analysis and Membership Categorization Analysis, to arrive at a finer understanding of how test-takers themselves used a range of semiotic resources to attend to these aspects of CoCo. This methodology paper therefore showcases that although NLSs can generate invaluable insight into a fuller account of CoCo, their criteria are oftentimes not readily usable for assessment, which requires precise description of ability indicators. However, complementing NLSs' criteria with discourse analysis that explicates speakers' linguistic conduct can address the under-specificity of NLSs' criteria and assist assessment developers to finesse assessment constructs and rubrics.

Paper and Demo Summaries – Wednesday, July 3

Multimodal EAP assessment reconceptualised

Sathena Chan (University of Bedfordshire), Nahal Khabbazbashi (University of Bedfordshire), Tony Clark (Cambridge University Press & Assessment)

Wednesday, July 3, 09:00 to 09:30 am Location: Aula

The widespread use of digital technologies in higher education (HE) means that the nature of communication may be shifting. There is a pressing need to better understand the novel ways in which language is used in HE contexts and explore the ways in which assessment tasks can reflect these new constructs. To this end, we present the results of our three-stage research project: (1) a scoping review to identify major trends of language use in HE, (2) design and implementation of new digital multimodal task formats which include a combination of linguistic, visual and aural modes and (3) evaluation and validation of the new task(s). Based upon evidence from the scoping review, a series of prototype digital multimodal tasks, which required test takers to navigate multimodal resources on digital platforms and to draw on them to engage in multimodal composing were designed and trialled. Eleven graduate students from two proficiency groups completed the multimodal test online. The presentation will report on the language functions – communicative purposes achieved through language use – elicited by the two groups of participants and feedback from test takers and examiners. The findings show promise for the future of assessments, as the new task types were successful in expanding EAP construct(s) beyond what is typically operationalised by independent skill-based tasks or single integrated tasks. The presentation will conclude with the implications of the findings for the operationalisation of multimodal solutions that incorporate technology to expand constructs of language proficiency and instigate assessment reforms in higher education.

The sound of one hand clapping: what monologues can tell us about interactional competence

Carsten Roever (University of Melbourne), Naoki Ikeda (University of Melbourne)

Wednesday, July 3, 09:00 to 09:30 am Location: HS1

Assessment of second language (L2) interactional competence (IC) has seen little implementation of IC assessment in operational tests mostly due to low practicality of dialogic IC tests. Practicality would be markedly increased if monologues could be used for IC assessment, and several existing IC assessments contain monologic tasks. The high correlations between scores on such monologue tasks and dialogue tasks suggest an appreciable degree of overlap. However, no previous study has described which interactional features relevant to the assessment of IC in dialogic interactions might also be present in monologues.

In this study, we collected data from 150 university-level ESL learners on an IC test containing three dialogues and three monologues. Their productions were scored using the criteria developed by Ikeda (2017), and we identified features in the monologic data which indicated orientation to an interlocutor, and compared how participants at different IC levels differed in these features.

Pearson correlation of dialogue and monologue scores showed a high correlation ($r = .829$, $p < .001$). Qualitatively, we found orientation to epistemics and intersubjectivity in monologues. Participants managed instances of potential conflict with the interlocutor by showing dispreference organization and showed affiliation by demonstrating empathy with the interlocutor. Test takers at different IC levels were differentially competent in the features identified.

Our findings suggest that monologues contain measurable features indicative of a speaker's IC. While this is promising for IC assessment, it is important to note that monologues do not provide sufficient information about turn taking and engagement with the interlocutor.

Reducing language barriers and improving diversity and inclusion in participant recruitment to randomized trials: A role for language assessment

Talia Isaacs (University College London), Andrea Vaughan (University College London), Eva Burnett, Zsofia Demjen (University College London), Marie-Anne Durand (Dartmouth College & Hanover NH; L'université Toulouse-III-Paul-Sabatier, Toulouse, France), Kate Gillies (University of Aberdeen), Kamlesh Khunti (University of Leicester), Jamie Murdoch (Kings College London), Nuru Noor (University of Cambridge & University College London), Leila Rooshenas (University of Bristol), Frances Shiely (University College Cork), Harpreet Sood (University College London Hospitals NHS), Fiona Stevenson (University College London), Matt Sydes (University College London), Shaun Treweek (University of Aberdeen), Katie Biggs (University of Sheffield)

Wednesday, July 3, 09:00 to 09:30 am Location: HS2

Randomized controlled trials (RCTs) are widely considered the gold standard of health intervention research, testing the safety/efficacy/efficiency of new medical treatments (Hedlin et al., 2021). Ethnic (including linguistic) minorities have been chronically underrepresented in trials, in part due to language barriers. Trial recruiters encounter competing tensions when making participant gatekeeping decisions during recruitment. On one hand, they need to ensure that prospective participants have the requisite language ability to understand the conditions and implications of trial participation—an ethical imperative (Davies et al., 2015). On the other, they need to adhere to language-related inclusion/exclusion criterion that is either explicitly stated (e.g., "must speak English"), implied, or assumed. This opens the door to potentially biased assessments that may also be detached from the trial's linguistic demands.

This presentation shares insights from two UK-based studies. The first uses computational tools and discourse analysis to analyse a corpus of 27 information sheets and 23 consent forms for cancer RCTs, with linguistic quality measures benchmarked against graded readers. Textual complexity metrics far exceeded medical associations' recommended reading levels for patient-facing materials—a likely deterrent for linguistic minorities' participation. The second study is a systematic review of 32 extended reports of depression and type 2 diabetes RCTs. We assessed the communication demands of the treatments (e.g., talking therapies) and primary outcome measures (e.g., questionnaires) and found haphazard language screening practices across studies and a dearth of language-related accommodations. A purpose-built tool that offsets recruiters' reliance on gut feeling could potentially lead to fairer, more consistent assessments.

An eye-tracking study of response processes on C-test items in the Duolingo English Test

Ruslan Suvorov (University of Western Ontario)

Wednesday, July 3, 09:00 to 09:30 am Location: HS3

The importance of validity evidence based on response processes is recognised among researchers and documented in the Standards for Educational and Psychological Testing. Process-based evidence is essential for cognitive validity, an indispensable component of Weir's (2005) socio-cognitive framework. Gathering cognitive validity evidence based on response processes is particularly important for newer language proficiency tests such as the Duolingo English Test (DET). Given the DET's short length and adaptive nature, it is paramount to understand how test takers interact with the DET items and how their response processes are related to scores and item difficulty levels. To address this need, the present study leveraged eye tracking to examine the relationship between 40 L2 learners' response processes on the DET's C-test items and item difficulty level and scores for these items. Different types of regression models were built to estimate the relationship between six eye-tracking measures and difficulty estimates (predictor variables) and scores (response variable) for C-test items both at the item level and the level of individual blanks. A larger number of eye fixations on C-test items was positively correlated with the possibility of answering the item correctly, whereas spending more time on C-test items led to lower scores on those items. The findings revealed individual differences among L2 test takers' response processes and highlighted the importance of complementing outcome-based evidence of language proficiency with process-based evidence. The presentation will conclude with the discussion of study limitations, implications for validation research, and future directions.

Clarifying Links Between Actionable Feedforward and Remediation in Diagnostic Language Assessment: Insights from Medical and Dynamic Assessment

Yong-Won Lee (Seoul National University)

Wednesday, July 3, 09:00 to 09:30 am Location: UR3

Learning-oriented assessment approaches, including diagnostic language assessment (DLA), have drawn much attention from language testers in recent decades. DLA is designed to identify learners' strengths and weaknesses in a targeted domain of language proficiency (Alderson, 2014, Lee, 2015) and provide diagnostic feedback (or feedforward) to guide subsequent remediation. In other words, DLA has explict goals of positively impacting future language learning by assessing the difficulty of learners and provide appropriate remediation to facilitate their further growth. While the procedures for validating test score interpretation and use have been relatively well-documented in the fields of assessment (Bachman & Palmer, 2010; Chapelle et al., 2010; Kane, 2006), however, those for evaluating the quality of diagnostic feedforward and remedial learning have not been sufficiently developed. With this as a backdrop, the major goals of the current study are to: (a) review the relevant literature and previous research on diagnosis, feedforward, treatment/intervention/remediation, social mediation, and scaffolding in the fields

of medical, dynamic, and educational assessment; (b) compare theories, methods, and techniques of diagnosis, feedforward, and remediation across these related fields; and (c) develop a unified, theoretical framework for evaluating the quality of diagnosis, feedforward, and remediation in DLA and eventually validating DLA.

Multimodality: a new construct in writing assessment

Duygu Candarli (University of Southampton)

Wednesday, July 3, 09:30 to 10:00 am Location: Aula

This presentation focuses on target language use (TLU) domain analysis of successful multimodal graduate-level writing. Despite the increasing multimodal nature of written assignments in higher education (e.g., Lim & Polio, 2020), high-stakes language proficiency tests currently involve no multimodal composing. In order to ensure the authenticity of L2 writing tasks in language proficiency tests, it is crucial to analyse multimodal writing assignments that students are typically asked to do at an English-medium university, which is the TLU domain of interest in this study. Drawing on Bachman and Palmer's (2010) framework for conducting TLU domain analysis, this paper addresses the following questions: (1) What are the characteristics of multimodal resources and their rhetorical functions in multimodal discipline-specific written assignments? (2) How do assessment criteria refer to multimodality in discipline-specific written assignments?

The data consist of a representative corpus of multimodal discipline-specific assignments written by master's students at UK universities and the assessment criteria of those assignments. The corpus of multimodal assignments was analysed to determine the frequency and type of multimodal resources and rhetorical functions performed by the multimodal resources. Qualitative coding of the assessment criteria uncovered themes linked with the use of multimodality in written assignments. The findings show that multimodal resources were frequently used in the written assignments for meaning-making in different disciplines, underscoring the importance of the construct of multimodality in writing. The assessment criteria most frequently demanded communication competence demonstrated through multimodal resources. The findings will be used to provide guidelines to develop a construct of multimodality in writing tasks in language proficiency tests.

Human- versus artificial-intelligence-based role-play tasks for the assessment of interactional competence: An applied conversation analytic study

Masaki Eguchi (Waseda University), Kotaro Takizawa (Waseda University), Fuma Kurata (Waseda University), Mao Saeki (Waseda University), Yoichi Matsuyama (Waseda University)

Wednesday, July 3, 09:30 am to 10:00 am Location: HS1

Role-play tasks have gained increasing attention in language assessment as a task format to elicit interactional performance (Roever & Ikeda, 2023; Youn, 2015). One persistent challenge of this type of assessment includes the scalability due to the lack of computerized assessment technology (cf. Ockey & Chukharev-Hudilainen, 2021). Recently, with the rise of spoken dialogue

systems (SDS) and Large Language Models, engineering a role-play interlocutor has been made possible for specific assessment contexts (Saeki, et al., 2021). However, very little is known how an AI-based interlocutor compares to its human counterpart in terms of detailed moment-by-moment interaction. In this study, following the applied conversation analytic (CA) studies (e.g., Galaczi, 2008; Seedhouse & Nakatsuhara, 2018), we examined the nature of the interaction in the state-of-the-art SDS for role-play assessment tasks by comparing it with human examiners.

A total of 75 EFL speakers in a private university in Japan participated in this study. They completed a total of three role plays and one discussion task twice (once with human interlocutors and with SDS for another time) in a counterbalanced order. Then, we analyzed 20 interactions (10 from each interlocutor type) for their interactional features, such as turn-taking, sequential organization, repair sequences, etc. The analysis revealed features that are shared between the two interlocutor type (e.g., overall sequential organization; delayed request sequences by the user) while highlighting aspects of interaction that deserves attention (e.g., turn-taking). These findings are discussed in relation to the designs of the AI-based role-play tasks and the target construct representation.

Exploring the language and communication demands of early childhood and school teachers in Australia: Implications for language assessment for teacher registration

Xiaoxiao Kong (University of Melbourne)

Wednesday, July 3, 09:30 to 10:00 am Location: HS2

Language assessments are increasingly used for professional accreditation purposes, however the validity of this practice is rarely investigated. From 2011, both overseas teachers and international graduates of teaching degrees are required to achieve set scores in an approved English proficiency test to register and work as a teacher within the Australian early childhood and school contexts (Australian Institute for Teaching and School Leadership, 2011).

The current study investigates the linguistic and communicative demands of early childhood and school teachers in Australia, as well as the appropriateness and adequacy of the IELTS Academic, the only English language proficiency test approved by all Australian jurisdictions for teacher registration, for assessing English language proficiency for teacher registration. Data were collected from an analysis of teachers' position description documents at the three education levels (i.e., early childhood, primary, secondary; n = 106) as well as focus groups with teachers (n = 37) on their workplace language demands as well as their views of the appropriateness and adequacy of the IELTS Academic for the assessment of teacher language proficiency. Findings suggested differences in teachers' language demands across education levels, as well as a mismatch between the IELTS Academic and teachers' language demands in terms of task format and characteristics of expected response, raising concerns over the validity and appropriateness of the IELTS Academic for teacher registration purposes. Such investigations provide implications for policy formulation as well as the design and implementation of language assessment for teacher registration, which could in turn contribute to student outcomes.

Young EFL learners' cognitive processes of taking digitalized picture-based causal explanation speaking tasks: Linking eye gaze with speech production

Wenjun (Elyse) Ding (University of Bristol & Oxford University Press), Guoxing Yu (University of Bristol)

Wednesday, July 3, 09:30 to 10:00 am Location: HS3

There is a strong need to make standardized language assessments cognitively appropriate for young learners and consistent with the current instructional foci in English as a Foreign Language (EFL) education. This study investigated the cognitive processes of picture-based causal explanation speaking tasks (CESTs). Ninety-six Chinese primary-school EFL learners in Grade 4 and 6 (ages 9 to 12) completed two computerized CESTs in Chinese (L1) and English (L2) under examination conditions. Their eye movements during the test were recorded. We examined to what extent L1 performance scores and L1 cognitive processes differ across two grades (Grade 4 and 6), and to what extent L2 cognitive processes are related to L1 cognitive processes, L2 performance scores, and L2 productive and receptive vocabulary sizes, and grade levels. It was found that 4th and 6th graders had high scores in L1 performance and had similar L1 cognitive processes, which confirmed that it was within the young learners' L1 capacity to complete the CESTs. In the L2 tasks, participants with fewer linguistic resources in L2 were found to view both the content-relevant area and the areas not directly related to the content significantly longer and more frequently. The findings of this study pointed to the dynamic complexities of the interactions at different stages of speech production between young language learners' cognitive ability and L2 proficiency and the visual and textual stimuli of the CESTs. We discussed the implications of the findings and directions for future research with reference to task design and cognitive processes of CESTs for young language learners.

Implementing Formative Assessment in the Chinese University EFL Classroom: Understanding Students' Perceptions

Qiaozhen Yan (Chongqing University), Xiangdong Gu (Chongqing University)

Wednesday, July 3, 09:30 to 10:00 am Location: UR3

Formative assessment has gained increasing prominence in the field of L2 education, owing to its great potential in facilitating students to become self-regulated learners. While research on formative assessment in L2 contexts has extensively examined teachers' beliefs and practices, how L2 students perceive formative assessment remains under-explored. This is a critical gap because students' perceptions of assessment can exert a significant impact on their approaches to learning and learning outcomes (Van der Kleij & Lipnevich, 2021).

Situated within a college English course at a Tie One university in southwest China, this study investigated the implementation of formative assessment from students' perceptive, focusing specifically on their perceived benefits and challenges of formative assessment. Five major formative assessment activities were implemented, including writing learning journals (a self-assessment tool), clarifying assessment criteria, conducting peer-assessment of students' course essays, conducting self-assessment of course essays, and providing teacher feedback on course essays. These activities were designed on a basis of Black and Wiliam's (2009) theoretical

framework for effective formative assessment. Adopting a qualitative approach, the researchers collected semi-structured written reflections from 47 students at the end of the course. The data was analyzed using a thematic analysis approach.

The study demonstrates the potentials of formative assessment in improving EFL students' self-regulation and language proficiency, as well as in enabling them to develop new perspectives on language assessment. It also highlights an array of challenges that EFL students face in formative assessment, pertaining to students' characteristics, assessment tasks, and the school and system issues.

Processing of multimodal input – Towards a more comprehensive definition of integrated writing assessment

Sonja Zimmermann (g.a.s.t./TestDaF-Institut)

Wednesday, July 3, 10:00 to 10:30 am Location: Aula

Even though scholars have called for it already a decade ago (e.g. Cumming, 2013; Knoch & Sitajalabhorn, 2013), a fundamental theory or model of writing from sources in the L2 is still missing. To contribute to a more comprehensive understanding of integrated writing, the current paper takes a closer look at an integrated writing task with multimodal input material in the context of a digital large-scale language test for admission purposes in Germany. The following research questions were addressed:

RQ1: What are the cognitive processes test takers use when summarizing information from written and graphical input material?

RQ2: How are the different sources processed with regard to source use and integration style?

A mixed-method approach consisting of a combination of eye-tracking and stimulated recall as well as text analysis was applied to investigate the cognitive processes and task performance of 14 international study applicants. The analysis of process data focused on viewing behaviour in relation to different Areas of Interest (AOIs), the written products were analyzed linguistically and with respect to content.

Findings indicate that the task elicits a specific interaction of basic reading and writing processes, and that the use of cognitive processes and the utilization of the two sources varied at different stages of the writing process. A closer look at the written products showed differential effects on task performance depending on the mode of task input.

The paper finally discusses implications regarding the expansion of integrated writing models across modalities, genres and language settings.

Exploring the Potential of Conversational AI for Assessing Second Language Oral Proficiency

Yasin Karatay (Cambridge University Press & Assessment), Jing Xu (Cambridge University Press & Assessment)

Wednesday, July 3, 10:00 to 10:30 am Location: HS1

Interactional competence is an integral part of oral proficiency, but it is missed by computer-delivered, semi- Interactional competence is an integral part of oral proficiency, but it is missed by computer-delivered, semi-direct speaking assessment (Author, 2021). Oral interviews facilitated by an interlocutor can simulate real-life social interactions, but they are resource-intensive, costly, and thus difficult to scale up (Timpe-Laughlin et al., 2022). Additionally, it remains a challenge to balance between interlocutor standardisation with authenticity, ensuring that candidates are examined under similar conditions while enabling natural and free-flowing interactions (Seedhouse & Nakatsuhara, 2018).

Conversational AI or spoken dialogue systems (SDS) designed to mimic the role of an interlocutor in real-time digital oral communication seem promising to mitigate the above-mentioned challenges that speaking assessment face. Employing a mixed-method research design, this study investigates how well a prototype SDS-mediated speaking test can elicit interactive oral language functions and simulate the face-to-face counterpart of the test. Thirty non-native English speakers sat the prototype test, and their oral performances in response to SDS were marked by two trained raters using a mark scheme adapted from the face-to-face test. Results of retrospective interviews are reported to explore candidate and rater perception of the SDS performance, the rateability of the speech samples, and the authenticity of the human-computer interaction. Interactive language functions identified in the speech transcriptions are compared with those observed in the human-mediated test. Finally, inter-rater reliability, descriptive statistics of survey data and thematic analysis on the interview data shed light on the efficacy of conversational AI for speaking assessment. Implications for using SDS in low- and high-stakes speaking assessments are discussed.

Assessing the language proficiency of internationally-graduated professionals: The intended vs. actual interpretations

Shahrzad Saif (Université Laval)

Wednesday, July 3, 10:00 to 10:30 am Location: HS2

This study investigates the language proficiency assessment of newly arrived health professionals in Canada as part of the recognition process of their credentials. Adopting a ‘discrepancy view of needs’ (Brown, 2016), the study explores the discrepancies between the tasks and constructs currently assessed and those that should be targeted to ensure the quality of healthcare communication. Data was gathered from multiple sources (including entry-level competencies, professional standards, federal and provincial publications) using a target-situation analysis method. The results were summarized thematically reflecting different aspects of the language proficiency assessment of health professionals as they transition from the pre-arrival phase to

early arrival, and then to full integration into the system. Stakeholders' feedback on key themes emerging from this analysis was also sought.

The findings provide insights into the complexity of choosing assessment tools whose score interpretations can lead to decisions with beneficial consequences for the quality of healthcare communication and for medical professionals' licensure (Bachman & Palmer, 2010). They point to a multitude of currently used language tests with diverging goals, varying tasks, and differing contents. More importantly, they underscore language abilities—integral to the successful functioning of health professionals—that current general-purpose testing tools fail to measure. Results also show unnecessary duplications due to overlapping language requirements by different government agencies. The study recommends that, in addition to the target situation communicative needs, the nuanced pattern of practice influenced by, or in response to, government policies/requirements be considered when choosing language tests for assessing health professionals' proficiency.

Young EFL Students' Writing Performance: Patterns by CEFR Levels and Task Types

Mikyung Kim Wolf (Educational Testing Service), Michael Suhan (Educational Testing Service)

Wednesday, July 3, 10:00 to 10:30 am Location: HS3

The present study examined the performance on a standardized writing assessment of young EFL students (aged 8-12) from various countries. Specifically, we aimed to investigate how students' writing differed across different Common European Framework of Reference for Languages (CEFR) levels and task types. Although the CEFR is widely used, its relation to young learners' writing patterns remains relatively underexplored. Identifying salient characteristics in their writing across different proficiency levels and tasks will enhance our understanding of their writing development and inform appropriate instructional and assessment approaches.

The data for this study (N = 1,157) were sampled from the computer-based field-test portion for the development of the TOEFL Primary® Writing test. The field-test form consisted of five task types including different response formats. An automated writing evaluation engine, trained on human ratings with a 0-3 rubric, scored students' responses. Based on their total scores, students' writings were classified into CEFR levels. Natural language processing (NLP) analyses were performed across CEFR levels and task types.

The results indicated that students' performances varied by task type and CEFR level, with different patterns emerging for higher-level students. NLP analyses further highlighted task-specific trends and linguistic complexities in their writing. For instance, clausal complexity did not increase linearly across CEFR levels, whereas phrasal and lexical complexity, grammatical accuracy, and use of cohesive features did, suggesting young EFL students' developmental patterns. This presentation will share the detailed results and implications of using automated tools in young learners' writing assessment and research.

Reforming teacher education to enhance language assessment literacy: New insights from pre-service teachers' reflections

Armin Berger (University of Vienna), Helen Heaney (University of Vienna)

Wednesday, July 3, 10:00 to 10:30 am Location: UR3

Exam reforms in educational systems necessitate corresponding changes in teacher education. In Austria, for example, the nationwide introduction of standardized school-leaving exams, coupled with a shift from impressionistic to analytic rating practices, has significantly influenced secondary-level teacher education in English. Some fifteen years after the initial reform, the rating scales for assessing writing in foreign language exams have been overhauled, underscoring the imperative to align teacher education with ongoing exam reforms.

Such reforms have spurred meta-reflection on language assessment literacy (LAL) for teachers. While early research concentrated on componential perspectives of LAL, recent approaches emphasize developmental views, focusing on teacher competencies at specific levels. This paper, situated at the intersection of developmental perspectives, the localization of LAL practices, and the quest for successful LAL pedagogies, complements previous measurement-driven research into the difficulty of assessment-related activities. It utilizes qualitative content analysis to explore 848 guided reflections from pre-service English teachers at the University of Vienna, Austria. The thematic coding scheme centres on the reflection topics, the quality of the reflections, and the accuracy of the propositions, particularly in relation to rater training and analytic rating scales, enabling further conclusions on the difficulty of specific LAL aspects.

Aligning with LTRC's 2024 vision, the paper addresses exam reforms through the lens of teacher education. It illustrates how the research findings can be used to reform LAL instruction, emphasizing the catalytic nature of exam reforms and the transformative impact of LAL research.

Cooccurrence of Disfluency Features of L2 Speech across Proficiency Levels in Controlled and Spontaneous Tasks

Yulin Pan (University of Illinois Urbana-Champaign)

Wednesday, July 3, 03:30 to 04:00 pm Location: HS2

Cognitive validity has been extensively studied in receptive skills but spare in productive skills. Speech disfluency, providing a window to the cognitive processes of speaking, can be used to investigate the cognitive validity of speaking tasks. Traditional analyses often use isolated (dis)fluency features to predict proficiency, while natural speech disfluencies often occur in chains.

The study used Elicited Imitation Tasks (EIT) and oral Listen-to-Summarize Tasks (LST) to examine disfluency co-occurrence in 58 L1 Chinese and 13 L1 English speakers. The methodology involved four steps: (1) manual coding of disfluency chains, (2) Python scripts for type and length extraction of disfluency chains, (3) Principal Component Analysis (PCA) for extracting patterns of disfluency chains, and (4) Hierarchical-based, K-means Cluster Analysis (CA) for identifying speakers' profiles on the use of different disfluency patterns.

On both tasks, the results revealed four disfluency patterns and three speaker profiles. All disfluency patterns and profiles could be interpreted functionally as broad categories of disfluencies associated with type and length of speech breakdown and repair, although the specific categories differed. Contingency tables showed that higher oral proficiency was associated with fewer disfluencies in both tasks. However, while oral proficiency was negatively associated with recall-like/monitoring behavior on the EIT and repair behaviors on LTS, it was positively associated with hesitation behaviors on the LTS. The study highlights that while the two tasks elicit similar speech production processes, they induce unique disfluency features. The results have meaningful implications for using disfluency features to examine the cognitive validity of speaking tasks.

Innovating constructs and assessments: The development and investigation of multimodal viewing-to-write tasks

Tineke Brunfaut (Lancaster University), Judit Kormos (Lancaster University)

Wednesday, July 3, 03:30 to 04:00 pm Location: HS3

Given the pervasiveness and accessibility of technology nowadays, multimodal language use has become increasingly prevalent. Consequently, insights into the nature of multimodal language processing, production, and assessment are vital to ensuring that language testing remains relevant to 21st century communication needs (Shohamy, 2022). In this presentation, we report on a study that aimed to extend the repertoire of tasks and constructs represented in language assessments by developing multimodal viewing-to-write tasks and investigating L2 learners' performance on these.

Viewing-to-write tasks provide integrated auditory-visual-written input and require a written output from test-takers. In our study, we explored two types of tasks: viewing-to-describe tasks and viewing-to-compare-and-contrast tasks. One hundred and thirty-four English-L2 learners aged 15-19 and at CEFR A2-C levels of proficiency each completed two to three viewing-to-write tasks. Task version and order were counterbalanced. Task performances were marked by two raters, using a purpose-developed rating scale. Participants also completed the Aptis listening, writing, grammar and vocabulary tests, as measures of independent skills ability.

To gain insights into the learners' performance on the two types of multimodal tasks, descriptive statistics of students' scores were established for the assessment criteria of viewing-for-writing, organization and structure, language use, and mechanics. Additionally, linear mixed-effects modelling was conducted to understand the interrelationship between learners' scores on the multimodal viewing-to-write tasks and their independent listening and writing abilities, and any potential moderating role of task-type. In our presentation, we report the findings and share insights into the practical feasibility, validity and usefulness of multimodal viewing-to-write tasks for assessment purposes.

How reliable were human raters when assessing second language English prosody? A Bayesian meta-analysis

Yuanyue Hao (University of Oxford)

Wednesday, July 3, 04:00 to 04:30 pm Location: HS2

As a linguistic feature that poses a challenge to both language learners and teachers, prosody (i.e., stress, intonation, and rhythm) has been found to be a significant predictor of intelligibility, comprehensibility, and communicative success in second language (L2) English speech. However, there are several potential issues that could undermine rater reliability, such as different operationalizations of prosody construct, rater background, and rating scales. To understand how these variations might influence rater reliability, this meta-analysis aims to 1) investigate the overall rater reliability in L2 English prosody assessment, and 2) explore to what extent reliability varied according to different operationalizations of construct, rater background, and scale features.

A literature search was conducted in 12 databases and 7 websites of language test providers. 109 studies were judged as eligible. A total of 441 reliability estimates were reported in these studies, and this paper focuses on the inter-rater reliability as assessed by Cronbach's alpha (k = 127) for meta-analysis. A Bayesian meta-analysis was conducted to investigate the overall inter-rater reliability. The overall estimate was 0.92, with 95% credible interval ranging from 0.87 to 0.96. However, the between-study suggests great variations among studies. Results from subgroup analyses indicated that inter-rater reliability was higher when prosody was assessed at the global level than at the specific level, and when the scales were accompanied with specific descriptors than with labels at either endpoint. This meta-analysis calls for further improvement in prosody assessment by clarifying construct and refining rating scales and has implications for automated prosody assessment.

Investigating cognitive strategy use in an intertextual reading-into-writing Summary task through online think-aloud interviews

Nathaniel Ingram Owen (Oxford University Press), Haiyan Xu (University of Leicester), Oliver Bigland (Oxford University Press)

Wednesday, July 3, 04:00 to 04:30 pm Location: HS3

This paper reports on the development and research of an innovative intertextual reading-into-writing summary task for use in academic admissions. Literature on tests of English for academic purposes increasingly advocate integrated assessment (Xi and Norris, 2021). However, tests of English used for university admission largely avoid intertextual reading (Weir and Chan, 2019) and do not ask test takers to synthesize information from multiple texts into a single piece of writing.

We designed a summary task which requires test takers to read two texts on the same topic (a total of 300 words) and to summarize the information (80-100 words). We recruited fifteen university students to participate in online continuous think-aloud interviews (Alhejaili, Wharrad and Windle, 2022) to identify cognitive strategy use. Students' verbalizations were coded in relation to on-screen objects (referents) and the type of strategies that students were using,

evidenced through verbalizations and referent engagement. Strategies were coded using a framework developed from CEFR Companion volume (2020) mediation descriptors and cognitive models of reading (Khalifa and Weir, 2009) and writing (Weir and Shaw, 2007).

We found the task was highly effective at eliciting a range of reading strategies for careful and expeditious reading (Khalifa and Weir, 2009), macro and micro planning to organize ideas and paraphrasing. The findings suggest that intertextual reading requirements elicit appropriate organizational and synthesizing strategy use (List and Alexander, 2020) so can support academic writing based on multiple texts (List, Du and Lee, 2021) and potentially have a positive washback effect (Green, 2013).

The validation and usability of an L2 Chinese prosody rating scale in three speaking task types

Sichang Gao (Shanghai International Studies University), Mingwei Pan (Shanghai International Studies University)

Wednesday, July 3, 04:30 to 05:00 pm Location: HS2

This mixed-methods study examines the effects of task types on raters' judgments of L2 Chinese speech prosody using an empirically developed rating scale. The evaluation of the scale was based on the raters' perception of the descriptors and the application of the three-category rating of learners' prosody performance, which was analyzed using the many-facet Rasch model(MFRM). Thirteen trained raters assessed speech samples of 38 learners from a degree-based L2 Chinese program on numerical rating scales for prosodic strategic competence, prosodic naturalness, and fluency. Three task types are used: read-a-passage, listen-and-repeat, and free talk. A quantitative analysis of raters' perceptions and task effects was undertaken through MFRM. Interviews were administered to elicit the raters' perceptions of the efficacy of the rating scale. MFRM results reveal that the 14 descriptors in the rating scale have good fitness in rating three task types. The scale effectively differentiated between levels of prosodic skills, with fluency descriptors proving easier than those related to prosodic strategic competence. This study confirmed variations in speakers' prosodic abilities across task types, with free talk emerging as a preferred assessment method. Besides, Inconsistencies in severity levels among raters highlighted a disparity between theoretical and perceived L2 prosody constructs. For instance, descriptors related to prosodic naturalness were intricately perceived with fluency and prosodic strategic competence. Recommendations, including "provide benchmark performances," "use more communicative/interactional tasks," and "provide sufficient score subdivision," were proposed for rating scale refinement and future rater training.

Source use patterns in integrated writing tasks: The role of discourse synthesis quality and linguistic features

Atta Gebril (The American University in Cairo)

Wednesday, July 3, 04:30 to 05:00 pm Location: HS3

Discourse synthesis skills, which involve selecting and transforming others' texts and eventually integrating them in one's writing, are essential for adequate performance on integrated writing tasks. While research has suggested that text quality and reading ability could affect decisions made by writers when selecting various source use patterns, surprisingly little research has looked into the relationship between these patterns and quality of discourse synthesis. Along the same line, little research has considered how source use quality could be influenced by both vocabulary. Given this gap, the purpose of this study is to investigate how discourse synthesis quality and language ability could affect source use patterns in integrated writing tasks. A total of 202 Students in an English-medium university completed an integrated reading-based writing task that was holistically and analytically scored by two raters. To analyze the research questions, regression models for count data were applied. The results indicated that integrated writing scores were negatively correlated with relative total source use and relative direct source use, but not significantly correlated with the other source use variables. Reading and source use quality did not predict total source use beyond the other independent variables. A similar pattern was found for indirect source use as a dependent variable, with a slightly stronger effect for vocabulary and a slightly weaker effect for integrated scores. For direct source use, only integrated scores had a significant effect. The study provides a number of implications for language programs, writing instructors, and language testing professionals.

Engagement, emotional valence, and attention: Investigating the impact of facial behavior on speaking test scores

J. Dylan Burton (University of Illinois Urbana-Champaign)

Wednesday, July 3, 05:00 to 05:30 pm Location: HS2

In order to address a long-standing question about the role of non-linguistic features of communication in test performance (Jenkins & Parra, 2003; Plough, 2021), this study investigated the impact of three measures of nonverbal behavior on language proficiency test scores. One hundred novice raters evaluated 30 recordings of individuals taking a language proficiency test. These recordings varied in proficiency levels, and raters scored them based on fluency, vocabulary, grammar, and comprehensibility using seven-point scales. Nonverbal behavior was measured using a software called iMotions, producing three metrics: facial muscle activation, behavioral valence, and attention. Results indicated that the impact of nonverbal behavior on scores differed based on the proficiency level of test-takers. For instance, higher variability in attention corresponded to lower scores in the lower proficiency group but higher scores in mid and high proficiency groups. Similarly, positive emotional behavior related to higher comprehensibility scores but only in the lower proficiency group. However, the influence of nonverbal behavior on scores was minimal, explaining only about 2% of the score variance. Although nonverbal behavior did not uniformly affect all proficiency groups, these findings suggest a potential for behavior to subtly influence raters' perceptions of proficiency, potentially

impacting borderline candidates' scores. This study has important implications for generating a more nuanced understanding of non-linguistic features that relate to communicative language ability as well as methodological implications for the use of behavior analysis software in research and practice.

Sequence analysis of log data: an application example from a study of integrated writing

Ximena Delgado-Osorio (DIPF; Leibniz Institute for Research and Information in Education), Valeriia Koval (University of Bremen), Johannes Hartig (DIPF; Leibniz Institute for Research and Information in Education), Claudia Harsch (University of Bremen)

Wednesday, July 3, 05:00 to 05:30 pm Location: HS3

With the increasing interaction of test-takers with technology during language testing, the complexity of the interactive processes also increases. Therefore, the study of test-takers' processes can benefit from systematic data extraction and analysis approaches. In this context, the extraction of log data allows researchers to understand the complex performance of test-takers in an online environment. In writing research, the use of keystrokes to explore writing processes and strategies has increased and has shown associations with test-taker characteristics and outcomes (e.g., Talebinamvar & Zarrabi, 2022; Révész et al., 2022; Zhu et al., 2019). The sequence analysis method is an exploratory and descriptive approach for analyzing long-term or process data (Raab & Struffolino, 2023), allowing the identification of trajectory patterns that explain processes of interest over a period of time (Studer & Ritschard, 2016).

In our study, sequence analysis was used to investigate the processing of integrated writing. After extracting log data from 601 integrated writing tasks completed by high school and university students, we conducted sequence analyses using the R package TraMineR (Gabadinho et al., 2011) to identify clusters and perform descriptive and variance analyses. The results showed four patterns in the processing of integrated writing tasks: rapid, strategic, note-taking, and low-interaction. We also found a relatively high association between the patterns shown in two integrated tasks by the same test-taker. Using our study as an example, we highlight the potential advantages and limitations of sequence analysis of log data for use in language assessment research.

Paper and Demo Summaries – Thursday, July 4

Language testers as policymakers

Laura Schildt (Ghent University)

Location: HS1

Thursday, July 4, 08:30 to 09:00 am

Language testers wear many hats including those of test developer, test administrator, statistician, psychometrist, and, increasingly, policymaker. Considering these expanding roles, it is worth exploring the identities of language testers. Identity is central to SLA research, yet virtually no studies have explored the identities of language testers. This paper seeks to address this research gap by using narrative analysis to examine the identity navigation of language testers in the policy context. Identity navigation refers to the ways individuals manage and adapt their identities in response to different social or professional environments. From 26 interviews, 38 narratives about interactions between language testers and policymakers were identified. The narratives were positioned on two dimensions of identity navigation theory: sameness/difference (between self and others) and agency/control. Four narratives were chosen to represent emblematic positionings of language testers. The analysis of these narratives was guided by discourse analysis and took into account a wide variety of linguistic and multimodal features. The findings reveal the importance of LAL to language tester identities and group belonging. Language testers mark out affiliations and position themselves as aligned with or in contrast to other actors in the policy world. Their linguistic choices point to the kinds of agency they attribute to themselves as they navigate a spectrum of neutrality/partiality regarding the use of their tests and their role as experts.

Construct relevant or irrelevant? The impact of background noise on listening comprehension

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

Thursday, July 4, 08:30 to 09:00 am Location: HS2

Listening comprehension in authentic communication settings presents a challenging skill for L2 speakers. Previous research has explored the impact of input complexity and task authenticity on listening comprehension. However, the impact of background noise, a key element of authentic oral communication, on L2 listening comprehension remains understudied. Research in speech and hearing science has indicated that processing of speech in noise imposes challenges to both native and non-native speakers, but the challenge intensifies as proficiency level decreases. This experimental study investigates how various types and levels of noise affect L2 listening comprehension and the psychometric qualities of associated tasks.

The study employed retired monologic and dialogic listening passages spanning A2 to C1 levels from the Cambridge English Assessment Suite. Each passage included three to five multiple-choice listening comprehension tasks, totaling 112 items. Noise-infused materials were created using primarily speech-shaped noise, known for its stationary nature resembling speech without containing language information. We designed a web-based platform to facilitate assessment

delivery, performance recording, and post-test questionnaire response collection. 237 undergraduate and graduate students from three Chinese universities participated in the study, and their item responses were subjected to Rasch analysis to examine listening comprehension between noise-infused and noise-reduced conditions.

We will first present results regarding participants' performance in noise-reduced versus noise-infused conditions, supplemented by perceptions of task difficulty and face validity. This exploration aims to deepen the understanding of how L2 listeners navigate speech in noise and whether background noise contributes construct-relevant variance, potentially enhancing the psychometric quality of listening tasks.

Building an argument for test score interpretation and use for a fully automated online assessment of L2 spoken interaction

Yasuyo Sawaki (Waseda University), Yuya Arai (Waseda University), Masaki Eguchi (Waseda University), Shungo Suzuki (Waseda University), Yoichi Matsuyama (Waseda University)

Thursday, July 4, 08:30 to 09:00 am Location: HS3

Recent advancements of conversational virtual agent and speech recognition technologies now enable AI-based conversational assessment task designs that elicit synchronous and reciprocal human-computer interaction, better reflecting the traditionally under-represented construct of interactional competence (Galaczi & Taylor, 2018) than conventional monologic speaking tasks. At the same time, however, such novel assessment designs generate various questions concerning construct validity, requiring careful investigations into the soundness of intended score interpretation and use.

This paper presents an overview of the validation framework based on Bachman and Palmer's (2010) assessment use argument (AUA) and summarizes initial empirical evidence to support the design of a fully automated placement assessment of L2 spoken interaction for a campus-wide English language program at a private university in Japan. In this multi-stage adaptive speaking test launched in spring 2023, students interact with a conversational virtual agent to complete interview and roleplay tasks online. Through automated scoring, CEFR-based (Council of Europe, 2001) analytic scores are reported immediately for diagnostic feedback and placement. In this session, we will focus specifically on two AUA claims directly related to construct validity: score interpretations and assessment records. We will present warrants and rebuttals associated with these claims and corresponding empirical evidence, featuring a factor analytic study examining the degree to which two analytic scores (Range and Fluency) generated from the machine-learning scoring engine can be explained by SLA-research-based microfeatures obtained from task performance data. Study implications and pivotal roles the AUA has played in this complex project involving multiple stakeholder groups will be discussed.

Delayed measures of speaking proficiency: Questioning assumptions

Anastasia Ulicheva (Pearson), Sumita Ishaque (Pearson), Rose Clesham (Pearson)

Thursday, July 4, 08:30 to 09:00 am Location: UR3

Test-takers are often given time to prepare their speaking responses in high-stakes language tests. It is generally believed that the delay discourages impulsive responding and thus creates a truer picture of the examinee's language proficiency.

Interestingly, cognitive studies of bilingual language processing show that L2 learners differ considerably from native speakers in their immediate responding, but in delayed responding, those differences fade away (Broos et al., 2021). Thus, it can be hypothesised that discriminating among different L2 proficiency levels may require immediate and not delayed responding, at least in some situations.

In this study, we tested L1 and L2 speakers of English from across a range of proficiency level as assessed by a high-stakes test. They completed several speaking activities in English, including immediate and delayed picture naming alongside extended-response tasks. Firstly, we tested whether proficiency differences are reflected in immediate as opposed to delayed naming latencies. We then established what set of measurements (immediate or delayed) correlated best with the scores on speaking tasks.

Our findings shed light on whether specific psycholinguistic measures can be useful at discriminating language proficiency levels and for revealing processes that underpin speaking performance. We also discuss what influence the addition of preparation time may have on the way we interpret measurements obtained from those responses, in theoretical and practical terms.

A Theory of Action in Working for Social Justice

Cecilie Hamnes Carlsen (Høgskulen på Vestlandet), Lorenzo Rocca (Società Dante Alighieri), Nick Saville (Cambridge University Press & Assessment), Graham Seed (Cambridge University Press & Assessment)

Thursday, July 4, 09:00 to 09:30 am Location: HS1

For more than 30 years, the Association of Language Testers in Europe (ALTE) has pursued its mission in the world of multilingual language assessment to set standards while sustaining diversity. ALTE engages with different stakeholders, including politicians, policy-makers, employers and language testers themselves, to extend existing concerns for equity, diversity and inclusion with issues of social justice. To achieve this, ALTE has developed many tools, courses, discussion forums, international events, and publications, to provide a theory of action. In this paper, this approach is illustrated by the work of ALTE's Special Interest Group, Language Assessment for Migrants' Integration (LAMI).

Since the early 2000s, LAMI has dealt with language assessment in the context of migration and integration and has grappled with issues of fairness and social justice from theoretical and practical perspectives. In this paper, we refer to three recent LAMI projects that have produced

'tools' that aim to achieve not only fairer assessment practices, but more socially just uses of language assessment.

The first is the development and validation of a new minimum quality standard (MS18) within the ALTE audit system, which requires test developers to provide validity evidence for mitigating against test misuse. The second are the multilingual LAMI-LASLLIAM Assessment Tools (LLAT), four multilingual tools to illustrate the use of assessments for low-literate vulnerable learners. The third is the Uneven Profile Report, detailing results of a survey on uneven profile testing requirements and practices across Europe, together with examples of good practice.

Equality, Diversity, and Inclusion in Practice: Candidate Reactions to Global English Accents in a Listening Test

Gemma Bellhouse (British Council)

Thursday, July 4, 09:00 to 09:30 am Location: HS2

While variation in English includes a host of linguistic and non-linguistic features, the significant role that accent plays as a marker of identity, socio-economic disparity and overall prestige, and the potential it has to be exclusive is broadly recognised, see Sung (2016). Within the area of language testing and assessment, the issue of accent, both from the developer's and test takers' point of view is critical (Nishizawa, 2023).

The presentation addresses the issue of heard accent in the listening section of an English language test from the perspective of Equity, Diversity and Inclusion (EDI). The first section of the presentation will briefly outline a pair of EDI policies. The candidate focused policy details special accommodations, while the test-focused policy relates to test content (e.g. representation in images, accents in audios). The listening paper we focus on includes tasks with audio input comprised of monologues and dialogues using a multiple-choice format and including a variety of accents heard across the UK. We then present a study in which over 700 candidates were asked to reflect on the accents they were exposed to during the test, revealing a range of reactions to the accents used.

The presentation aims to encourage discussion on issues of EDI policy implementation and test standardisation from the perspective of the key test stakeholders. While recognising the practical considerations inherent in producing listening tests that are accent-inclusive, the data from this study supports our commitment to an incremental shift towards this goal within the tests we develop.

Automated scoring and validity: Expanding evidence through explainability

Sarah R. Hughes (University of Cambridge)

Thursday, July 4, 09:00 to 09:30 am Location: HS3

Automated scoring (AS) technologies are increasingly prevalent in language assessment. These technologies have advanced greatly in the past decade, however, the methods available for evaluating the validity of scores produced by AS systems have not advanced at the same pace.

The primary measure of automated assessment quality continues to be agreement with human raters. If the AS system produces the same score as human raters produce, this is offered as evidence that the AS system accurately predicts human scores and is therefore working as intended. However, evidence of predictive accuracy is not the same as evidence of construct validity, particularly when the considerable predictive power of AS systems often comes at the cost of transparency. The most highly performing models are complex, opaque, and proprietary.

This presentation explores recent developments in eXplainable Artificial Intelligence (XAI) research that have the potential to offer new ways to evaluate the construct validity of AS decisions by increasing their transparency and explainability. A study was designed to evaluate the utility and limitations of these techniques in the context of essay scoring. The study used the Automated Student Assessment Prize (ASAP) essay dataset, a publicly available deep learning AS algorithm, and an XAI technique known as SHAP to generate explanatory information. This presentation reports on the findings from this study and discusses the alignment between the SHAP-generated explanatory information and human raters' understanding of the assessed construct. This presentation will be relevant to anyone interested in the responsible use of AS systems for language assessments.

Analyzing Argumentative Skills in Foreign Language Learners: Integrated Task Assessments and Rhetorical Moves Analysis

Jorge Luis Beltran Zuniga (Teachers College, Columbia University)

Thursday, July 4, 09:00 to 09:30 am Location: UR3

This paper explores two integrated tasks assessing argumentative skills in second language learners. Task A (Evaluate Arguments) required critiquing forum post arguments based on guidelines, showcasing learners' comprehension, analysis, and topical knowledge application. Rhetorical move analysis identified four moves within this task, delineating stages like framing, evaluating, backing, and closing the evaluation. Task B (Address Counterargument 2) focused on orally responding to a peer's counterargument, examining learners' abilities to comprehend content and persuasively defend their stance. Responses followed stages: acknowledging the counterargument, refuting with concession and rebuttal, and reinforcing the argument or concluding. Both tasks assessed higher-order argumentative competencies alongside language proficiency, offering insights into learners' comprehension, evaluation, and response skills. The analysis provides valuable implications for language assessment and teaching strategies.

Language and knowledge of society tests for citizenship: implications for vulnerable migrant groups

Marieke Vanbuel (Ghent University), Edit Bugge (HVL Bergen)

Thursday, July 4, 09:30 to 10:00 am Location: HS1

In many Western countries, aspiring citizens are required to pass oral second language tests (L2) at specified proficiency levels and Knowledge of Society Tests (KoS), often administered in the L2. This study investigates the impact of these requirements on migrants, particularly Low-

Educated Second Language and Literacy Acquisition (LESLLA) learners with limited prior formal education and literacy experience in their home countries. For LESLLA learners, language and KoS tests can serve as significant barriers to obtaining citizenship.

Anchored in the IMPECT research project, this research uses an administrative dataset from Norway covering test results of 79,794 migrants between 2017 and 2020, to address questions like 'To what extent does educational background influence the probability of passing the language test and KoS test?' and 'What is the role of second language (L2) skills in determining success on KoS tests?'.

Using two-level logistic regression analyses, the findings reveal that LESLLA learners have significantly lower pass probabilities compared to candidates with secondary or tertiary degrees. The disparity in pass probabilities is more pronounced for KoS tests than for the language test. Language skills strongly predict success on KoS tests, confirming the concept of implicit language tests. Furthermore, the influence of language skills on KoS test success varies by educational backgrounds, with the language test being a more reliable predictor for LESLLA learners compared to candidates with tertiary degrees. These results emphasize the need for nuanced policy considerations and support mechanisms to facilitate LESLLA learners' integration and attainment of citizenship.

What makes listening comprehension difficult?: A feature-based machine learning approach to understanding item difficulty

Huiying Cai, Ping-Lin Chuang, Yulin Pan, Mingyue Huo, Xun Yan (University of Illinois Urbana-Champaign)

Thursday, July 4, 09:30 to 10:00 am Location: HS2

Item difficulty in L2 listening assessment can be affected by textual and acoustic features of both listening inputs and items. Traditional statistical approaches use regression models to predict difficulty based on various features. In contrast, machine learning (ML) allows for more generalizable models with consistent predictive performances across datasets, offering a broader understanding of item difficulty. This study builds a feature-based ML model, incorporating textual and acoustic features and extra-linguistic features, to predict difficulty in 225 multiple-choice listening items from Taiwan's General English Proficiency Test.

We extracted 950 textual (i.e., lexical/syntactic complexity and similarities among options, stems, and stimuli) and acoustic features (i.e., pronunciation and fluency) at the option, stem and stimulus levels. By considering different item types, we selected two types of features: item-type generic features and item-type specific features. For each feature type, we further reduced data dimension through manual removal and principal component analysis, yielding four different feature sets. We subjected each feature set to mixed-effects ridge regression models, along with extra-linguistics features (i.e., test focus and item type), and compared their performances. The best-performing model employed 28 item-type generic raw features ($R2 = 0.88$).

Results indicated meaningful relationships between item difficulty and lexical/syntactic complexity, similarities among options, stem and stimuli, pronunciation, test focus, and item type. The findings highlight how these features influence item keys and distractors, offering

insights for item difficulty modeling and distractor writing. This study underscores the effectiveness of integrating computational linguistics and ML in L2 listening assessment research.

Evaluating score accuracy for an automated scoring system in a high-stakes writing test

Trevor Breakspear, Edmund Jones, Shilin Gao, Trevor Benjamin, Jing Xu (Cambridge University Press & Assessment)

Thursday, July 4, 09:30 to 10:00 am Location: HS3

We evaluated an automated scoring system for a high-stakes English writing test. In our approach, the "reference score" for a response is the Rasch fair average of the scores given by multiple examiners. We compare the scores given by single examiners to the reference scores, and compare the scores given by the automated system to the reference scores. The automated system is only accepted if it performs at least as well as the single examiners.

Our research questions were: How well do scores from the automated system agree with reference scores, overall and in different parts of the score scale? How similar is the distribution of scores from the automated system to the distribution of the reference scores?

We measured agreement using root mean squared error (RMSE) and agreement on CEFR levels. For agreement in different parts of the range we calculated RMSE within bands (to ensure fairness across the spectrum of candidate ability). For distribution similarity we used earth-mover's distance.

We also used importance weighting, which makes it possible to use the same evaluation dataset for different populations. This involves choosing a target distribution; for example, a uniform target distribution would imply that all parts of the score range are equally important. Our target distribution was the empirically observed distribution of scores.

The automated system equalled the performance of individual examiners. Nevertheless, for the operational test we plan to use a hybrid system. Examiners are used when we are less confident that the automated system's scores are sufficiently accurate.

Evaluating General Language Proficiency Speaking Test Assessment Criteria: Evidence From Non-Language Specialists

Curtis Gautschi (Zürich University of Applied Sciences)

Thursday, July 4, 09:30 to 10:00 am Location: UR3

The General Language Proficiency Testing (GLPT) industry provides certification of foreign language ability, playing an important role in the personal mobility and employment opportunities of non-native speakers. GLP speaking criteria, however, are not based on the perceptions of those representing post-test settings, but rather on language-expert intuition and judgments. Since a) the implied context of such tests is beyond the language classroom, b) non-language professionals best represent those involved in communicative acts in such contexts and

c) the perspectives of non-language professionals are not considered, the counterclaim that current assessment criteria may not represent the intended context of post-test context of language use is a potential risk. This, in turn, may compromise the meaningfulness of test scores, which is crucial to test validity.

This paper presents a multi-phase mixed-methods study which investigated the judgments of non-language professionals (both native and non-native speakers of English) as the basis for a new approach to the validation of current GLP speaking test assessment criteria. The main findings call into question the meaningfulness of the GLPT construct for the non-teacher non-native speaker group (NTNNS). This, in turn, suggests the need to develop an English as a lingua franca (ELF)-specific test construct, with rating criteria and test factorial structure that better represent NTNNS perceptions of language proficiency and communication ability. This study adds to ongoing research that advocates for greater integration of post-test language use context representation in the definition of testing constructs.

The sufficiency question: untangling relevance, representativeness and sufficiency

Ute Knoch (University of Melbourne), Susy Macqueen (Australian National University)

Thursday, July 4, 10:00 to 10:30 am Location: HS1

Fundamental principles in language testing and assessment are that the construct should capture all the important aspects of language skills/knowledge (i.e., avoid construct under-representation) and that nothing irrelevant should interfere with the sampling of these aspects (i.e., construct-irrelevant variance) (Messick, 1996). Although these principles of "relevance and representativeness" are widely accepted, when designing or evaluating tests for particular uses, a further question is how much of any domain-relevant skill or knowledge constitutes an adequate sample as a basis for score meaning: the sufficiency question. Sufficiency is "the degree to which the interpretation provides enough information for the decision-maker to make a decision" (Bachman & Palmer, p. 119). In this paper, we present a sufficiency framework in language tests and then apply it to four standardised English language tests used for study and migration purposes. The sufficiency framework sets out various dimensions of sufficiency for productive and receptive skills which are evaluated in relation to both test purpose and key levels in scored-based decisions. Our review of four standardised English language tests in light of the sufficiency framework shows that all fall short on some dimensions of sufficiency for migration and study decisions. To explain our findings, we draw on regulatory capitalism (Levi-Faur 2005), whereby market pressures to offer tests at competitive rates interact with the length and method of a test, its stated purpose, and its use as a regulatory instrument in migration policy.

The road to understanding in lecture listening: how students integrate auditory and textual information

Nicola Latimer (CRELLA; University of Bedfordshire), Daniel Lam (University of Glasgow), Chihiro Inoue (CRELLA, University of Bedfordshire), Sathena Chan (CRELLA, University of Bedfordshire)

Thursday, July 4, 10:00 to 10:30 am Location: HS2

Academic lectures nowadays are predominantly delivered with lecturers' speech and accompanying slides. However, many tests of academic listening still have audio as the primary source of information. Technological advancements in test delivery afford the opportunity to develop new test tasks that more closely replicate real-world academic listening behaviour. However, this necessitates an in-depth understanding of the nature of the lecture listening construct, which concerns how auditory and textual information are integrated in lecture delivery and in lecture comprehension.

This presentation reports on a mixed-methods two-phase study of the lecture listening construct. In Phase 1, the study collected recordings of five academic lectures. The discourse relations between the lecturer's speech and the slide text were analysed using a framework adapted from Hallewell and Crook (2019). In phase 2, the study used eye-tracking and stimulated recall to investigate how students integrated auditory and textual information to develop a mental representation of lecture content. Eight university students at CEFR B2 and C1 levels watched two short online lecture clips from Phase 1, with slides containing a) topic headings only or b) topic headings and summaries of teaching points.

We will discuss two discourse relations in lecture delivery identified in Phase 1. We will then discuss Phase 2 findings from the eye-tracking and stimulated recall which provided new insights into the complex interplay between reading and listening, and factors impacting the difficulty of lecture listening comprehension. The presentation concludes by considering the study's implications for the development of new assessment formats for lecture listening.

Exploring two novel applications of Generative AI in Automated Essay Scoring

Jing Wei (MetaMetrics), Alistair Van Moere (MetaMetrics, The University of North Carolina at Chapel Hill), Steve Lattanzio (MetaMetrics)

Thursday, July 4, 10:00 to 10:30 am Location: HS3

Large scale field testing is a logistically expensive and time consuming part of the test development cycle. Often, field test data are sub-optimal due to sampling, class imbalance, or participant motivation. This creates a challenge if the goal is to build automated essay scoring (AES) models using students' writing data. This study has two research questions: 1) Can synthetic data generated by GPT effectively replace real student data for training AES models? 2) Can we use GPT to directly predict student writing scores, thus bypassing traditional AES modeling altogether?

The baseline data was drawn from writing responses from 2,050 EFL students participating in a large-scale assessment. To address the first question, we utilized GPT-3.5-turbo prompt engineering to generate a synthetic dataset of 1,000 responses. Two regression models were trained, one with the synthetic data and the other with real student data. For the second research

question, GPT-3.5-turbo was fed the rating criteria and instructed to directly score student writing, and then GPT scores were compared with human scores. Preliminary results demonstrate a comparable performance on models trained with synthetic data (r = 0.79) versus real student data (r = 0.80) and a similarity in scores generated by GPT versus by human raters (r = 0.84).

This research is the first empirical study that demonstrates the potential of generative AI to enhance efficiency of automated writing assessment. We finish by discussing factors for consideration in a validity argument for assessments integrating generative AI.

Conceptualizing and operationalizing the construct of critical thinking in EAP speaking: The development and validation of a rating scale

Shengkai Yin (Shanghai Jiao Tong University, University of Melbourne)

Thursday, July 4, 10:00 to 10:30 am Location: UR3

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

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

A collaborative approach to examining BESTEP's impact on tertiary EAP in Taiwan

Jessica R. W. Wu (The Language Training & Testing Center), Heng-Tsung Danny Huang (National Taiwan University), Shao-Ting Alan Hung (National Taiwan University of Science and Technology), Anita Chun-Wen Lin (The Language Training & Testing Center), Joyce Shao Chin (The Language Training & Testing Center), Ali Shuhsuan Ke (The Language Training & Testing Center)

Thursday, July 4, 04:00 to 04:30 pm Location: HS1

As part of the Bilingual Education for College Students (BEST) Program, the Ministry of Education in Taiwan has supported the development of the BEST Test of English Proficiency (BESTEP). Aligned with the Common European Framework of Reference for Languages (CEFR) A2-B2 levels, BESTEP assesses English for Academic Purposes (EAP) abilities required at the tertiary level. This paper presents an ongoing validation study conducted jointly by BESTEP developers and two universities in Northern Taiwan. The study involves six EAP course designers and instructors and 200 students. Aiming to explore the connection between the BESTEP speaking and writing tests and EAP courses over two semesters, the study begins by analyzing and comparing course descriptions and test specifications, supplemented with classroom observations and teacher interviews. It then tracks changes in students' abilities using BESTEP as pre- and post-tests, alongside instructor evaluations from classroom assessments. The research highlights the vital integration of learning, teaching, and assessment, offering insights into means of improving language assessment practices and informing education system reforms. Additionally, it contributes to the conference theme of reforming language assessment systems by advocating for an effective approach to establishing a comprehensive learning system at the tertiary level in Taiwan.

Aligning Proficiency Level Descriptors with Audiences and Uses: Enhancing Equitable Communication in a K-12 Language Assessment System

Lynn Shafer Willner, Margo Gottlieb (University of Wisconsin-Madison)

Thursday, July 4, 04:00 to 04:30 pm Location: HS2

Reforming communication within a language assessment system necessitates a systematic focus on the information in proficiency descriptors. Our study examined the use of descriptors to interpret results from a redesigned K-12 language test (anchored in the 2020 WIDA language development framework). This study gauged the usability of Proficiency Level Descriptors and their aligned derivatives for three key audiences: language specialists, collaborating subject area teachers, and families of multilingual learners, who may require translator assistance.

Employing mixed methods, we compared the descriptors across various language dimensions, technical versus plain language, and levels of detail to determine their suitability for different users. Both language educators and subject area educators struggled with the shift from a conventional, linear view of language development to a functional one that is contextualized and contingent on genre, purpose, audience, and topic. Additionally, to support communication with subject area educators and families, not only should Plain Language be applied to technical terms (e.g., cohesive devices), descriptor information should be reframed and displayed using different

layouts. Thus, theoretical framing, as well as technical writing and information use influence descriptor design.

Advancements in digital technology allow for more interactive, classroom-based language assessments, necessitating enhanced practices for communicating about the assessment of student language performance, as highlighted by Harding (2021). By adjusting the theoretical framing, technicality, level of detail, and focus of a set of aligned descriptor tools within an assessment system, developers can facilitate equitable communication to diverse stakeholders about the performance of multilingual learners.

Analyzing the Variances in Two Test Administration Modes: Time for a change in the assessment paradigm?

Linda Nepivodova, Simona Kalova (Masaryk University)

Thursday, July 4, 04:00 to 04:30 pm Location: HS3

This paper investigates the impact of computer-based language achievement tests in supervised and unsupervised settings, particularly focusing on university student scores in both on-site and off-site conditions. Conducted at one of the Czech Republic's largest English departments, where computer-based language testing has been integrated since 2003, the study addresses the challenges and changes prompted by the shift to unsupervised testing during the COVID-19 pandemic and their consequences.

Despite smooth implementation and a university policy favouring trust over strict security measures during challenging times, a noticeable score discrepancy emerged between supervised on-site and unsupervised off-site tests, raising concerns about academic integrity.

In a mixed-method study, student scores were analysed across supervised and unsupervised settings, accompanied by a questionnaire surveying student preferences and perceptions regarding the two modes. While students generally favoured computer-based tests, stress levels decreased further in unsupervised conditions. However, due to significant differences in improvement, the paper advises caution in employing unsupervised modes in high-stakes testing, advocating for some level of monitoring to ensure result reliability. The paper highlights the heightened risk to academic integrity in tests administered in unsupervised settings, with students admitting to dishonest behaviour, motivated by the need to check answers or fear of failure. These results indicate that testing in unsupervised conditions needs to be carefully considered and measures must be taken to prevent cheating.

Exploring a new method for multi-lingual alignment of language frameworks: Developing a Global Scale for Multiple Languages Using Comparative Judgement

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

Thursday, July 4, 04:00 to 04:30 pm Location: UR3

This study presents an intriguing exploration into the applicability of the Comparative Judgment method for the alignment of multi-lingual Can-Do statements. To answer the question: To what extent can a set of Can-do descriptors provide a framework to describe language learning and levels of language proficiency across multiple Languages, 320 Learning Objectives (LOs) were translated into Spanish.A panel of 20 qualified raters conducted 25 CJ comparisons per LO in both English and Spanish, resulting in 16,000 data points. A subsequent CJ study was conducted to consolidate the outcome obtained, using the same set of 320 LOs in German. A panel of 20 qualified raters conducted 25 CJ comparisons per LO, resulting in 8,000 data points.

A series of analyses, including rater fit statistics and LO item fit statistics, were performed to gauge the Learning Objective difficulties. Original LO difficulty as indicated by Rasch logits were compared with the CJ estimates from both English and Spanish versions across four language skills. Transformation equations were derived from these comparisons to align the outcomes of Spanish LOs with the existing English Scale, leading to the creation of a new Global Scale of Languages (GSL) for Spanish. Further analysis on the German data alone, then the combined Spanish and German data were performed to validate this alignment.

Using the innovative CJ approach, this study provides empirical evidence supporting the view that the CEFR itself is a language-neutral framework which “can be adapted and used for multiple contexts and applied for all languages” (CoE, 2001).

Where the Lines are Drawn: A Survey of English Proficiency Test Use in Admissions among U.S. Research-Intensive Universities

Nicholas Coney, Daniel R. Isbell (University of Hawai‘i at Mānoa)

Thursday, July 4, 04:30 to 05:00 pm Location: HS1

This study examines the English Language Proficiency (ELP) requirements for international students seeking admission to Carnegie R1 universities in the United States. These institutions are research-intensive and represent a significant proportion of international student enrollment in the country.

Between September 2022 and January 2023, we collected data on ELP test usage from 146 Carnegie R1 university websites. In total, there were 32 different ELP tests being used for admissions, with the most common tests being TOEFL iBT, IELTS Academic, Duolingo English Test (DET), and Pearson Test of English Academic (in that order). In many cases, the cut scores for less commonly used tests were aligned with TOEFL iBT scores, although DET often had lower requirements. Subscore requirements were relatively rare; when implemented writing requirements were most common.

For undergraduate admissions, institutions with higher TOEFL iBT cut scores tended to be more prestigious, selective, and had larger proportions of enrolled international students, as well as

higher 6-year graduation rates for international students. Interestingly, a small handful of prestigious institutions, including Harvard, Stanford, and the University of Chicago, did not require ELP scores for undergraduate admissions. Graduate admissions were more conservative, with fewer ELP tests accepted, higher score requirements, and more common subscore requirements. Conditional admissions were more prevalent for graduate students. Institutional prestige correlated with TOEFL iBT cut scores and international graduate student enrollment. Study findings provide important context for evaluating test use in real-world higher education settings.

Investigating score reporting systems and practices: Content and genre analyses of parent versions of standardized language test score reports

Monique Yoder (Michigan State University)

Thursday, July 4, 04:30 to 05:00 pm Location: HS2

Educational test developers use score reports to communicate test score information to test user stakeholders (e.g., school administrators, educators, parents). As such, these reports play a vital role in the valid use of test scores (Lu et al., 2021). In the context of standardized K–12 English language proficiency tests in the U.S., there are professional guidelines (AERA, 2014), industry standards (Zenisky & Hambleton, 2012), and national education policies (ESSA, 2015) that guide and constrain the content of score reports and how schoolchildren's scores are reported to parents. To what extent and how these score reporting conventions are manifested in parent versions of standardized language test score reports has yet to be explored in educational measurement and young learner language assessment. In this presentation, I summarize how I inventoried the type of information and how information is presented on parent versions of young learners' score reports through iterative document analysis (Donaldson et al., 2021) and genre analysis of state-level English Language Arts (ELA) standards exams from all 50 U.S. States, Washington, DC, and the Virgin Islands. I report on the most common types of information that states include in parent versions of Grades 3–8 test score reports (e.g., relating a child's scores to performance standards). I also showcase the different rhetorical moves and layout techniques states use to present score and performance information to parents. I discuss the implications my findings have on high-stakes test score reporting systems and parent version score report design for young learner assessments.

Assessment method reform: Examining the comparability of linguistic features of communication elicited in virtual and physical settings

Slobodanka Dimova (University of Copenhagen)

Thursday, July 4, 04:30 to 05:00 pm Location: HS3

Locally-developed tests used for oral English assessment of English-medium instruction (EMI) lecturers' language skills for teaching purposes are often based on physical classroom observations, simulated lectures, or locally-contextualized speaking tasks. Rapid changes of teaching practices that are increasingly embedded in virtual and multimodal environments

require assessment method reforms using digital technologies. However, concerns are raised regarding the comparability of EMI lecturers' communicative behavior in the physical and the virtual classroom, and, therefore, the validity of test results based on virtual test administration. This study examines the comparability of EMI lecturers' linguistic skills and discourse modalities elicited during virtual testing sessions based on a simulated lecture and when teaching EMI courses face-to-face. For that purpose, both test performances and teaching practices of 15 EMI lecturers were video recorded (total of 13hrs.), transcribed, and coded for modality (embodied and disembodied). EMI lecturers' linguistic skills were measured in terms of lexical diversity and syntactic complexity. Incidence of connectives was also examined as a measure of cohesion. Based on the non-parametric Wilcoxon signed-ranks test, no meaningful differences in the linguistic features of EMI lecturers' language use were observed between the two contexts. Differences were found primarily in the presence of embodied modes, while occurrence of disembodied modes were similar in virtual test performance and face-to-face teaching practices. Drawing on the results, we discuss the validity of virtual administration of assessments based on a simulated lecture and the role of multimodality in the comparison of the characteristics of face-to-face classroom communication and virtual test performance.

Using AI to enhance JEDI: multilingual constructs to reform monolingual tests

Graham Seed (Cambridge University Press & Assessment)

Thursday, July 4, 04:30 to 05:00 pm Location: UR3

In the era of the 'multilingual turn' (May, 2014), adoption of constructs which recognise the plurilingual competence of test-takers is one way to enhance principles of JEDI within language assessment. But critics have been quick to point out the lack, or slow speed, of recognition or adoption of plurilingual, code-switching and/or translanguaging constructs and practices within language assessment. As AI starts impacting language testing practice, this paper therefore investigates ways in which testing bodies might recognise a test-taker's plurilingual competence within assessment, by using AI technology.

This paper reports on a study of examples of code-switching behaviour found in test responses (N=540), taken from an automated learning-oriented assessment platform 'Write&Improve'. Responses were coded according to different reasons a test-taker may use non-English words and phrases in an English text, building on categories suggested by Nguyen, Yuan & Seed, 2022. The results provide insights as to how and why test-takers may code-switch. These include reasons such as the desire to include real-life elements from their home culture in their response, as a strategy to maintain communication despite a language breakdown, or to provide a draft of content in L1.

Finally, the paper considers that development of AI-driven digital capabilities will enable personalised, learning-oriented language assessment and feedback, facilitating the display of plurilingual competence. These developments are however in a nascent state, and this paper provides thought as to the future, digital, direction of plurilingual assessments, beginning with the need for proper training data in AI (Caines, Seed & Buttery, forthcoming).

Supporting Higher Education institutions through language assessment reform: Evaluating the impact of change on admissions tests

Tony Clark (Cambridge University Press & Assessment), Emma Bruce (British Council), Karen Ottewell (University of Cambridge)

Thursday, July 4, 05:00 to 05:30 pm Location: HS1

Since language assessment continues to adapt to a context of rapid change, a major challenge for Higher Education Institutions (HEIs) is how to ensure that high-stakes decisions around admissions tests are informed by empirical data. In recent years a broader range of English tests has been used by HEIs as proof of language proficiency. Whilst providing enhanced choice and accessibility, concerns have been raised about the proficiency levels of international students and the ensuing impact on their ability to engage and thrive academically (Wood, 2023). Low language proficiency and lack of English support are often cited as reasons why students may have an inferior academic experience (Russell et.al., 2022). However, the potentially negative impact of the tests themselves should not be overlooked.

This paper reports on a mixed-methods study, investigating different tests at institutions, perceptions of university personnel towards the various language tests used, and transparency around decision-making for test acceptance. The study involved: i) desk-based research on institutions' admissions tests, their required scores and changes in accepted tests (n=50 institutions); ii) survey data (n=300) and interviews (n=20) with university personnel (faculty, recruitment, admissions, EAP). Focus group discussions (n=20) with international students followed.

Preliminary results indicate that processes around test acceptance are not uniform across institutions, and apprehensions around key factors which underpin new forms of online testing were apparent. The presentation concludes by outlining how findings may underpin a framework for evaluating tests used for HEI admissions, in addition to developing assessment literacy for stakeholders (Baker, 2016, Taylor, 2013).

Intersecting Voices: A Sociocultural Exploration of Test-takers' and their Parents' Experiences and Perceptions in English Tests of Young Learners

Jia Guo (Queen's University), Liying Cheng (City University of Macau)

Thursday, July 4, 05:00 to 05:30 pm Location: HS2

English tests for young learners' proficiency have seen a notable increase, significantly influencing English education for children (Cheng, 2008; Bulter, 2015). While many studies have delved into the perspectives of test-takers and parents, the sociocultural dimension remains largely unexplored (Carless & Lam, 2014). This research delves into individual family units to intertwine the experiences of both stakeholders, aiming to understand the sociocultural nuances linked to these tests for young learners, reflecting the intricate testing environment in early language testing (Moss et al., 2006).

The study primarily investigates the consequences of English tests on young learners by capturing voices from both the test-takers and their parents. It centers on three research questions: the

experiences of test-takers, the perceived uses and consequences by parents, and the overall test consequences on individual family units. The research employs Hofstede et al.'s (2004) layered structure of culture as its theoretical base, incorporating the pyramid and onion models to contextualize test use and consequences. A qualitative methodology was adopted, involving 15 family units. Young test-takers visualized their testing experiences through comic strip-style drawings and underwent subsequent interviews to discuss their experiences in depth. Parents, on the other hand, participated in semi-structured interviews about their perceptions and perceived test outcomes. Thematic analysis from a social constructivist viewpoint (Creswell & Creswell, 2018) illuminated the sociocultural norms influencing these testing dynamics.

Preliminary findings spotlighted four themes, underlining the cultural factors associated with test registration, preparation, the test-taking process, and post-test results. Parents' interactions with the tests are molded by cultural symbols and societal anticipations. Test-takers' compliance is heavily influenced by cultural competition and parental aspirations. An outstanding observation was the emotional synchronization within families, emphasizing cultural values like obedience, and gratitude, and the emphasis on academic excellence and familial cohesion.

How does extended time affect dyslexic test-takers with different item types in an online English test?: An exploratory study

Chihiro Inoue, Lynda Taylor (University of Bedfordshire)

Thursday, July 4, 05:00 to 05:30 pm Location: HS3

Extended time is a common accommodation for test-takers with specific learning difficulties in language tests (Kormos & Taylor, 2021). Two recent studies investigated the effect of this accommodation with learners with literacy-related difficulties, and highlighted the need for looking into different types of items/tasks (Kormos and Ratajczak, 2019) and for gathering qualitative data into the usage of extended time (Pastorino et al, 2023).

Four English learners with a confirmed diagnosis of dyslexia participated in this study. They completed an online English test with operational and extended timing (up to 50% more time per item/task). The test covered eight item/task types assessing various language/literacy skills, including C-tests, lexical decision tasks (textual and aural), dictation, reading, and opinion-giving tasks. Data included overall test scores, subscores, raw scores per item/task type, time spent per item/task, and perceived usefulness of extended time per item/task and reasons behind it.

Results showed variation among the four participants. Two scored higher in the time-extended test, particularly in reading and writing. However, the other two did not; one showed no score change, while the other's scores declined due to fatigue, especially in items/tasks related to listening and speaking tasks. Participants did not always use the extended time, and perceived usefulness varied according to item/task types, their characteristics of dyslexia, and previous training in English. This study highlights the complex and individualised nature of how time-extension accommodation help (or not help) dyslexic learners of English.

The role of policy actors' agency in test impact: Assessment of languages other than English in China's senior secondary education

Chenyang Zhang (University of Melbourne)

Thursday, July 4, 05:00 to 05:30 pm Location: UR3

Following the promulgation of the Belt & Road initiative in 2013, increasing attention has been devoted to supporting languages other than English (LOTE, i.e., foreign languages other than English) education in China. Amid this trend, the national assessment policy encourages senior secondary students to select a LOTE subject instead of English in the National College Entrance Examination (i.e., Gaokao). This policy shift, however, has captured scant attention from the language assessment community. Drawing on the dialogical approach to agency, this study aims to develop a test impact model to account for the role of policy actors' agency in generating the LOTE test impact. The methodology of this study adopted the constructivist grounded theory. In the context of senior secondary education in Shanghai, China, data were collected from multiple one-on-one semi-structured interviews with 1 Shanghai Municipal Educational Examinations Authority officer, 2 school administrators, 4 LOTE teachers and 27 senior secondary students. Data analysis involved initial coding, focused coding, and categorising, to inductively generate theories. Findings showed that policy actors' agency – their felt experiences expressed by voices – played a key role in generating test impact in the three interrelated themes – (a) the political context (the context within which policy actors continuously participate in dialogue with others and their social settings), (b) the ontogenetic responses (individuals' inner evaluations and expectations), and (c) concrete actions. This test impact model contributes to language testing theories since the current theorisation of test impact does not pay sufficient attention to agency of policy actors at multiple levels.

An investigation of the alignment of national language teaching policy with the advanced-level secondary school leaving examination in foreign languages in Hungary

Katalin Piniel, Gyula Tankó, Zsuzsanna Andréka (Eötvös Loránd University, Budapest, Hungary)

Thursday, July 4, 05:30 to 06:00 pm Location: HS1

Since the reform in high school final examinations in Hungary in 2005, the Foreign Language Matura exam has undergone many changes. While research on the school leaving examination itself is rather scarce, there seems to be even less work available on the coherence of the education system as far as the link between policies regulating foreign language learning in secondary school and the school leaving examination is concerned. This would also be important, as curriculum-assessment-instruction models (Bunch, 2012), including that of Biggs and Tang's (2011) idea of constructive alignment suggest that intended learning objectives, the assessment task, and learning activities should be seen in unity.

Thus, the research question guiding the study was the following: How do intended language learning outcomes as included in the national core curriculum and other regulatory documents compare to the goals and purposes of assessment stated in the Hungarian advanced level Matura examination in foreign languages test specifications? To this end, we conducted an exploratory

study, using document analysis (Bowen, 2009) on a set of publicly available regulatory. For data analysis, we used Atlas.ti 9 to establish categories relevant to our aims and code the texts. Our findings show that there has been a shift on emphasis in the regulatory documents regarding the detail about the aims of the examination itself. In terms of intended learning outcomes, those spelled out in regulations overlap with but are not identical to the aims of the assessment.

Shedding Light on the Test-Taking Experiences of Francophone African Learners of English in High-Stakes English Proficiency Testing

Kadidja Koné (Ecole Normale d'Enseignement Technique et Professionnel (ENETP), University of Letters and Humanities Bamako, Mali), Paula Winke (Michigan State University)

Thursday, July 4, 05:30 to 06:00 pm Location: HS2

Africa has the fastest growing and youngest population of any continent. The New York Times (2023, October) estimated that by 2050, one in four people will be African. To quote the Times, "The world is becoming more African" (p. 4), and thus, we believe, standardized, international language tests must become more African too. Toward this end, we investigated 63 francophone West African test takers' beliefs and reflections after taking a standardized, international English test in Africa. We asked: (1) What do West African, francophone learners of English report as their perceptions of a standardized, international English test? (2) What can test designers and English teachers do so that the African test takers can have positive test taking experiences? The 63 graduate student participants learning English included teacher trainees and pre-service English teachers. Each was given a coupon to take the Duolingo English Test (DET). After taking the test, each student partook in a focus group session, with audio and transcripts uploaded to MAXQDA for analysis. Emergent themes suggest: (1) significant electrical grid instability and technology limitations affected West African test-takers' experiences; however, (2) the test takers indicated that they enjoyed participating in international English-language testing, which most indicated was a novel experience for them. They discussed that their purposes for assessment may include more diagnostic and formative purposes than university-entrance or immigration/study-abroad purposes. They discussed West African culture and how it interplayed with their exam experiences. What the themes may mean for international and global English-language testing is discussed.

A literature review on the ordering of test components

Ramsey Lee Cardwell, Ben Naismith (Duolingo)

Thursday, July 4, 05:30 to 06:00 pm Location: HS3

In designing language tests, the sequence of items and sections can impact scores, test-taker experiences, and psychometric properties, but current practices seem to lack empirical support. Trends of at-home and adaptive testing creates an opportunity to reevaluate these practices.

We conducted a systematic literature review on test item and section ordering, searching PsycINFO and ERIC databases for relevant studies, and identified 64 studies published since 1942 with varied contexts, test constructs, and test-taker demographics. Studies were reviewed for

independent variables/experimental manipulation, dependent/outcome variables, mediating variables, and findings/conclusions of statistical significance.

The most studied topics were (1) how difficulty-based item ordering (e.g., easy-to-hard) affects scores and perceived performance, and (2) anxiety/affect as both a mediating and outcome variable. Findings show a slight positive effect of easy-to-hard ordering on scores and an interaction between anxiety levels and difficulty-based ordering. However, there's a lack of research on section ordering, adaptive tests, and language proficiency tests.

Our findings support a few test design best practices. Starting with easy items can modestly improve test-taker performance and experience, though the underlying reasons are unclear. Test item/section ordering also appears to affect individuals differently based on characteristics like proficiency and anxiety. Ending tests with difficult items could leave test-takers feeling negatively.

More than conclusive answers, our literature review revealed many unanswered questions. We propose topics needing further research, which will facilitate leveraging the full potential of digital testing technology in pursuit of assessments that are accessible, equitable, inclusive, secure, and valid.

Comparing reading item difficulty: Does A1 equal A1?

Katharina Karges (University of Leipzig & University of Fribourg)

Thursday, July 4, 05:30 to 06:00 pm Location: UR3

This study explores the sources of foreign language reading item difficulty in three foreign languages (German, French and English, A1.2/A2.1), based on test results collected in a large-scale assessment in Swiss primary schools (N=19,357). The study operationalises foreign language reading items in terms of their linguistic characteristics (e.g. lexical diversity, difficulty, cognates) and the answering process (e.g. item types, distractor falsifiability, lexical overlap), and puts these in relation to their empirical item difficulty. Results highlight the nuanced sources of item difficulty, contributing to a deeper understanding of item difficulty in foreign language reading assessments. The analysis also underscores the influence of individual linguistic repertoires on assessing foreign language competences. This research addresses a gap in the literature, especially for languages other than English, offering insights that could significantly impact the design and validation of reading assessments.

Paper and Demo Summaries – Friday, July 5

Human-Centered AI for Test Development

Alina A. von Davier, Andrew Runge, Yigal Attali, Yena Park, Geoffrey T. LaFlair, Jacqueline Church (Duolingo)

Friday, July 5, 08:30 to 09:00 am Location: HS1

In this presentation we will be presenting a human-centered AI framework that supports and regulates the application of AI for content generation. Human-Centered AI recognizes the importance of involving educators, experts, and students in the development process, ensuring that AI systems enhance, rather than replace, the role of teachers and developers. It also places a strong emphasis on transparency, fairness, and ethical considerations in educational AI.

We will be also proposing a new framework for managing the item development stage for a high volume, high-stakes test which we refer to as an item factory, a system where human expertise and AI prowess are combined for an efficient and high-quality test development process. This system is based on the concept of intelligent automation used in manufacturing. We integrate item and test design and subsequent piloting activities into this framework and discuss how these activities could be made more efficient. The item factory framework relies on automation, AI, and engineering principles to make the item development process scalable for high-volume, high-stakes testing in the digital era.

Speaking of reform: introducing large-scale speaking assessment into a lower-secondary school system

Johanna Motteram (British Council), Jamie Dunlea (British Council), Barry O'Sullivan (British Council), Fumiyo Nakatsuhara (Centre of Research in English Language Learning & Assessment, University of Bedfordshire), Akihiro Matsuura (British Council), Robin Skipsey (British Council)

Friday, July 5, 08:30 to 09:00 am Location: HS2

This presentation reports on a major reform initiative to introduce speaking tests for all junior high school students in the Tokyo Metropolitan area. While revisions of the Courses of Study curriculum guidelines by the Ministry for Education, Culture, Sports, Science and Technology (MEXT) have emphasised the teaching and learning of both productive and receptive skills, the reforms have not delivered the expected outcomes. Central to the debate has been the role of assessment, with the lack of large-scale, effective speaking assessment often cited as an impediment to increasing the focus on the teaching of speaking in the classroom.

In response to these challenges, the Tokyo Metropolitan Government introduced the English Speaking Achievement Test for Japanese Junior High School Students (ESAT-J). Tests for 3rd year students were introduced operationally in 2022. In 2023, the British Council was invited to design and deliver formative speaking tests for years 1 and 2, while continuing the summative test for year 3 from 2024 onwards (targeting approximately 240,000 students).

This presentation outlines the solution advocated by the British Council, focusing on the tests for years 1 and 2. The approach is grounded in the integrated arguments approach (Chalhoub-Deville

& O'Sullivan, 2020). We will present results from a series of data collection steps, from trialling to operational delivery for 160,000 students in January to March 2024. These are integrated into a discussion of the Theory of Action and the Communication model and how these drive our interactions with the key stakeholders crucial to this reform's success.

Reforming sign language assessment: setting up a longitudinal learner corpus of rated elicited imitation performances to develop an AI-driven sign language assessment system

Franz Holzknecht (University of Teacher Education in Special Needs Zurich), Tobias Haug (University of Teacher Education in Special Needs Zurich), Alessia Battisti (University of Zurich), Katja Tissi (University of Teacher Education in Special Needs Zurich), Sandra Sidler-Miserez (University of Teacher Education in Special Needs Zurich), Sarah Ebling (University of Zurich)

Friday, July 5, 08:30 to 09:00 am Location: HS3

Sign languages are the preferred means of communication among deaf people and they bridge the communication gap between the deaf and hearing communities. Despite an increasing number of studies in sign language education, there is a lack of research on sign language assessment, particularly on the role of artificial intelligence in the automated assessment of sign language production.

This paper reports on the collection and analysis of a longitudinal learner corpus as the basis of the automated assessment system. We administered 12 elicited imitation (EI) tasks to 14 L2 learners of DSGS across four data collection points over two and a half years, as well as feedback questionnaires on the learners' perceptions. The 672 videos were scored by eight trained deaf L1 raters on six different rating criteria. The rater data were collected in a randomised overlapping design, wherein 56 videos were scored by all raters and the remaining videos were scored by two raters each. The data were analysed using MFRM.

We discuss detected differences in rater severity and rater reliability and outline how these differences are accounted for in the automated system. The paper focusses specifically on the ratings of non-manual components, as the current system will be the first to automatically score non-manuals in sign language performances. Finally, we report on the extent to which the learners' performance improved over the two and a half years and how their perceptions of the EI tasks changed, to draw conclusions with regards to improved sign language learning and teaching.

Humans vs. LLMs: How good are LLMs in generating input texts for reading tasks on B2/C1 levels of the CEFR?

Anastasia Drackert (Gesellschaft für Akademische Studienvorbereitung und Testentwicklung e. V. / TestDaF-Institut), Andrea Horbach (Universität Hildesheim; CATALPA, FernUniversität in Hagen), Anja Peters (Gesellschaft für Akademische Studienvorbereitung und Testentwicklung e. V. / TestDaF-Institut)

Friday, July 5, 09:00 to 09:30 am Location: HS1

One of the most challenging and time-consuming aspects of developing test tasks for reading comprehension is finding appropriate input texts. At the same time, linguistic features of input texts are an important factor in determining task difficulty and should therefore be comparable across exam versions. Using large language models (LLMs), such as ChatGPT, for generating input texts seems to be a promising yet so far understudied way to develop test materials.

To study the potential of generative artificial intelligence for generating input texts, we investigated the comparability of texts used in a standardized high-stakes B2/C1 German exam and those generated by an LLM. To this end, we used a sample of authentic input texts from a reading task and generated the same number of texts on the same topics and genre using ChatGPT. All texts were analyzed according to a variety of linguistic features.

Results showed some statistically significant differences between the two text types. For example, texts written by humans tend to have higher lexical variation and higher TTR, while AI-generated texts seem to have higher readability indices. Other measures, e.g. word length or frequency of passive sentences, showed no significant differences. A further analysis by assessment experts revealed that ChatGPT texts tend to include fewer names and numbers, are more structured and redundant.

We conclude the presentation by discussing the consequences of the usage of ChatGPT for input text generation, thus contributing to the discussion of the role of generative AI in assessing core language skills and assessment development.

Nuanced approach to the English Language Examination Reform in Japan

Noriko Iwashita (University of Queensland), Megan Yucel (Australian Council for Educational Research)

Friday, July 5, 09:00 to 09:30 am Location: HS2

The study investigates stakeholder perceptions of the examination reform, focusing on data from university lecturers, senior and junior secondary teachers, and students. The study employed a questionnaire survey (N=96) and semi-structured interviews (N=21). The survey results are mixed but with more positive than negative responses.

The semi-structured interviews asked questions about the current and proposed changes in the examination system concerning the assessment of communication skills. The analysis revealed a perceived gap between four-skill-based tests like IELTS and Japan's current English language policy. Respondents expressed that excelling in commercial four-skill-based tests requires additional skills not covered in the existing English curriculum. Many believe that the assessed

skills in commercial tests are either unnecessary or beyond the needs of all Japanese students according to current educational policies.

These findings challenge the government's intended washback effect on the curriculum, suggesting that tests may inform policy development but may have a different impact. Drawing upon Shohamy's claim (2007) that language tests serve as de facto language policy, this study argues that the examination reform, including the introduction of commercial tests, necessitates consideration beyond just equity concerns. Instead, the focus should shift towards curriculum reform aimed at enhancing communication skills, taking into account the prevailing school environment and the specific needs of students. The study provides valuable insights into stakeholders' perceptions of examination reform in Japan and highlights the complexities of aligning language testing with educational policies. It sheds light on the intricate challenges of harmonising language testing with educational policies.

Automatic CEFR classification of written learner texts using Natural Language Processing

Torsten Zesch, Jeanette Bewersdorff, Josef Ruppenhofer (FernUniversität in Hagen)

Friday, July 5, 09:00 to 09:30 am Location: HS3

CEFR levels are a central analytic variable in language testing research. While automatic systems analyzing language productions based on Natural Language Processing (NLP) have been proposed, they are still not widely used. Existing NLP systems are usually trained based on CEFR datasets annotated either on document or sentence level. Early experiments with large language models (like GPT) indicate that they can also be used when appropriately prompted.

Given all this research there are relatively few readily usable classifiers. In this presentation, we give an overview of the research landscape within automated CEFR classification of written learner texts using NLP methods. We re-implement and compare different approaches according to classification performance, interpretability and practicability. We then zoom in and discuss the specific issues arising from automatically dealing with learner language. For example, orthography is in itself an important predictor in many automated systems, but spelling errors might also interfere with the identification of lexical items. Here, we present experiments on separating features into either working on the learner language or on a target hypothesis. Another practical issue, especially in source-based writing, is re-use of material. Here, we report results on the impact of identifying (and later excluding from analysis) material that was taken from a specific source text.

Our presentation takes a nuanced view on the use of AI (and NLP in particular) in language testing, outlining the potentials but also carefully examining the conditions and requirements necessary for a successful integration into existing language testing workflows.

An Online Diagnostic Assessment System for English Language Teaching and Learning at Schools, Colleges, and Universities

Yan Jin (Shanghai Jiao Tong University), Zunmin Wu (Beijing Normal University), Liping Liu (Foreign Language Teaching and Research Press)

Friday, July 5, 09:00 to 09:30 am Location: UR3

Diagnostic language assessment provides detailed insights into learners' language proficiency, shedding light on their strengths and weaknesses in the sub-skills of major language skills. Therefore, the development of diagnostic assessment emphasizes the importance of accurate and timely diagnostic feedback, as well as its utility for teaching and learning. This presentation demonstrates how the UDig Diagnostic Assessment (UDig) system addresses the needs of English language teaching and learning at schools, colleges, and universities in Chinese educational settings. Developed by the Foreign Language Teaching and Research Press (FLTRP) in China, UDig offers a comprehensive suite of online diagnostic tests, covering reading, listening, speaking, writing, grammar, and vocabulary. Currently, it serves over 465,000 students across 220 universities and 1000 K-12 schools. In this demonstration, we first showcase UDig's structure and functions, highlighting its three core features: accuracy, timeliness, and comprehensiveness. We will then take a closer look at sample feedback reports for teachers and students, illustrating how they monitor and track student progress and guide remedial instruction effectively. We will end the demo with an overview of research programs funded by the publisher to validate the assessment system and promote learning-oriented assessment practices and feedback-based remedial instruction (Fan, Song, & Guan, 2021; Jin & Yu, 2023). UDig represents a significant advancement in diagnostic assessment, offering educators and learners valuable tools for enhancing English language teaching and learning in China. Its accurate, timely, and comprehensive approach ensures that diagnostic feedback is not only informative but also actionable, supporting more effective teaching and learning strategies.

Can GPT write good items? Comparing item characteristics of human-written and GPT-4-written items

Yena Park, Jacqueline Church, Yigal Attali (Duolingo)

Friday, July 5, 09:30 to 10:00 am Location: HS1

Writing good items is difficult. An illustrative example is 43 rules in Haladyna and Downing's (1989) multiple-choice item writing guidelines. Item writing is not only cognitively taxing, potentially impacting the quality of the items produced, but also challenging for testing programs operating within the constraints of limited resources in need of a large number of items.

The resource-intensive nature of item writing may be alleviated by leveraging large language models (LLMs) (e.g., OpenAI, 2023). Not only can LLMs generate items that adhere strictly to all and any rules, LLMs have also generated construct-relevant items for language tests (Attali et al., 2021; Zu et al., 2023). Questions remain, however, as to the psychometric quality of automatically generated items and its comparability to that of items written by human experts.

To answer this research question, we collected test taker responses on expert-written and automatically generated sets of items on a test readiness platform of a standardized language

proficiency test. Classical item analysis showed that GPT-4 generated items had comparable item difficulty and item discrimination indices to expert-written items, except for cloze-type items. A higher proportion of expert-written items met the threshold of acceptable item quality, but conducting distractor analysis helped decrease the gap.

These results lend support to using large language models within the test development process to create test items. Large language models are capable of producing items of comparable psychometric quality that facilitates test development at scale, while staying faithful to the item specification and item writing guidelines.

Developing an evaluation framework for proficiency testing for education and employment in Taiwan

Richard Spiby, Emma Bruce (British Council)

Friday, July 5, 09:30 to 10:00 am Location: HS2

Test users are expected to bear certain responsibilities when selecting language tests and subsequently interpreting scores derived from those tests (AERA, 2014). However, in reality, the expertise of score users to evaluate information made available by test providers can vary widely, as can the information itself, negatively affecting subsequent decisions.

Since the early 2000s, the Taiwan Ministry of Education has set recommended proficiency levels, as recently updated in the 2030 Bilingual Programme (NDC, 2020). While no specific tests have been authorised, an implicit range of tests has become established among test-score users.

The aim of the present research is to provide an evidence-based framework for evaluating tests for educational institutions, government departments and public/private employers in Taiwan. Stage one surveyed current practice in the selection of tests used by universities, the civil service and employers. Stage two was a pilot project to establish a framework to evaluate proficiency tests used for different purposes in the Taiwanese context. Key considerations are test design, technical properties, CEFR alignment, and information made available for users. The framework is constructed based on the principles of the socio-cognitive model (Chalhoub-Deville & O'Sullivan, 2020) with reference to frameworks for test comparison and recognition.

This presentation reports on findings for the surveyed universities, civil service departments and financial services sector, and on the evaluation framework, which was piloted with a sample of established tests. Results will be presented and the challenges of creating evaluation frameworks applicable for different purposes for use by diverse stakeholders will be discussed.

Fairness of TCF Writing using human raters and a hybrid automated rating model: from construct validity to psychometrics, to an argument-based approach

Vincent Folny (France Education International), Rodrigo Souza Wilkens (Université de Louvain / UC Louvain), Rémi Cardon (Université de Louvain / UC Louvain), Thomas François (Université de Louvain / UC Louvain)

Friday, July 5, 09:30 to 10:00 am Location: HS3

The Standards for Educational and Psychological Testing (2014) state that the fairness of a test is defined not only in terms of psychometrics and bias, but also in terms of accessibility, intended use or construct validity. The fundamental difference between bias studies and fairness is not always shared outside the field of education measurement (Klebanov & Madnani, 2022, Tay et al., 2022).

For the TCF writing a solution called FIDELIA was developed in collaboration with UC Louvain to automate one of the two human ratings of the TCF. Prior to the project, a literature review revealed the lack of state of the art for bias studies and fairness for French. The UC Louvain worked on a model that promotes explainable artificial intelligence. A set of 48 linguistic features was created. Using this feature extraction, an evidence synthesis procedure (Bejar,2011) and the deep learning CamemBERT model (Martin et al., ACL 2020), UC Louvain created 24 different hybrid models. France éducation international (FEI) created a gold standard corpus of 961 tasks scored by 55 raters and calibrated with a Many Facet Rasch model (Folny, 2023).

A study was carried out using mainly the bias analysis procedure proposed in the Facets software (Linacre, 2023). The conclusions of this study were used for the TCF model card (O'Sullivan, B., Breakspear, T & Bayliss, W., 2023). Finally, FEI has decided to adopt a vision of fairness that is not set in stone but is dynamic, linked to validity and the context of use (Lane, 2023). This is what we might call an ongoing retrofit of fairness.

Probing attribute structures in testlet-based listening assessment: An application of cognitive diagnostic models

Lidi Xiong, Lianzhen He (Zhejiang University)

Friday, July 5, 09:30 to 10:00 am Location: UR3

This paper explored the structural relationship of listening attributes in the testlet-based diagnostic listening assessment. Specifically, six cognitive diagnostic models (CDMs), the independence model, the higher-order model, the hierarchical model, and their testlet counterparts, were applied to a testlet-based listening dataset and then rigorously examined for their model-data fit, parameter estimates and diagnostic results. Test-takers in this study included 973 Chinese EFL learners at tertiary level who were assessed by the listening section of a diagnostic test provided by the UDig platform. flexMIRT 3.51 (Cai, 2017) was utilized to fit the six CDMs. R was employed for further evaluation of model performance and item analysis. Regarding attribute structures, the results showed that the higher-order models consistently outperformed others in terms of goodness-of-fit and classification accuracy, regardless of testlet effects considered or not. Regarding the influence of the testlet effects, all testlet CDMs exhibited better model-data fit than their non-testlet counterparts, albeit with a slight trade-off in

classification accuracy. A pronounced alleviation of local item dependence was also observed when using the testlet CDMs. However, all models, except for the two independence models, demonstrated high classification consistency at the pattern and attribute levels. Additionally, the parameter estimates of the six models exhibited moderate to strong correlations on average. These findings suggest that higher-order models effectively capture the underlying structure of listening, yet in practice, CDMs, apart from the independence models, demonstrate robustness even in the presence of theoretical and data misspecifications.

Cloning Tasks with GPT Models for Automated Difficulty Estimation

Sylwia Macinska (Cambridge University Press & Assessment), Andrew Mullooly (Cambridge University Press & Assessment), Luca Benedetto (ALTA Institute & Computer Laboratory, University of Cambridge), Hannah Bouteba (Cambridge University Press & Assessment), Mark Elliott (Cambridge University Press & Assessment)

Friday, July 5, 10:00 to 10:30 am Location: HS1

Content calibration is significant bottleneck in educational content development, relying on expert judgment and pre-testing, which are resource-intensive. Recent computational models, notably Large Language Models (LLMs), have shown promise in predicting item difficulty directly from text (Benedetto et al., 2023). However, they typically require additional pre-training.

This study investigates the use of the latest GPT models through a zero-shot learning approach. However, instead of directly estimating item difficulty from text, this study proposes a cloning method for generating new items by modifying key features of validated tasks. The objective is for these new items to either mirror the originals' difficulty or to follow a predictable performance pattern.

We applied this method to a subset of an Open-Cloze dataset (Felice et al., 2022), focusing on B1 and B2 CEFR levels. The method involved manipulating variables such as task topic, item order, and contextual words around gaps in the original content. This resultant pool of tasks was then administered to a sample of learners.

Our research addresses the following core aspects: (i) tailoring prompts for quality task cloning, (ii) comparing cloned and original tasks in readability, linguistic features, and similarity metrics, (iii) and finally, predicting learner performance on cloned tasks using original task data.

Our presentation will cover the prompting strategies used, showcase cloned task examples, and provide a comparative analysis between original and AI-generated tasks. We will discuss the predictability of performance on cloned tasks and explore the implications of our method for construct coverage and task predictability.

Revising the ILTA Code of Ethics, and the impact of ethical consensus in the global language testing community

Bart Deygers (Ghent University), Margaret Malone (ACTFL)

Friday, July 5, 10:00 to 10:30 am Location: HS2

The Code of Ethics (COE) by the International Language Testing Association (ILTA) is the benchmark of ethical testing practice and deontology that all language testing associations subscribe to. It was ratified by the ILTA community in 2001 and twenty years later ILTA has started the process of revising its COE. In this presentation we will discuss findings stemming from the research-driven revision process and answer two research questions. The first RQ investigates to what extent the academic and privatized language testing communities consider ethical testing practices to be a priority? The results of a survey administered to the global community of language testing researchers and practitioners (n = 284) show that ethical considerations are important for language testers, regardless of their affiliation, but not regardless of culture: not all regions of the world consider all ethical principles listed in the current COE equally important. Based on these findings we formulated a second research question: To what extent is it possible to come to global agreement on central ethical principles for language testing? To answer this question we organized ten cross-cultural focus groups. Relying on Nagel's principles of multiperspectivity each focus group included participants from diverse regional, cultural and professional backgrounds and discussed two ethical principles. The results of the thematic analysis of these focus groups show the difficulty of ethical universality in language testing practices. In the discussion we will identify the challenges ahead and propose ways to deal with them in a revised and effective code of ethics.

Use of a technology-assisted rating tool for assessing integrated English academic writing ability

Haeyun Jin (Korea National Open University, Korea, Republic of South Korea)

Friday, July 5, 10:00 to 10:30 am Location: HS3

Reading-to-write integrated tasks have been adopted widely in many L2 English writing assessment contexts as tools for assessing L2 proficiency (Knoch & Sitajalabhorn, 2013). Despite the augmented authenticity of these tasks, numerous challenges have been reported pertaining to the rating of source use: raters struggled with distinguishing between test takers' own language and language cited from the source texts (e.g., Weigle & Parker, 2012). Less is known about how the affordances of technologies can be employed to address the complexity of rating reading-to-write tasks. In response to such a gap in research, this study investigated how a computer-based rating tool, called the Scaffolded Rating Tool (SRT), could help raters produce more accurate and generalizable scores and reduce raters' perceived complexity of rating reading-to-write tasks. Six trained raters rated summary essays written by 90 international students in two conditions: SRT and non-SRT conditions. The test takers' performance was rated based on an analytic rubric composed of five criteria: viewpoint recognition, text engagement, organization, development, and language use. The Many-facet Rasch Measurement analysis indicated that rating accuracy and consistency were much improved in the SRT rating. A G-study analysis further showed that the SRT rating resulted in more dependable scores, especially for

the two source use criteria. Lastly, all raters found the SRT rating cognitively less demanding and more efficient and were overall more confident in their decisions. The implications of this study will be discussed as well as some recommendations for developing a technology-based rating tool for integrated writing tasks.

Integrated Diagnostic Grammar Assessment: A Systemic Functional Linguistics Approach

Roz Hirch (Medicine Hat College, Canada)

Friday, July 5, 10:00 to 10:30 am Location: UR3

Recent pedagogical research underscores the importance of diagnostic assessments in language education, identifying areas of learners' language ability that are strengths and pinpointing those areas where the learner needs to work. Yet, considerable research is needed to better understand diagnostic language assessment. This includes understanding the nature of information that should be derived from a diagnostic assessment, and the language theory used. Additionally, the form of the diagnostic feedback given to teachers and students and how it will be used by teachers and students must be considered.

This presentation explores these issues through validating the use of a diagnostic assessment designed for English grammar. The argument-based validation process was guided by two main research questions: What types of diagnostic information are useful for ESL classrooms? And how should this information be presented to users? The grammar test was administered to 84 college students across 8 classes in a first-year ESL course. The test results were then analyzed using a framework developed built on Systemic Functional Linguistics (SFL), which offers a nuanced understanding of grammatical proficiency, incorporating both semantic and syntactic information. These results were then used to develop feedback forms, which were then reviewed by 8 teachers, who responded to questionnaires asking which type of feedback they preferred and how they would use the feedback with their students and classes. These results, including the test, framework, and automated feedback developed for this study will be discussed in this presentation, along with implications for the future of diagnostic language assessment.

"Context-limited" or "boundary-crossing"? The essential contribution of case study research in language assessment

Beverly Baker (University of Ottawa), Lynda Taylor (CRELLA, University of Bedfordshire)

Friday, July 5, 11:00 to 11:30 am Location: Aula

In this conceptual/theoretical presentation, we aim to invite a critical reconsideration of the role and value of case study research in language assessment. In our presentation, we draw on a range of recent examples of high-quality case studies in the language assessment literature, highlighting the insights from this work that we believe could not have been attained through any other approach. Case study research is still sometimes regarded as inferior - in terms of both quality and importance. A common convention in scholarly contributions has been to downplay

the value of such studies on the basis that they are not “generalisable”. We argue, however, that this conception of the term "generalisable" represents a positivist and context-independent orientation at odds with the interpretive nature of the case study research that was undertaken in the first place. In short, we challenge perceived limitations of case studies from an epistemological standpoint—sharing the work of scholars who argue that the context-dependent case study is the most valuable tool for knowledge generation and practical application in the social sciences. In doing so, we hope for a productive discussion on how case studies may be better perceived and reported in our field in the future.

Investigation of Differential Item Functioning Analyses Due to Multiple Manifest Grouping Variables: Rasch Perspective

Sanshiroh Ogawa, Hong Jiao (University of Maryland, College Park)

Friday, July 5, 11:00 to 11:30 am Location: HS1

Measurement invariance is fundamental to test fairness in high-stakes standardized tests. One common way to test measurement invariance is via differential item functioning (DIF) analysis as a method to identify potential biases in test items. Traditional DIF analysis considers one grouping variable at a time, such as gender or ethnicity, but this approach may not capture DIF resulting from the interaction of multiple grouping variables (Jiao & Chen, 2014). Thus, the main aim of the current study is to develop a model that detects DIF caused by two grouping variables and their interaction in the context of language assessment.

To this end, the proposed model extends the Rasch model by incorporating a group specific parameter for each group and a parameter for the interaction effect between the two grouping variables. The performance of the proposed model is first investigated with DIF in simulated data. The research questions focus on comparing the DIF detection results obtained from two different methods: the traditional IRT-based approach and the model proposed in the current study. Primary results show that the proposed model detects DIF caused by multiple grouping variables as well as their interaction simultaneously. The analysis of data from the reading test in the Programme for International Student Assessment 2018 will also be reported.

Communicating ELP Assessment Changes to K-12 Educators

Ahyoung Alicia Kim (WIDA, University of Wisconsin-Madison), Lorena Alarcon (Univ. of Illinois Urbana-Champaign), Jason Kemp (WIDA, University of Wisconsin-Madison), Fabiana MacMillan (WIDA, University of Wisconsin-Madison)

Friday, July 5, 11:00 – 11:30 am Location: HS2

English language proficiency (ELP) assessments are closely tied to English Language Development (ELD) Standards in the kindergarten to 12th grade (K-12) educational setting in the U.S. The standards function as the language construct measured in ELP assessments, and inform the design of test items and tasks (Gottlieb & Chapman, 2022). When changes are made to the standards, assessments need to be updated. These updates need to be effectively communicated to educators who support English learners (ELs). This communication should serve the dual

purpose of sharing changes to assessments, but also enhancing educators' language assessment literacy regarding the changes.

This study examines (1) K-12 EL educators' understanding of the impact of an updated ELD Standards on an ELP assessment and (2) ways to enhance educators' language assessment literacy regarding the changes to the assessment. We conducted a two-phase mixed-methods study. In Phase 1, 1,568 educators completed an online survey on their understanding and needs regarding ELD Standards, ELP assessment, and their connection; data were analyzed using primarily descriptive statistics. Phase 2 involved two focus groups of nine educators to discuss their needs in greater detail and to get feedback on educator resources; data were analyzed using inductive-deductive approach (Creswell, 2014). Results guide the communication of assessment changes to K-12 EL educators and the development of resources that aim to enhance educators' language assessment literacy. This study has implications for the public dissemination of reforms to large-scale language assessments, and it highlights some benefits of increased collaboration among testing agencies and educators.

Exploring the moderating role of assistance in assessing speaking ability for argumentation

Jorge Luis Beltran Zuniga (Teachers College, Columbia University)

Friday, July 5, 11:00 – 11.30 am Location: HS3

The current study explored the assessment of English Language Learners' argumentative speaking ability. Specifically, the study involved the development and administration of a scenario-based test that aimed to measure learners' ability to display competency in L2 argumentative speaking ability by building an argument from factual evidence and presenting it to a simulated audience. In addition, the study aimed to examine whether embedding assistance in the test can help overcome gaps in background knowledge. Following a mixed-methods approach, results from the test administration were analyzed using robust statistical analyses (e.g., MG-Theory, Rasch Analysis, Multiple Regression) and qualitative analysis of the responses.

The Effects of Linguistic Features and Genre of Test Prompt as Predictors of College Writing Placement for L2 Students

Weejeong Jeong (Indiana University)

Friday, July 5, 11:00 – 11.30 am Location: UR3

This study is an investigation of the effects of linguistic features on quality of second language (L2) writers' essays for writing course placement at a mid-western university in the U.S., and by implication at other universities and colleges. This study addresses the following research questions: (1) To what extent do selected linguistic features of narrative (Model I) and argumentative (Model II) essays predict L2 students' performance on a representative university holistic writing placement test?; (2) To what extent is the writing topic related to L2 students' argumentative writing quality as revealed by holistically scored ratings?; (3) Can differences of

linguistic features predict performance in the production of narrative and integrated argumentative writings of L2 students? The data in this study are L2 students' narrative and argumentative scores and their corresponding writing texts at an in-house English placement writing test (N=962). The ordinal logistic regression method was employed to answer the research questions. The results demonstrated that there was no evidence that selected linguistic indices had statistically significant effects on the quality of L2 students' narrative writings (p>.05), but there was evidence that some were significant linguistic features (all associated p-values were less than .01) on the effects of L2 students' integrated argumentative writing quality. Statistical effects of topic on L2 students' argumentative writing were not identified. The results of this study shed light on more rationale for including integrated argumentative writing tasks on a L2 writing placement test and provide meaningful insights for L2 writing instructors and test developers.

"Father brings books; son writes; mother worries; daughter volunteers." Gender representations in Chinese Gaokao English (2014-2023)

Xiaoqin Huang, Xiangdong Gu, Yong Wang (Chongqing University)

Friday, July 5, 11:30 to 12:00 pm Location: Aula

Contrary to the extensive research on gender representation in language teaching textbooks (Sunderland et al. 2020), there is a dearth of research on gender representation in language testing, especially in China. This study aims to provide valuable insights into gender equity in the field of educational testing through the gender representation of listening and reading comprehension sections of English in the national-level Chinese Gaokao (i.e. College Entrance Examination) in the past ten years (2014-2023). To investigate the roles of females and males in Gaokao listening and reading, we draw on content analysis and corpus analysis (Lee 2021) to examine the 31 Gaokao test papers from the perspective of the experiential and relational functions of systemic functional linguistics. Results show the visibility of female pronouns has been increasing significantly for the past decade, but the female roles in family, society, and profession are still far from equal to male ones. In conversations, the number of speeches by males is far greater than that of females, while both genders participate in speaking turns, initiating conversations, and dominating topics. In terms of verb transitivity, while the female roles are increasing in material process, males are still more involved in them. This paper provides suggestions for test paper designers and examination authorities. It also draws implications for future research studies.

The differential impact of COVID-19 on EL proficiency: unpacking language domains

Narek Sahakyan (University of Wisconsin-Madison/Wisconsin Center for Education Research)

Friday, July 5, 11:30 to 12:00 pm Location: HS1

About 10% of K-12 students in the United States are identified as ELs, and they retain this label until they meet state-established criteria for reclassification as Fluent English proficient (FEP). ELs

are also one of the fastest growing student populations: estimates project that by 2025 almost one in four students will be an English Learner. Meanwhile, the COVID-19 pandemic had a profound impact on the K-12 education system nationally, with schools forced to close or adopt remote learning approaches. English learner students in particular have faced significant challenges during and after the pandemic due to a myriad of factors. Applying an "Interrupted Time Series" research design and a theoretical lens of Intersectionality focused on justice, equity, diversity and inclusion for all English learners, I leverage large-scale assessment data from the entire population of students taking WIDA's ACCESS for ELLs Online high-stakes annual English language assessment throughout 2017-2023, a rich set of student-level covariates, and a "seemingly unrelated" (SUREG) regression framework that includes complete test information from all four domains of Reading, Writing, Listening and Speaking totaling about 35 million student by domain by year observations, to document systemic, substantial and consistent gaps in the educational outcomes for several disadvantaged English learner subgroups. Worryingly, and perhaps expectedly, there is strong evidence that some of these within-EL disparities have further widened after the COVID-19 pandemic.

Lost in translation? Reporting the results of a CEFR linking study to educators

David MacGregor, Katie Schultz, Mark Chapman, H. Gary Cook (WIDA, University of Wisconsin-Madison)

Friday, July 5, 11:30 to 12:00 pm Location: HS2

We report on the challenges in linking a language test designed for students in K-12 English-language medium international schools to the Common European Framework of Reference (CEFR) for Languages (Council of Europe, 2001) and communicating those results to a diverse group of educators internationally. CEFR is internationally meaningful and therefore of particular interest for international educators who use assessment results to understand student learning and inform their teaching.

The study used well-established standard-setting methodologies (Cizek & Bunch, 2007; Council of Europe, 2009). Four panels of five international school educators were convened, each focusing on one grade-level cluster (1-2, 3-5, 6-8, or 9-12). For the Listening, Reading, and Speaking tests, we used a bookmarking method; for Writing, a body-of-work method. The study established scale-score cuts for each of the four language-domain tests recommended for classifying learners according to CEFR levels. A technical report (MacGregor, Chapman, & Cook, 2023) describes the methods, results, and implications for test users. In this presentation we address the challenges of making this information consumable for the primary audience of educators in international schools, who may find interpreting technical reports difficult.

Details and findings of the study were documented in the technical report, but additional resources geared toward educators were required to communicate effectively. We report on the development of these resources, including concordance tables showing cut scores by CEFR level for each domain, a short interpretive publication, and a supporting video expanding on topics covered in those resources.

You may say this better: Consequential validity evidence for diagnostic speaking assessment on lexical use

Shungo Suzuki, Hiroaki Takatsu, Ryuki Matsuura, Mao Saeki, Yuya Arai, Yoichi Matsuyama (Waseda University)

Friday, July 5, 11:30 to 12:00 pm Location: HS3

Diagnostic assessments often adopt a discrete-item format to ensure the interpretability of assessment results (Alderson et al., 2015), while admitting the importance of assessing learners' global skills for successful diagnostic assessments (Alderson, 2005). With technological advancement in natural language processing (NLP) and machine learning techniques (e.g., Chen et al., 2018), detailed information about learners' profile can be extracted from performance data, allowing for diagnosing learners' weaknesses through direct performance assessment. The current study evaluates an automated approach to diagnostic assessment through an authentic speaking test in terms of the effectiveness for learning and instruction (i.e., consequential validity; Lee, 2015), focusing on lexical aspects of speaking performance. Sixty Japanese learners of English whose proficiency levels ranged from A2 to B2 on the CEFR scale were randomly assigned to a control and an experimental group. In both groups learners repeated the same oral interview task six times, whereas only the experimental group received the diagnostic feedback on lexical use immediately after each test. Our feedback system identified utterances that lower the probability of being estimated as one level above the current CEFR level (i.e., weaknesses) and minimally paraphrased the utterances with a few vocabulary items that are one level above their current CEFR level. Results showed that both groups improved their range CEFR score across time with a medium effect size. Analysis of students' uptake revealed that learners with higher proficiency were more inclined to adopt suggested expressions, possibly because they were more likely to partially acquire the items.

Comparative judgement as a foreign language assessment tool: an overview of the Crowdsourcing Language Assessment Project

Peter Thwaites (UC Louvain), Magali Paquot (UC Louvain & Fonds De La Recherche Scientifique – FNRS)

Friday, July 5, 11:30 to 12:00 pm Location: UR3

Assessing L2 writing typically means employing professional raters to evaluate texts using a rubric. There are several issues with this approach: it is expensive and time-consuming; attempts to increase reliability can lead to the development of "hyper-specific" rubrics (Pinot de Moira et al., 2022); and the ordinal grades (e.g. CEFR levels) which emerge from rubric-based assessment lack precision.

In Comparative Judgement (CJ), groups of judges compare pairs of performances, such as learner essays, drawing on their knowledge, expertise, and understanding of the target construct to decide which item is "better". By compiling many such decisions, made by many judges, a scale can be generated which ranks each performance from the strongest to the weakest, and reflects an emergent, collective conceptualisation of the target construct.

The Crowdsourcing Language Assessment Project (CLAP) explores the potential of CJ as an alternative to the rubric-based assessment of L2 writing. While existing studies suggest CJ's reliability for assessing relatively short L2 texts representing a broad proficiency span (Paquot et al., 2022), the CLAP project explores CJ's efficacy under more testing conditions. In this presentation, we report on three studies involving CJ tasks involving (a) texts which are relatively long and homogeneous in proficiency; and (b) judges are recruited either from the applied linguistic community, or through the crowdsourcing platform Prolific. The results of these studies suggest that CJ can yield accurate and reliable results under a broad set of conditions, and therefore suggest CJ's broad utility as a method for assessing L2 writing.

Works-in-Progress – Wednesday, July 3, 01:30 to 03:00 pm

AI for dynamic and diagnostic assessment: Automatic task design and mediation to support development of L2 English reading and writing

Ari Huhta (University of Jyväskylä), Dmitri Leontjev (University of Jyväskylä), Roman Yangarber (University of Helsinki), Matthew Poehner (The Pennsylvania State University)

Location: Kaiser-Leopold-Saal (Theologie building)

Our presentation reports on a four-year project aiming to integrate Dynamic Assessment (DA) and Diagnostic Assessment (DiagA). DA is envisaged to help link DiagA results with teaching and to expand the scope of assessments to include language abilities that are still forming. For its part, DiagA provides elaborate procedures for defining and operationalising language constructs. The project is carried out in upper secondary schools in Finland, and it focuses on the development of students' English reading and writing skills and, thus, prepares them for a national school-leaving examination. In the project, an online AI system is leveraged to inform learner progress. Our presentation introduces the AI system, which employs NLP to create grammar and vocabulary exercises from texts input by learners or teachers and supports learners through graduated hints. We report on how we defined and operationalised the constructs that the online exercises target and on the development of DA-based prompting, or mediation, for each construct. Defining the constructs was based on reviews of the literature on L2 reading and writing and extensive surveys of key stakeholders (English language teachers, students, and item writers) for their conceptions of reading and writing in English. These were complemented with teacher and student interviews, think-aloud data and textbook analyses. We will share examples of the reading and writing tasks with their related mediation to illustrate the constructs. We will also discuss the challenges in the design of the tasks, mediation, and the NLP / AI procedures used in the online system.

Diagnosing Chinese EFL Learners' Speaking Proficiency: A Machine Learning-Based Cognitive Diagnostic Modeling Approach

Shuting Zhang, Lianzhen He (Zhejiang University)

Location: Kaiser-Leopold-Saal (Theologie building)

Despite the growing interest in diagnostic assessment, to date, no study has applied cognitive diagnosis models (CDMs) to diagnose EFL speaking ability. The main reason lies in the complicated data collection and labor-intensive human-rating process. Some researchers suggested using machine learning models to improve scoring efficiency (Li, 2021; Lee 2015).

Therefore, this study intends to develop a machine learning-based CDA scheme to diagnose Chinese EFL learners' speaking proficiency. A dataset of 400 examinee responses to the speaking test of an in-house English Proficiency Test is utilized in this study. Three hundred of these responses are used as the training set. In the training process, human raters are hired to score student responses on a scale of 0-2 based on a CDM-checklist. Researchers then feed the computers the students' constructed responses and human scores to develop algorithmic

models for each checklist item. Once satisfactory machine-human agreement is achieved, the scoring models will be applied to score the remaining 100 samples in the testing set to validate the scoring models.

After automatic speech scoring of all the speech samples, appropriate polytomous CDMs will be applied to examine to what extent diagnostic results can be used to provide useful diagnostic information. The study will highlight the advantage of combining machine learning and cognitive diagnostic modeling in assessing the fine-grained cognitive strengths and weaknesses of test takers' EFL speaking with the aspiration that future student responses to the speaking test can be diagnosed automatically using the proposed machine learning-based CDM scheme.

Process and Product in Diagnostic Assessment of Writing: What Do Experts See?

Michelle Czajkowski (Radboud University, Netherlands)

Location: Kaiser-Leopold-Saal (Theologie building)

The diagnostic assessment of writing typically examines only the writing product. However, advancements in technology have made it easier to access the writing process. By examining both the final product and the writing process, assessors may be able to provide richer diagnostic feedback, enhancing test-takers' ability to address specific issues in their own learning and writing.

This study explores the diagnostic potential of examining both the writing product and process, focusing on the academic writing skills of first-year university students at a bilingual Dutch university. The study poses three key questions: 1) What problems do experts identify in the writing samples of first-year undergraduates? 2) What difficulties in the writing process do experts observe, potentially linked to these problems? 3) How consistently do different raters associate problems with observed difficulties?

To address these questions, writing samples from three groups of first-year undergraduates will be collected, and their writing processes recorded. Raters from the university's language and writing centre will holistically assess the texts, identifying problems, and then analyse video recordings of the writing process to understand how observed behaviours contribute to identified issues.

As a work in progress, feedback would be welcome on planned data analysis and how keystroke log data can provide insights into students' writing processes and raters' evaluations. By drawing on the language testing community's experience, the methodology and analyses of the study will be strengthened, along with its potential applications in further research and diagnostic test design.

What Inferences can we Draw from Scores on Paired Discussion Tasks Delivered Through Spoken Dialog Systems? A Study on Construct-Relevant and -Irrelevant Factors

Nazlinur Gokturk (Republic of Turkey Ministry of National Education), Evgeny Chukharev (Iowa State University)

Location: Kaiser-Leopold-Saal (Theologie building)

Research suggests that Spoken Dialog Systems (SDSs) may provide an alternative way to deliver different types of oral assessment tasks (e.g., Ramanarayanan et al., 2016). While several studies have explored the appropriateness of SDS-delivered tasks for second language (L2) oral communication assessment, most focused on comparing task performance between two delivery modes: human versus SDS (e.g., Timpe-Laughlin et al., 2022). To comprehensively understand the construct of L2 oral communication as measured by SDS-delivered tasks, additional research with various sources of validity evidence is needed. Following an argument-based approach to validation (Chapelle, 2020; Kane, 2013), this study aims to examine the extent to which scores on a prototype SDS-delivered paired discussion task (SDS-PDT) relate to performance on another test of L2 oral communication ability (a construct-relevant factor) and two construct-irrelevant factors associated with human-computer interaction: previous experience with SDSs and partner models (i.e., speakers' perceptions of the communicative ability of the systems). Participants will include 250 test-takers (L2 English students at Turkish universities) and eight trained raters. Six instruments to be used for data collection involve a background questionnaire, the Partner Model Questionnaire (Doyle, 2022) translated into Turkish, the SDS-PDT (Gokturk, 2020), the IELTS Speaking section, and relevant rating scales. Factor and regression analyses will be employed to examine the contribution of construct-relevant and -irrelevant factors to scores on the SDS-PDT. By bridging the fields of language assessment and human-computer interaction, this study will enhance our understanding of the construct underlying SDS-delivered tasks and provide implications for task design and administration.

The Role of L1 in L2 Models Adopted to Assess L2 Learners' Writing Quality

Ping-Yu Huang (Ming Chi University of Technology, Taiwan)

Location: Kaiser-Leopold-Saal (Theologie building)

Using L2 corpora contributed by learners with mixed L1s, Monteiro et al. (2018) demonstrated that L2 lexical sophistication indices better evaluated/predicted L2 writing quality than those from L1 corpora. This study further focuses on the role of L1 of L2 corpora, investigating whether L2 output generated by speakers of the same L1 would offer even stronger indices. To address this, we adopted the 11 sub-corpora of TOEFL11, and examined whether the Spanish sub-corpus generated indices best predicting the essay scores of 200 Spanish learners of English. According to the regression analyses performed, the Spanish model was not the one most predictive; two other models explained larger amounts of variance of the writing scores. Lexical frequency, additionally, was found to be a better predictor of writing proficiency than lexical dispersion.

To account for these findings, we further evaluated the similarity of word frequency distributions of those L2 sub-corpora. Based on an X2-based measure proposed by Kilgarriff (2001), we found that the frequency distributions were rather similar, which then made the L2 models exert similar effects on the quality evaluation of the 200 essays. Our current results generally confirm the usefulness of L2 norms as benchmarks to assess L2 writing proficiency, and suggest that and explain why the norms should not necessarily come from output produced by speakers of the same L1.

We will also describe a large-scale "same-L1" L2 corpus that we are currently developing to re-examine these results, and discuss implications for research on L2 writing and automated writing assessments.

A Mixed-Methods Investigation into Raters' Perceptions and Challenges about Rating Prosodic Features

Meng-Hsun Lee (University of Toronto)

Location: Kaiser-Leopold-Saal (Theologie building)

The current study aimed to explore how human raters perceived the assessments of prosodic features and the associated challenges. Enrolled in education-focused master's or doctoral programs in Canada, seven English-L1 and seven Mandarin-L1 raters assessed 330 students' oral reading performances using an adapted 5-point prosody rubric from Rasinski (2004). The rubric evaluated four prosodic features, including expression and volume, phrasing, smoothness, and pace. This research adopted a two-phase explanatory sequential mixed-methods design (Creswell, 2014). Phase 1 focused on quantitative analyses using a many-facet Rasch measurement (MFRM) approach to investigate the difficulty levels of the four prosodic criteria, and Phase 2 focused on qualitative analyses drawing on a think-aloud approach and semi-structured interviews. The MFRM results demonstrated that smoothness (0.53 logits) and phrasing (0.22 logits) emerged as two challenging criteria. The difficulty level of expression and volume (0.05 logits) was close to zero. Pace (-0.80 logits) was the only criterion with a negative logit value. The think-aloud data from eight raters highlighted that the severity levels of the prosodic features, as shown by the MFRM analysis, impacted the order in which raters addressed the categories. Typically, raters began with the less demanding features (i.e., pace and expression and volume), postponing the more challenging criteria, such as smoothness and phrasing, until later in the scoring process. These observations from the think-aloud sessions were further supported by feedback obtained during the semi-structured interviews, which underscored the raters' perceived difficulties in scoring across the four prosodic features.

Building a corpus of academic writing in EMI contexts: Exploring applications for language assessment

Dana Gablasova, Luke Harding, Raffaella Bottini, Haoshan Ren, Vaclav Brezina (Lancaster University)

Location: Kaiser-Leopold-Saal (Theologie building)

Despite an established tradition of corpus-based research revealing insights into academic English writing in Britain and North America, corpus methods remain rare in English-medium instruction (EMI) research, with no large-scale corpus-based analysis of EMI student writing conducted to date. The aim of our project is to address this issue by building the first large corpus representing student writing across multiple EMI contexts. We are currently collecting data (students' written assignments) at seven universities in four countries with large EMI provision (China, Italy, Thailand, UK). When complete, the corpus will contain two million words from over 1,000 students in different disciplinary areas. Currently, the corpus contains over 1.4 million words from 650 texts drawn from students in Social Sciences and Humanities; Science and Technology; Business and Management; and Life Sciences. For language testers, the corpus will allow a deeper understanding of academic writing in these new and emerging target language use domains.

In this work-in-progress presentation, we have two aims. First, we will describe the background and status of the corpus, sharing current findings and research challenges. Second, we will discuss how the EMI writing corpus would have specific benefits for test developers. We anticipate that the corpus will provide valuable foundational insights for domain analysis of EMI contexts, informing the design of appropriate tasks and rubrics for admissions tests. However, we would like to learn what other uses the corpus might have among practitioners and researchers, to ensure that our ongoing research is responsive to the needs of the field.

Accommodations in listening assessment: Exploring the effect of self-paced listening on test scores and anxiety of learners with differing L1 literacy skills

Elisa Guggenbichler (Universität Innsbruck)

Location: Kaiser-Leopold-Saal (Theologie building)

Administrator-controlled listening tests are challenging for learners with low-level L1 literacy skills, who demonstrate decreased L2 listening comprehension (e.g., Kormos et al., 2019) and higher levels of anxiety (e.g., Sparks, 2023). To accommodate these learners, test providers may allow candidates to self-pace, i.e., pause and rewind, the recording. However, previous research into the effect of self-paced listening as a test accommodation yielded inconclusive results (Eberharter et al., 2023). This study presents preliminary findings on the differential effect of self-paced listening on (1) test scores and (2) anxiety of students with varying L1 literacy skills, L2 vocabulary knowledge, and metacognitive awareness across response formats. It further discusses first insights into the use of self-pacing strategies.

A sample of 200 EFL learners completed a subset of items from a standardized B2 listening test in a counter-balanced design. In session 1, participants took an anchor task and two listening tasks in four different conditions regarding administration mode (double-play vs. self-paced) and

response format (multiple-choice vs. short-answer), while the interaction with the audio player was tracked. In addition, participants completed questionnaires on test-taking strategies, listening anxiety, self-pacing use, and the MALQ (Vandergrift et al., 2006). In session 2, participants took standardized L1 reading tasks and the Updated Vocabulary Levels Test (Webb et al. 2017), which served as a proxy for L2 proficiency. Data was analyzed using Generalized Linear Mixed-Effects Modelling. The findings expand research on the complexity and viability of self-pacing to increase fairness in L2 listening tests for students with reading-related learning difficulties.

Exploring the impact of test mode on test takers' turn management in paired discussion tasks

Yaqian Zhang, Yan Jin (Shanghai Jiao Tong University)

Location: Madonnensaal (Theologie building)

Driven by technological advancements and the impact of the COVID-19 pandemic, the shift in spoken interaction from face-to-face (F2F) to online audio-call (AC) and video-call (VC) modes has highlighted the need for research on assessing interactional competence (IC) in technology-mediated contexts. As an essential aspect of IC, turn management (TM) has not been specifically investigated in language testing literature, and little is known about test-takers' TM performances across test modes.

This study, therefore, aims to explore the impact of test mode on test-takers' TM performance, focusing on both quantitative and qualitative TM features. Specifically, the study adopts a convergent parallel mixed-methods design to look into test-takers' TM performances in a paired discussion task delivered in AC, VC, and F2F modes and the effect of oral proficiency on TM performances. The research questions are: 1) how does test mode affect TM features? 2) how does oral proficiency of the test-taker and his/her partner affect TM features? 3) how do TM features affect overall interactional performance? In the main study, 84 participants (42 dyads) of three proficiency groups will complete a paired discussion task given in three modes. Following the test, the participants will be surveyed, and one third of them will engage in retrospective verbal reporting. Conversation analysis and thematic analysis, together with statistical analysis, will be used to analyze the data. The findings of the study are expected to provide validity evidence for/against the online modes as alternatives to the traditional F2F mode in future IC assessments.

Academic language socialization: Transforming research findings into a self-assessment/diagnostic tool for students and teachers

Heike Neumann (Concordia University), Saskia Van Viegen (York University), Sandra Zappa-Hollman (University of British Colombia)

Location: Madonnensaal (Theologie building)

According to academic language socialization (ALS) research on multilingual students in higher education settings, the role of socialization agents is key to university students' academic success and general wellbeing (Duff et. al, 2019). In turn, student self-assessment of their language learning processes promotes active engagement in learning (Andrade, 2019). However, a self-assessment tool designed to guide students in their ALS journey has yet to be developed. The Multilingual University Student Experience (MUSE) Project—a multi-year, multi-site project examining the ALS of multilingual students at three Canadian universities—can address this need. Findings provide insights into the ALS process of multilingual students and should allow us to determine demographic and contextual characteristics that are associated with certain ALS outcomes. It is our goal to leverage these findings into a self-assessment tool for students that generates feedback and recommendations based on the information entered or selected by students about their personal situation and ALS strategies and experiences. The objective of such a tool would be to provide multilingual students in minority language contexts to take charge of their ALS journey. For the MUSE project, we collected questionnaire responses from participating students and conducted student and faculty interviews and focus groups. We will share a snapshot of key findings to date and preliminary plans for the self-diagnostic tool. Following this, we will seek attendees' input, perspectives, and advice on how to proceed with the development of the ALS self-assessment tool.

Exploring English writing proficiency among 15-year-old students in Sweden

Eva Olsson, Linda Borger, Sofie Johansson (University of Gothenburg)

Location: Madonnensaal (Theologie building)

In this presentation we outline a planned research project with the purpose of gaining in-depth knowledge of Swedish students' writing proficiency in English, as this information is essential for future English education and for supporting teachers' assessment of writing proficiency.

The project has three aims:

Create a corpus of 4000 student texts, collected between 2000 and 2022 from the Writing section of the National Test of English for school year 9, including the teacher-assigned ratings.

Identify potential changes in linguistic and textual characteristics of students' writing competence over twenty years.

Determine which linguistic features most strongly predict teachers' holistic ratings of student texts.

NLP tools (e.g., TAALES and TAACO) will be used to investigate central aspects of writing quality including lexical sophistication and diversity, syntactic complexity and cohesion. We intend to

employ statistical analyses on the data, such as factor analysis, regression analysis, and structural equation modeling.

In addition to providing essential information for English education at the policy level regarding changes in writing proficiency over time, the results will indicate how automated text analysis tools may support and complement teachers' holistic assessment of writing proficiency. The national tests will be digitalized in the near future, and, in the longer perspective, the outcomes of this project may provide data for creating an automated tool to support teachers' assessment of the national tests of English.

AI-Supported Automated Scoring of Constructed Response Tasks for Second-Language Academic Reading Proficiency Assessment

Marcello Gecchele (Tokyo Institute of Technology; University of Chicago), Ahmet Dursun (Tokyo Institute of Technology)

Location: Madonnensaal (Theologie building)

This presentation addresses the challenges inherent in assessing advanced second-language reading skills among graduate students in Humanities and Social Sciences at U.S. research universities. The prevalent use of translation exams for this purpose often obstructs students' academic progress due to the misalignment between the exam's intended construct, format, and purpose (Dursun et al., 2021). In response, the University of Chicago Office of Language Assessment introduced the Academic Reading Comprehension Assessment (ARCA™) in 2016, featuring three constructed-response tasks designed to offer a vastly superior assessment of advanced reading.

The presentation describes how novel evaluation models from the field of Natural Language Processing were leveraged to automate the scoring of the ARCA constructed-response tasks. It first demonstrates the application of content alignment models based on Large Language Models to assess student-generated summaries by extracting Idea Units from student responses (Gecchele et al., 2022) and comparing them against an established scoring rubric. It then highlights challenges in scoring student responses to the open-ended questions and discusses the application of existing models for the evaluation of AI Question Answering models (Fabbri et al., 2022) when compared to human responses. Next, it demonstrates how techniques to score translations produced by Generative AI, such as BERTScore (Zhang et al., 2020), perform when compared to human translations, and whether they can substitute for human judgment in the assessment of student translation tasks. Lastly, it discusses potential avenues and challenges for utilizing AI to develop ARCA exams and generate the corresponding scoring rubrics derived from the source texts.

Developing a scenario-based test to assess the language assessment knowledge of EFL teachers in Chile

Salomé Villa Larenas (Universidad Alberto Hurtado, Chile)

Location: Madonnensaal (Theologie building)

While the use of self-reported knowledge surveys provides language assessment literacy (LAL) researchers with insights into what stakeholders' knowledge of language assessment might be, their results might strongly depend on stakeholders' self-perception and are highly sensitive to social desirability bias (Riazi, 2016). Some efforts to develop more direct instruments have been reported in the last years (e.g., the Language Assessment Knowledge Scale LAKS by Ölmezer-Öztürk & Aydin in 2018, and the Teachers' Language Assessment Literacy Test TLALT by Zhang in 2023). Yet, these instruments present some weaknesses in their type of item (true-false, LAKS) or their ultimate focus (still on self-reports of knowledge, TLALT). This Work-in-Progress paper will report on the development of an online instrument to assess the knowledge of language assessment of public-school English teachers in Chile. The online test consists of language assessment scenarios for which the teachers need to select the most suitable solution from a set of options. During the Work-in-Progress session, I will describe the development of the test and its challenges. In addition, initial results from the piloting administration to 50 English teachers in Chile will be shared, to shed light on their language assessment knowledge as well as on the quality of the instrument. Finally, it is hoped that the Work-in-Progress session will raise the discussion on the benefits and challenges of a language assessment knowledge test, the ways of approaching its results analysis and its potential to contribute to research on language teachers' LAL in different contexts.

There are C-Tests and C-Tests: Digitalised Formats and Reduced Times - Changed Constructs?

Anastasia Drackert (g.a.s.t. e.V. & Ruhr University Bochum), Anna Timukova (g.a.s.t. e.V. & Ruhr University Bochum), Franziska Möller (g.a.s.t. e.V.)

Location: Madonnensaal (Theologie building)

C-Tests have recently gained much attention in language testing. The new time-reduced speeded C-Test (S-C-Test) appears to promise even more in terms of efficiency, but has not yet been sufficiently investigated. In particular, many questions regarding the underlying construct of a S-C-Test remain unanswered. It is hypothesized that, in contrast to a canonical C-Test with a generous time limit of five minutes primarily measuring the amount of learners' declarative and procedural knowledge, a S-C-Test additionally gauges the level of automaticity of their skills and should thus be better at predicting learners' oral skills.

In the presentation, we report on partial results of a large study that aims to investigate the role of the time variable in the construct of computer-administered C-Tests in English, German and Russian. A total of 230L2 English learners took a canonical and a speeded C-Test along with seven instruments measuring their declarative and procedural knowledge, general oral language proficiency and typing skills. For the presentation, we will report on the findings for three research questions: How does the time variable influence the reliability of computerised C-

Tests?How does it influence learners' scores depending on their proficiency level?Which components of L2 proficiency (declarative, procedural knowledge and automaticity) are better predictors of differently timed C-Tests?Answers were gained through reliability analyses, ANCOVA and structural equation modeling.

The study contributes not only to our understanding of the C-Test construct but also to our understanding of new constructs and assessment formats measuring procedural and declarative knowledge.

Indigenous Assessment Criteria in a Test of English for Tourism Students: Adopting Pill's (2016) Approach

Gina Ward (Concordia University, Canada)

Location: Madonnensaal (Theologie building)

It is important to include indigenous assessment criteria when assessing language for professional purposes. Pill's (2016) approach for using these criteria to develop more authentic assessments in a healthcare setting do not appear to have been adopted in a tourism setting, another specific-purpose language testing context. These assessment criteria are especially useful in tourism training programmes, to ensure that there is positive washback on teaching and learning, i.e., that trainees are adequately prepared to enter the workforce. In the context of tourism students at a rural university in Central America, there is a lack of clear English language assessment criteria from tourism employers to guide teachers and learners as to the required proficiency levels for employment. Adopting the qualitative approach used by Pill (2016) to develop professionally relevant language assessment criteria using domain experts' judgments of video recordings and reports on trainee performance, the goal of this study is to develop assessment criteria for tourism professionals in a rural context in Central America. Findings from the analysis are expected to (1) generate professionally relevant criteria for English language assessment in tourism workplaces in a rural area, and (2) provide insights into the extent to which Pill's (2016) approach for developing indigenous assessment criteria in the workplace can be applied to other professional language testing contexts. In this WIP, I will report on the background, objectives, and data collection processes of the study. I would like to get input on suggestions for data analysis and subsequent development of assessment criteria.

Exploring test takers' experiences with instructions in reading-into-writing tasks

Lies Strobbe, Goedele Vandommele, Sterre Turling (KU Leuven, Belgium)

Location: Madonnensaal (Theologie building)

The CNaVT offers CEFR-aligned task-based exams and includes a C1-language test for educational and professional purposes. This test features integrated reading-into-writing tasks, reflecting real-world challenges in educational and professional settings, requiring test takers to interact with and summarize multiple information sources.At the C1-level, the CEFR emphasizes dialogic reading in the writing of well-structured texts, differentiating between main and secondary ideas

in descriptors for thematic development, written production, reading comprehension, and mediation skills. Feedback from examiners and teachers, however, suggests a preference for tightly controlled, specific instructions in Dutch as a Foreign Language, aligning with good practice guidelines for clear task completion criteria (ALTE, 2002).

This research explores test takers' experiences with open-ended and specific instructions, aiming to understand their impact on cognitive processes, overall performance, planning, response cohesion, stress, anxiety, motivation, and feelings of competence. These factors, if not addressed, could introduce construct-irrelevant noise, potentially affecting the validity and reliability of the C1-test.

The study, conducted in two phases, engaged twelve students evenly divided into two groups: six with open-ended instructions and six with highly concrete instructions guiding them toward key points. After the task, experiences were discussed to gauge the impact of instruction specificity on various aspects. Test takers' responses were scored analytically and holistically to determine the most effective approach. Follow-up phases involved three test takers for each instruction type, using think-aloud procedures and draft papers to delve deeper into cognitive processes and their interaction with provided instructions. Results from the study will assist in the design of better reading-to-writing test task in the future.

Validating Prompts and Rubrics in an Office-Hour Role-Play Task – a mixed method approach to local test reformation

Stephen Daniel Looney (Pennsylvania State University), Haoshan (Sally) Ren (Lancaster University)

Location: SR VI (Theologie building)

This project focuses on validating and improving a local test for International Teaching Assistants (ITAs), examining the validity issues of role-play tasks which are rarely addressed in ITA assessment literature. Informed by the evaluation inference argument-based framework, the study analyzes task prompts in an office hour role-play task from a locally-administered ITA test.

The quantitative investigation, using Many-Faceted Rasch Measurement, includes 521 examinees and 39 raters, revealing significant differences in prompt difficulties and rater behaviors. This analysis revealed significant differences between prompts and rubric dimensions resulting in differences in prompt difficulties and rater behaviors.

The qualitative analysis, utilizing Conversation Analysis, investigates language patterns elicited by the prompts, revealing distinct sequences of action based on prompt variations. Notably, the first action required of the test-taker differs across prompts. Further qualitative investigation aims to understand how these differences are influenced by prompt content or context.

Anticipated outcomes include refinements to the ITA test's rating criteria and prompt formulation, addressing validity concerns associated with role-play tasks. By offering insights into the impact of prompt variations, the research aids ITA test designers and raters, enhancing accuracy in evaluating ITA abilities and informing decisions on academic roles. These refinements have the potential to shape future iterations of similar assessments, improving precision in assessing ITA abilities and guiding decisions on academic placements.

ChatGPT versus human raters in integrated writing assessment: Comparing rating performance across test taker levels and rating criteria

Haeyun Jin (Korea National Open University, Republic of South Korea)

Location: SR VI (Theologie building)

The advent of ChatGPT has garnered attention for its potential application in second language writing instruction and assessment. While existing research has predominantly explored the utility of ChatGPT as a tool to assist students during the writing process, there remains a dearth of systematic investigation into its performance as a "reliable rater." In particular, there is a notable lack of empirical evidence regarding how the quality of ratings assigned by ChatGPT varies according to test takers' proficiency levels and distinct rating criteria. This work-in-progress study aimed to address such a gap by exploring the performance of ChatGPT as a rater for reading-to-write integrated assessments in comparison to human raters, focusing on two aspects: 1) the accuracy of ratings across test takers' proficiency levels and analytic rating criteria, and 2) the specific aspects of reading-to-write ability that affect rating decisions. First, a sample of 90 essays from a corpus of reading-to-write essays will be rated by ChatGPT and two trained human raters. Using Many-facet Rasch Measurement analysis, the accuracy of ratings assigned by ChatGPT will be compared with those of human raters in each proficiency level and rating criterion. Then, the human raters will rate ten essays while verbalizing their thought process. Similarly, ChatGPT will be asked to rate the same essays, verbalizing its rationale for assigning the ratings. Using the inductive thematic analysis method, the recorded rating process will be analyzed to examine the specific aspects of rating reading-to-write ability that received raters' and ChatGPT's attention during the rating process. The potential implications of this study for integrated writing assessments will be discussed.

Diagnosing L2 English Academic Reading Ability in the CEFR Context: A CDA Approach

Tugba Elif Toprak Yildiz (Izmir Democracy University; University of Bremen), Claudia Harsch (University of Bremen)

Location: SR VI (Theologie building)

Cognitive Diagnostic Assessment (CDA), a relatively new methodology, has attracted increasing attention in language assessment. CDA yields fine-grained feedback about examinees' mastery status in a given ability. So far, CDA applications have mostly been reverse-engineered retrofitting, where such feedback is extracted using non-diagnostic assessments. Nevertheless, there have been recurrent calls in the field for an inductive approach to CDA, in which tasks are created from scratch, based on a cognitive model of interest. Hence, this study aims to develop a set of diagnostic assessment tasks targeting L2 English academic reading in an EFL higher education setting, where the ability is vital to academic success. The study is inductive and based on the CEFR framework. The study features several stages that involve defining the construct and generating the cognitive model of L2 English academic reading, generating and administering the tasks, applying psychometric modelling, and interpreting results. The study uses various quantitative/qualitative techniques (e.g., systematic literature review, expert interviews, textual

analysis, think-aloud protocols, and diagnostic classification modelling) in these stages. Considering there is a paucity of inductive CDAs, one important contribution of this study is its inductive nature. Secondly, even though the CEFR has underlain many language assessments, to our best knowledge, the framework has not been used in the CDA context. Along with these contributions, the study bears useful implications for the use of a relatively novel CDA methodology in language assessment, specifically concerning assessment design/development and cognitive diagnosis.

Writing assessment literacy and the factors shaping its development: the case of pre-service and in-service English and French second language secondary school teachers in Quebec

Amira Ben Hmida (University of Montréal)

Location: SR VI (Theologie building)

Students' ability to write effectively in a second language has been a shared concern for researchers to investigate especially in Quebec. The increasing importance attributed to the quality of students' writing finds a part of its argument in the crucial role that writing assessment plays as a gatekeeping mechanism that allows or otherwise denies students access to certain educational resources or opportunities (Weigle, 2007). Hence, teachers' ability to fairly assess their students' attainment of and progress towards meeting different aspects of the writing competence, hereafter referred to as teachers' writing assessment literacy (WAL), is deemed essential especially within the current vision of assessment that "must contribute to improving the student's quality of spoken and written language" (MEQ, 2003, p. 20). When it comes to developing teachers' WAL, it is advocated that teacher training is the ultimate quality assurance to reach this aim (DeLuca & Klinger, 2010). Yet, as far as writing assessment is concerned, researchers believed that the field of second-language writing has overlooked the preparation of second-language teachers and focused mainly on students' writing competence (Hirvela & Belcher, 2007). Due to the lack of research examining teachers' writing assessment competence, our doctoral research aims to answer the following two research questions: What is the level of writing assessment literacy of pre-service and in-service secondary school English and French second language teachers in Quebec? And what are their writing assessment needs? What are the factors influencing the development of teachers' writing assessment literacy?

Poster Presentations – Thursday, July 4, 02:00 to 03:30 pm

A Digital Mapping of High Leverage Communicative Practices in School-Age Content-Area Contexts

Lynn Shafer Willner (University of Wisconsin-Madison)

As digital design and delivery expands the variety of modalities and interactions made possible in language testing, test contexts can more closely mirror real-world language usage. To identify these contexts, this poster presents findings from an extensive study identifying high leverage communicative practices for Kindergarten-Grade 12 (i.e., ages 5-18) in state academic standards for English Language Arts, Mathematics, Science, and Social Studies, used by approximately four-fifths of states in the U.S.

Findings provide a mapping of the prevalence of Key Language Uses by four content areas and six grade-level cluster clusters (K, Grades 1, 2-3, 4-5, 6-8, and 9-12). This crosswalk provided an evidence-base for the updates applied to the WIDA English Language Development Standards Framework, 2020 Edition and informed the redesign of test items for WIDA ACCESS, an annual summative English language proficiency test. Study research questions, methods, acceptability measures, findings, and evidence are available at https://www.wcer.wisc.edu/publications/abstract/wcer-working-paper-no-2023-3.

This poster not only shares the results of this extensive review of communicative practices embedded in U.S. K-12 state academic standards, it also allows conference attendees to learn about a method for sharing and connecting their own language standards, frameworks, and descriptors using an open-source, digital format. It is time to move beyond PDFs as the primary format for sharing standards, frameworks, and descriptors. Thus, as discussed by Harding (2021), test developers can support the spread of digitalization practices around language standards, frameworks, and descriptors, including their integration into tests, connections with aligned curriculum resources, and use during instructional planning.

A Multifaceted Investigation on the Assessment of French Language Competence of K-12 Teachers in Canada

Samira ElAtia (The University of Alberta), Komla Essiomle (The University of Alberta), Elissa Corsi (The Alberta Teachers Association, The University of Alberta), Pierre Rousseau (The University of Alberta, The Alberta Teachers Association), Danielle Dallaire (The Alberta Teachers Association)

French Immersion (FI) teachers must be language models for students. To this end, FI teachers need to have adequate French language ability. Universities providing pre-service teacher education must ensure and assess students' language skills to determine their level of proficiency at the beginning of the program, with the goal of providing opportunities to improve and enhance their qualifications upon graduation. In the workforce, principals and school board administrators must ensure through an evaluation process that the teachers they hire have sufficient language skills to provide a bilingual learning environment for students and are committed to their ongoing improvement. However, this is not necessarily the case. Through (a) a survey and focus group discussion administered to 12 pre-service teachers enrolled in a university education program in French and 10 junior teachers (b) an extensive survey

administered to 48 FI professionals in Alberta, followed (c) by semi-structured interviews with four principals, educators' perspectives on the needs and challenges of French assessment were analyzed. Findings indicated that a standardized assessment is used for all new students enrolled in the teachers training program, but no other formative assessment is used to determine their language progress. Conversely, in the workforce, a standardized language assessment process does not exist during teachers' recruitment process and the planning of language professional development activities is often affected by the size of the school board, the location of the school, the resources available, and teachers' language insecurities.

Augmented Assessment: Shaping EFL Speaking Assessment with Mobile AR Technology

Jung-Hee Byun (Gyeongsang National University High School)

This research explores using Mobile Augmented Reality (MAR) to assess language proficiency. The study was driven by the need for innovative and safe interaction methods, particularly during the COVID-19 pandemic. The research introduces a MAR-based speaking assessment called MAR-mediated speaking assessment (MARST) to evaluate its impact on performance, task types, and scoring. The study employed a mixed-method approach, investigating MARST's comparability with traditional speaking assessments, which has implications for assessment settings, test-takers' perceptions, and linguistic features of MAR communication.

The study engaged 200 Korean high school students and utilized the "Eco English Test" app for tasks such as dialog completion, poster description, expressing opinions, and explaining recycled devices. The tests aligned with national curriculum standards and aimed to simulate real-life communication scenarios.

The findings through Multi-Trait Multi-Method (MTMM) and factor analysis confirmed MARST's score reliability and unidimensional construct without a test method effect. Many-Faceted Rasch Measurement (MFRM) analysis revealed consistent scoring across raters, high reliability, and accurate discrimination of test-taker ability. Surveys showed that test-takers found MARST immersive and authentic, enhancing the representation of the speaking construct.

The qualitative analysis reflected increased engagement and utilization of language resources, supporting the validity of MAR in speaking assessments. Although technical and cheating issues were noted, the study concluded that MAR technology positively affects test-takers' cognitive and affective processes, providing a solid context for assessing speaking proficiency and supporting the validity argument for such innovative assessment methods.

ChatGPT in the Classroom: Pre-Service English Language Teachers' Perspectives on AI Integration in Language Assessment Training

Asli Lidice Gokturk-Saglam (University of South Eastern Norway)

ChatGPT, the artificial intelligence-powered language model, has garnered significant interest in the field of education worldwide. By delving into the challenges and benefits of integrating AI in language assessment training, this study reports the findings of a mixed-method research into exploring how this innovative concept of using AI-powered chatbot can be integrated into a

measurement and assessment course that is offered online in flipped learning to pre-service English language teachers at a Turkish university. The intervention unfolds in three cycles, with students progressively engaging with ChatGPT to explore its functions, gather information on course content, and actively use the chatbot in designing assessment tasks, guidelines, rubrics, and item writing. Qualitative analysis was conducted using NVIVO software to capture nuanced insights from written reflections, online surveys, and one-on-one discussions with 29 students. Additionally, Statistical Package for the Social Sciences (SPSS) was used to complement the qualitative findings. In each cycle, perceptions of students (regarding the benefits, pitfalls, concerns, and suggestions) were elicited and used to calibrate the assigned tasks in the following cycle. The findings, derived from a combination of qualitative and quantitative analyses, contribute significant insights into the educational application of ChatGPT. Presenting a novel approach to language assessment training, the study not only sheds light on the potential benefits and challenges of utilizing ChatGPT in the local context but also offers valuable insights for teacher educators and instructional designers globally in mitigating likely difficulties and charting pathways forward for using ChatGPT as a pedagogical tool.

Computerized Dynamic Reading Assessment as an Enhancer of Reading Development of Students with Lower Proficiency

Chansak Siengyen, Punchalee Wasanasomsithi (Chulalongkorn University, Thailand)

Emphasizing learning potential of learners and measuring response to teaching (Dixon et al., 2022), dynamic assessment (DA) is considered an alternative method to compensate for what traditional assessment may lack (Naeini & Duvall, 2012). Due to the large size of students enrolling in English reading class, a human-to-human mediation is impractical. The present study investigated the effects of a computerized dynamic reading assessment (CDRA) program of 30 undergraduate students in northern Thailand. Using an interventionist model, the iSpring Suit 10 software was utilized to develop the program that offered students pre-fabricated, and strategy-based mediational prompts for each item on the reading tests which were validated by the experts and selected based on the Fox index ranging from 8-12, with the reliability of all tests met the requirement at above .70 of KR-20. Following the test-train-test design, the study, spanning seven weeks, encompassed a training session, and data collection through non-DA pre- and post-tests, and four CDRA program tests. Qualitative data were obtained through an attitude questionnaire and semi-structured interviews. Quantitative findings indicated a positive impact of the CDRA program on students' reading comprehension, revealing both independent (ZAD) and assisted (ZPD) performance abilities (Vygotsky, 1998). The qualitative findings highlighted the students' positive attitudes toward the developed program. In this presentation, implications and recommendations regarding the design of the CDRA program's interface and in-class utilization of the program to promote reading comprehension of EFL students, particularly those with a lower level of English proficiency, will be discussed and exemplified.

Developing a new writing rubric as part of an exam reform project

Mark Derek Chapman (WIDA at the University of Wisconsin-Madison), Tanya Bitterman (Center of Applied Linguistic), Heather Elliott (WIDA at the University of Wisconsin-Madison)

In this paper we report on reforms to an assessment system as a result of the release of updated standards. We highlight how the need to redesign the assessment can provide the opportunity to improve long running issues. This paper reports on the process and final product of a writing rubric revision project. The aim of the project was to develop a new rubric, aligned to the 2020 WIDA ELD Standards, for scoring responses to writing tasks on two assessments for English learners in K-12 U.S. public schools. The rubric, in use since 2015 has had a number of issues, including score points that are rarely awarded, clustering of scores awarded in the center of the rubric, and issues with rater reliability.

We describe the processes followed to develop the new rubric (Banerjee et al., 2015; Becker, 2018; Turner & Upshur, 2002), which consisted of a data-informed approach to rubric development, working with a corpus of student responses (n=324). Then we recount the review and validation steps we undertook to refine the rubric. These steps included multiple rounds of subject matter expert review from diverse stakeholders. Subsequently, the new rubric was provided to educators for their feedback. Educator feedback confirmed the need for an updated rubric and the addition of detailed scoring guidance with an expanded glossary. Finally, a group of 51 trained raters used the new rubric to score 1,200 responses in a crossed design that allowed for a multi-faceted Rasch analysis validation phase (Li, 2022).

Developing an efficient EAP placement test using integrated tasks to assess receptive and productive skills

Rebecca Yeager, Alfonso Martinez (University of Iowa)

Although Integrated Assessment (IA) is well-supported by the assessment literature (Llosa & Malone, 2019; Rukthong, 2021), language programs sometimes hesitate to use it for local placement tests. Two objections are commonly raised: first, that developing and rating IA will be prohibitively impractical, and second, that integrated tasks cannot reliably identify support needs for receptive skills. This poster documents the development of an English for Academic Purposes placement test consisting of summary and argumentative writing tasks, an oral interview, and 20 selected-response items. The test maximizes efficiency through tightly-controlled task specifications, resulting in average rating time-to-decision of 23 minutes per sample. First, since lexical overlap from listening sources is associated with summary quality (Kyle, 2020), we chose to explicitly allow patchwriting from listening sources, reducing the burden on raters. Second, to assess receptive skills, we developed an analytic rubric (Ohta et al., 2018), addressing concerns about construct representation by having not one but multiple integrated tasks (Asención, 2008), supplemented by a few selected-response tasks targeting details, inference, and vocabulary skills (Rukthong, 2021). Internal validation measures, including descriptive statistics and many-faceted Rasch analysis, indicate that the test is targeting an appropriate difficulty level, discriminates well, and displays stable inter-rater reliability (exact agreement .54; adjacent .92; severity: mean 0.00, SD .24). External validation measures, including post-assessment surveys and diagnostic intake checklists, indicate that the test supports student learning and enables accurate placement decisions. We hope that our experience may encourage local test developers to incorporate IA for placement purposes.

Evolving Modalities: Exploring Changes in Language Assessment Practices in Higher Education

Michelle Reyes Raquel, Simon David Boynton, Wim Vergult, Grace Chang, Anne Hu (University of Hong Kong)

This presentation reports the results of a study that investigates the evolving landscape of multimodal language assessment in higher education, shaped by the integration of visual, digital artifacts, and GenAi tools. It aims to understand the types of course assessments, the integration of oral and written language abilities with visual and digital criteria, and the reasons behind faculty teachers' assessment choices.

A comprehensive survey of teachers across various disciplines, followed by in-depth interviews, unearthed the assessment types and reasons for their use. Student interviews captured their perceptions of these assessments. Our findings revealed a shift towards multimodal assessments, with lower weightage given to oral and written language abilities compared to content or creativity. Teachers' motivations for this shift included a desire for enhanced engagement and improved learning outcomes. Barriers included steep learning curves for new skills, personal preference, and access to necessary resources. Students found the changes challenging but appreciated acquiring transferable skills and took pride in their work. Overall, this study highlights the changing landscape of multimodal language assessment and provides insights to guide educators and institutions in supporting this pedagogical shift.

Examining the Writing Style of ChatGPT using AI-Generated Text Detection

Peter Kim (Cambridge Boxhill Language Assessment)

The AI revolution has democratized the use of large language models (LLM) by making them more accessible and available for the general public. This has led to an increasing reliance on such technology by language learners for writing assistance and, in some cases, generate text ostensibly as their own. Given this backdrop, the purpose of this study is to investigate ChatGPT's writing style using two approaches, 1) employing stop-words to conduct a stylometry analysis with Burrow's delta method. 2) In the second method, content words devoid of stop-words were used to detect AI generated text by using random forest classification (RFC). After fitting the model, feature importance was extracted for further analysis for both methods. The first method primary uses stop-words for classification while the second method uses content words only. These two complementary methods were designed to shed light on the distinct features and styles of writing employed by ChatGPT. Comparison between Burrow's delta and RFC revealed that while stop-word-based classification recognized most instances of GPT-generated text, it also had a high false positive rate. Taken together with the results of RFC, this study concludes that ChatGPT tends to favor a distinct style of writing essays. This has implications for the “teaching” of expository, persuasive, and argumentative essays by LLMs, suggesting a need for revisiting and reforming the current approach to writing. It is possible that as the use of LLMs become more pervasive, diversity of writing styles may be negatively impacted.

Exploring Language Assessment Literacy: What do Taiwanese CLIL teachers need to learn and relearn?

Yu-Ting Kao (National Cheng Kung University, Taiwan)

This study investigated Taiwanese CLIL teachers' assessment perceptions and practices and identified their training needs in designing assessment. A language assessment literacy (LAL) survey adopted from Wu (2014) and Kremmel & Harding (2020) was applied to explore 245 in-service teachers' perceptions and practices. Results of the survey first indicated that 68% of the teachers (N=167) did not have previous training in assessment-related issues, and 79% of them (N=194) responded with unfamiliarity with the designing principles of language assessments. Through exploratory factor analysis, major areas in which teachers report training needs include (1) the assessment forms and approaches to English speaking and listening skills, (2) item writing techniques, and (3) the interpretation of assessment results, particularly in identifying students' learning difficulties in either the content area or English language. Secondly, an inquiry-based approach was applied to examine the aspects that might be left unclarified in the survey. 50 teachers were invited to the semi-structured interviews, and the findings showed that teachers were uncertain about the design and use of assessment criteria in bilingual classes. They also reported parental concerns and administrative difficulties when implementing CLIL assessments. These results have pedagogical implications for developing language assessment literacy in CLIL and they could be used to inform the development of new textbooks and the provision of relevant training programs for CLIL teachers in Taiwan. This study contributes to the construction of a LAL knowledge base that helps teachers gain different perspectives toward the function of assessment and realize its relation to CLIL instruction.

Implementing a Learning-Oriented Academic Reading and Writing Assessment Model at a Tertiary Level in Thailand

Punchalee Wasanasomsithi (Chulalongkorn University, Thailand)

This presentation reports on the implementation of a learning-oriented academic reading and writing assessment model developed based on the concepts and frameworks of learning-oriented assessment proposed by Carless (2015), Jones and Saville (2016), Purpura and Turner (2014), and Turner and Purpura (2016), incorporating three key components, namely learning as assessing tasks, developing assessment expertise in learners, and learner engagement with feedback. The implementation of the learning-oriented reading and writing assessment model, divided into two reading and two writing modules implemented over a 16-week English for Academic Purposes (Science) course offered to 30 second-year science-majored students at a public university in Bangkok, Thailand, will be described, highlighting how both summative and formative assessments were utilized to promote learners' academic reading and writing ability and learning engagement. In particular, pre-reading tests, post-reading tests, pre-writing tests, and post-writing tests administered before and after the end of each of the four modules were used to gather quantitative evidence to determine if learning had actually taken place, as well as whether any of the students needed further instruction, practices, and assistance from the instructor, whereas learners' journals, teacher's classroom observation notes, and in-depth semi-structured interviews yielded qualitative data that shed further light on learners' learning performance and engagement while in the learning and assessing processes.

Study findings led to a conclusion that learners' reading and writing ability and learning engagement can be enhanced with a learning-oriented academic reading and writing assessment model, thus supporting the integration of learning-oriented assessment into classroom instruction.

Language testing and assessment academic production in Latin America: a bibliometric analysis

Gladys Quevedo (University of Brasília, Brazil)

The assessment of additional languages through standardized exams or summative and formative assessment practices used in classrooms has been on the agenda in many Latin American countries. Although they drink from the source of knowledge spread worldwide, these countries seek specific solutions for their socio-historical-cultural-educational realities through the combination of global and local knowledge (Dendrinos, 2013; Dimova, Yan, & Ginther, 2020). The concept of glocalization (Robertson, 1995; Trippestad, 2016) calls for the idea of a strong and intense connection between the local and the global, associated with the profound transmutations of everyday life that affect teaching practices and pre-existing modes of behaviour. Despite an immense socio-historical-cultural-educational diversity that characterizes the twenty Latin American countries, there are many common or similar problems. Bearing those questions in mind, this poster will present a bibliometric analysis (Donthu et at., 2021) of the academic production on language testing and assessment in Latin American countries. Using automated text analysis for the titles of all academic papers published between 1985 and 2023 in Latin American journals of relevance, we track the evolution of the academic knowledge on the assessment of additional languages produced in Latin America, identifying the emergence and growth of specific topics and groups of journals with similar agendas. We discuss current gaps in the literature and the consolidation of specific methodological practices, as well as the incorporation of local contextual knowledge into the language assessment scholarship produced in Latin America.

Language testing and language policy change: A case study from Ukraine

Karen Jeanette Dunn (British Council), Jamie Dunlea (British Council), Zhanna Sevastianova (British Council, Ukraine), Irina Umbetaliyeva (British Council, Ukraine), Martin Murphy (Australian Council for Educational Research)

The intersection between language testing, language education, research, and policy has long been of interest in the language testing community. This topic is rarely without controversy, and often takes us, as language testers, out of our comfort zone. This paper will report on a particularly challenging context to implement best practice in language testing as part of a bigger policy and research landscape, that of the proposed move to promote the use of English as one of the languages of international communication in public life in Ukraine.

Ministry-commissioned research was designed to provide critical information for understanding the current capacity of the English school-teacher population. Computer-based English language proficiency testing was coupled with an attitudinal questionnaire. The special circumstances of test administration required adaptation to the test-taker journey, including (but not limited to) registration, use of remote invigilation, and the setting up of special protocols to cover unforeseen disruption. This reflects a lesser-discussed aspect of test localisation, that of the necessity of adapting test delivery to local needs (Dunlea et al., 2019). In the current study this was balanced with a requirement to uphold methodological rigor in the research design. We will report on the logistically demanding sampling approach, essential for upholding the legitimacy of the study and for making meaningful inferences; communication surrounding the use and relevance of testing (cf. Chalhoub-Deville and O'Sullivan, 2020); and the impact of the study on individuals, with teacher participation taking place (often) in the face of personal struggle and loss.

Measuring verbal and non-verbal features of L2 learners' spoken interaction: Rethinking automated speaking assessment

Anna von Zansen (University of Helsinki, Finland)

Automated speaking assessment (Zechner & Evanini 2020) is often limited to individual performance. Moreover, scales used for assessing second language (L2) spoken interaction rarely include non-verbal behaviors such as gaze and gestures or intonational cues (see e.g. Council of Europe 2020), although they are important in non-test conversations.

This poster presents aims and starting points of the Aasis research project (2023–2027), which focuses on verbal and nonverbal features of L2 Finnish learners' spoken interaction and develop ways to assess these automatically. Aasis builds on a previous project, DigiTala, which developed an ASR-based (automatic speech recognition) online tool for assessing L2 Swedish and Finnish learners' speech automatically and providing automated feedback to the language learners.

The project aims to expand ASR-based L2 speaking assessment to cover also assessment of interaction skills. In addition to L2 speech, the research interests include non-verbal communication such as body language and interactional phonetics.

The methods include 1) videoing academic L2 Finnish learners' dialogues, 2) training human raters to assess and transcribers to annotate learners' performances, 3) analyzing the ratings using Many-facet Rasch measurement and 4) experimenting machine learning methods to predict human ratings. Where possible, the data collected and tools developed are published following the principles of open science.

Automatic assessment of interaction improves authenticity and reliability of spoken L2 assessment by enabling ASR-based dialogue speaking tests and supporting human raters' and teachers' work. Moreover, automatic scores produced by the machine could be used for providing automated feedback to the learners.

Scoring validity of an AI-powered essay-scoring system for a task-based writing test

Yoshihito Sugita (Meiji Gakuin University, Japan)

This poster mainly aims to demonstrate how far we can depend on the scores made by an AI-powered essay-scoring system for a task-based writing test (TBWT). The constructs of TBWT are Accuracy and Communicability, which form the basis of two elicitation tasks. The system workflow will be presented, and a clear explanation of the scoring the tasks will be given. To examine the scoring validity of the AI-powered essay-scoring system, fifteen human raters (HUM) scored each of the forty scripts collected from twenty undergraduate students who took TBWT, and their scores were integrated using FACETS. The same scripts were analyzed and scored by three automated essay-scoring systems: a rule-based system (AES), machine-learning system (AIM), and deep-leaning system (AID). FACETS analysis of the scoring results revealed that (a) the strongest correlation was found between the scores of HUM and AID; (b) the scores of AES and AID provided a reasonable fit to the Rach model; (c) the scores were also validated by bias analysis between the rating systems and samples; and (d) the correlation between the students' Criterion scores and the AID-predicted scores showed acceptable levels of validity related to the criteria. In additional study, 160 high school students exhibited consensual agreement in their grades, with A, B+, B, B-and C according to the cut-off score set for each grade. These findings were discussed from the point of view of further improvement of the AID system.

Test-Taker Insights in Language Assessment Literacy: The Road Less Travelled

Andy Jiahao Liu (University of Arizona)

In this presentation, I examine and highlight what I see as an important missing element in language assessment literacy scholarship: the voice of test-takers. By "language assessment literacy," I mean the knowledge, skills, and principles of using assessments to maximize beneficial learning washback in an equal and inclusive manner. During the past 10 years or so in which language assessment literacy has come to prominence in the language testing and assessment field, much has been examined. In particular, I have learned a great deal about the language assessment literacy of teachers (Baker & Riches, 2018), university administrators (Deygers & Malone, 2019), policy-makers (Pill & Harding, 2013), and other stakeholders. Though collective efforts have presented important steps toward addressing what Inbar-Lourie (2017, p. 267) advocates the "assessment circle should be expanded to include consumers of language assessment literacy, such as parents and students" in our field, much work remains to be done if we intend to close that gap. With this presentation, I aim to generate new insights for practice and expand the research landscape of test-takers in language assessment literacy scholarship that builds on the momentum gained by the 2017 LTRC and Jin (2022). I do so by identifying and briefly explaining what I see as common themes in language assessment literacy research, followed by an exploratory qualitative study demonstration on Chinese tertiary students' (N = 25) language assessment literacy. By way of conclusion, I discuss future research recommendations for a deeper understanding of test-takers in language assessment literacy studies.

The Process and Impact of Streamlining a Placement Test: Factor Analysis and Rasch Modeling in Practice

Jieun Kim, Maggie McGehee (University of Hawai'i at Mānoa)

This study bridges a research gap by delving into the comprehensive process of revising language placement tests and subsequently evaluating the impact of implementing changes. The focus is particularly on a reading placement test within an ESL program at a US state university. The existing practice of using two tests, Reading Comprehension (RC) and Gap-filling (GF), for a single placement decision raised concerns about practicality and potential construct misalignment.

To address these concerns, the study analyzed data from 559 test-takers using factor analysis and Rasch modeling, supplemented by qualitative approaches. The findings highlighted that RC and GF assess distinct abilities, leading to the decision to eliminate GF to streamline the test and improve practicality and validity. Four revision scenarios were developed, and two were selected for their reliability and assessing a single latent variable: reading ability.

The program promptly implemented one of these scenarios in Fall 2023 admissions, and newly collected data is currently under evaluation. Plans are in place to further refine the test for Spring 2024, proposing new cut scores. The second phase aims to compare data from Spring 2024 admissions with the original and Fall 2023 versions to ensure consistent placement decisions and reliability.

While studies often stop at analyzing existing tests, this research proceeds to illustrate the thorough process of revising a local test, from analyzing its existing factor structure and individual

items, to proposing and implementing revised versions, and then evaluating how well the streamlined versions improve practicality and validity while maintaining reliability.

Unveiling learners' perspectives during speaking disfluencies: Building learners' disfluency profiles across various proficiency levels in OPI assessment

Yu (Joyce) Wu (University of Rhode Island), Qiaona Yu (Wake Forest University)

Speech production is a complex cognitive process. While fluency has been widely researched (Segalowitz, 2010), few studies have examined disfluency features (Yan, 2023) and individual factors that led to disfluencies. Rather than taking disfluency as breaks or disruptions in the flow of speech, this study views disfluency as a dynamic, evolving, and individualized process across proficiency levels. It investigates three research questions: (1) How are disfluencies manifested at different proficiency levels? (2) What constructs the complex and dynamic process during disfluencies? (3) What is the relationship between disfluencies and speakers' Willingness to Communicate (WTC) and anxiety across proficiency levels?

Thirty L2 Chinese learners first took an Oral Proficiency Interview (OPI) and synchronously rated their WTC on each topic during the OPI. They then watched video-recorded episodes (N = 265) on their own disfluencies and shared cognitive-emotional perspectives through stimulated recalls. The results showed: 1) the number, length, and type of disfluencies (i.e., silence pause, self-repair) significantly distinguished oral proficiency levels. Qualitative analysis based on Levelt's Model (1999) revealed that disfluencies took sequential but overlapped steps of idea conceptualization, language formulation, and form articulation. 2) Learners' WTC ratings fluctuated within each OPI, with higher WTC ratings related to familiar and interesting topics, and lower WTC associated with unprepared topics, anxiety, lack of vocabulary, etc. 3) Such WTC contours, however, varied across proficiency levels. Low WTC ratings corresponded to disfluencies at the Intermediate but not at the Advanced/Superior levels.

Using ChatGPT as a tool for automated writing evaluation: impact on syntactic and lexical complexity

Bart Deygers (Ghent University), Liisa Buelens (Ghent University), Laura Schildt (Ghent University), Marieke Vanbuel (Ghent University & KU Leuven)

The goal of this study is to determine whether using ChatGPT as a AWE tool in first-year EFL writing classes at a university can positively impact writing products after an 8-week intervention. The study utilizes a pre-post experimental design in which the experimental group and the control group (n=36) receive in-class instruction and teacher feedback written assignments. In addition the experimental group (n=61) had access to ChatGPT during the in-class writing process after having gone through an online learning path on how to use ChatGPT as an AWE tool. All students wrote a baseline essay prior to the 8-week intervention, two essays during the intervention (the experimental group used ChatGPT) and one essay after the intervention (no group used ChatGPT). In addition, pre and post vocabulary levels tests were administered. The essays were analyzed for syntactic and lexical complexity using T-Scan. The primary analysis will consist of stepwise multilevel linear regression with the abovementioned measures as outcome

variables and with time as the independent variable with random slope. Interaction effects will be calculated between time and condition to examine the influence of the experimental condition on the outcome variables, controlling for student background variables. Based on the existing AWE literature, we hypothesize that the experimental group will show significant improvement in all measures compared to the control group, while using ChatGPT. We do not expect to see large effects in the final writing product, when the experimental group no longer has access to ChatGPT.

Virtual Administration of an Oral English Proficiency Test: Procedures, Challenges and Student Perceptions

Sharareh Taghizadeh Vahed (Center for Applied Linguistics -CAL)

When it comes to International Teaching Assistant (ITA) certification, oral English skills are usually the focus of assessment as the concern with ITAs' direct communication with undergraduates mostly involves the subskill of speaking. The assessment of ITAs' oral English skills is usually done in-person through a local language testing program at each specific institution, whether the assessment method is direct (i.e., in the form of an oral proficiency interview), or semi-direct (i.e., delivered by a computer in a test center). In the context of school and test center shutdowns during the COVID-19 pandemic, language testers and university officials encountered difficulties in assessing ITAs' oral English proficiency for ITA certification purposes. Local language testing programs, with limited staff and budget, needed to quickly come up with a solution to meet their institutions' need to assess their ITAs to either certify them or place them in post-entry language support programs. The poster demonstrates how a large public university in the Midwest was able to transfer their computer-delivered oral English proficiency test to an online test invigilation platform, what challenges were encountered, and what test-takers' perceptions were regarding the online administration of a speaking test that was originally developed for in-person administration. The presenter will discuss the logistics of partnering with online invigilation providers, troubleshooting and quality control procedures, and the local testing program's role in administering an oral English test virtually. The presenter will also discuss why their local testing program decided to continue administering their oral English proficiency test virtually post-pandemic.

A world ready for you, created by Cambridge

We believe that English can unlock a lifetime of experiences and, together with teachers and our partners, we help people to learn and confidently prove their skills to the world.

5.5m
assessments taken every year.

25,000+
organisations accept our exams worldwide.

2,800
exam centres in 130 countries.

50,000
preparation centres in more than 130 countries.

3m+
teachers and learners use Cambridge One for digital learning.

cambridge.org/english

CAMBRIDGE

Where your world grows

Innovative Assessments That Transform Language Learning

- Academic English Proficiency Assessment (AEPA™)
- Academic Reading Comprehension Assessment (ARCA™)
- Foreign Language Proficiency Certifications (FLPC™)

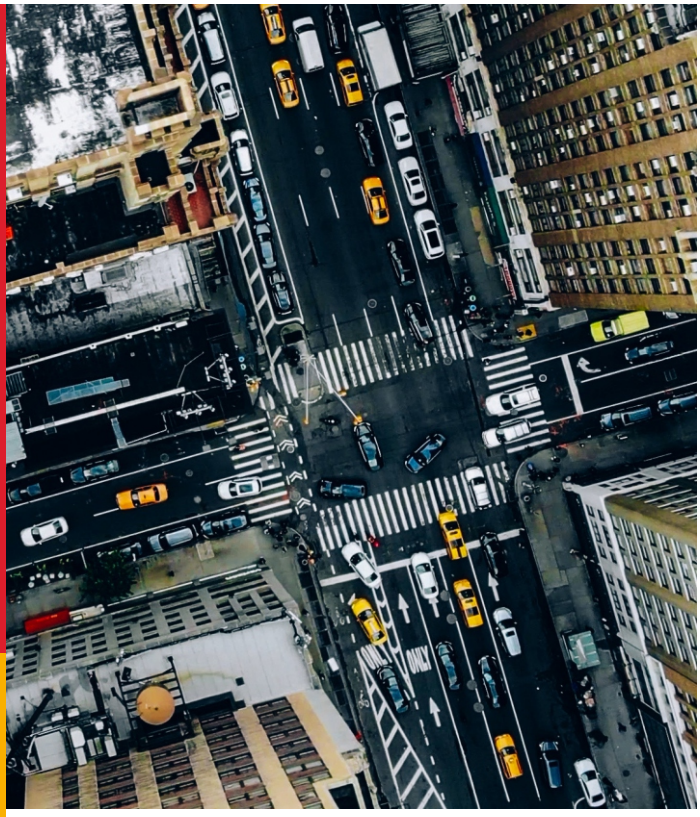
languageassessment.uchicago.edu

IELTS

Research Grants

Interested in conducting an applied research project related to IELTS? Consider applying for the IELTS joint-funded research programme. Successful individuals and educational institutions will be supported with a grant of up to £45,000/AU$80,000.

Scan to find out more!



BRITISH COUNCIL idp CAMBRIDGE English

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

---

Advance your career as a language testing professional

Masters in Language Testing

Lancaster University

By distance learning | Part-time | 2 Years

Scan here to find out more

Apply NOW for October 2024 start

Scholarships available
Attractive new fee structure
Learn about other courses in Language Testing

www.lancaster.ac.uk/linguistics/masters-level
Contact: postgraduatelinguistics@lancaster.ac.uk

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.**

Apply Now!

Contact widainternships@wcer.wisc.edu

WIDA™
UNIVERSITY OF WISCONSIN–MADISON

wida.wisc.edu

LTRC 2025

Language Assessment in Multicultural Contexts: West meets East

Chulalongkorn University

Bangkok, Thailand

June 4-8, 2025

ILTA
INTERNATIONAL LANGUAGE TESTING ASSOCIATION

HOSTED BY

สถาบันภาษา
LANGUAGE INSTITUTE
Chulalongkorn University

ENGLISH AS AN INTERNATIONAL LANGUAGE
Chulalongkorn University

INDIANA UNIVERSITY

More than five decades of research supports the validity of the **TOEFL® Family of Assessments**

Choi, J. S. & Loewen, S. (2022). Exploring Young Learners' Strategic Behaviors in a Speaking Test. *TESOL Quarterly,* 56(4), 1384–1396. **https://doi.org/10.1002/tesq.3136**

Hsieh, C.-N. (2023). The Role of task types and reading proficiency on young English as a foreign language learners' writing performances. *TESOL Quarterly.* **https://doi.org/10.1002/tesq.3286**

Hui, B., Wong, S. S. Y., & Au, R. K. C. (2022). Reading aloud listening test items to young learners: Attention, item understanding, and test performance. *System*, 108, 102831. **https://doi.org/10.1016/j.system.2022.102831**

Kim, M., Nam, Y., & Crossley, S. (2022). Roles of working memory, syllogistic inferencing ability, and linguistic knowledge on second language listening comprehension for passages of different lengths. *Language Testing*, 39(4), 593–617. **https://doi.org/10.1177/02655322211060076**

Papageorgiou, S., & Manna, V. F. (Eds.) (2023). *Meaningful language test scores: Research to enhance score interpretation.* John Benjamins. **https://doi.org/10.1075/illa.1**

Roever, C. & Ikeda, N. (2023). The Relationship Between L2 Interactional Competence and Proficiency. *Applied Linguistics*, pp. 23-. **https://doi.org/10.1093/applin/amad053**

Cushing, S. T., Ren, H., & Tan, Y. (2024). *The use of TOEFL iBT in admissions decisions: Stakeholder perceptions of policies and practices* (TOEFL Research Report No. RR-101). ETS. **https://doi.org/10.1002/ets2.12375**

Suhan, M., Papageorgiou, S., & Wolf, M. K. (2024). *Mapping the scores of the TOEFL Primary® Writing test to the Common European Framework of Reference levels* (Research Memorandum No. RM-24-03). ETS. **https://www.ets.org/Media/Research/pdf/RM-24-03.pdf**

Suhan, M., Papageorgiou, S., Davis, L., & Palmer, M. (2024). *Mapping the TOEFL ITP® Speaking scores to the levels of the Common European Framework of Reference* (Research Memorandum No. RM-24-04). ETS. **https://www.ets.org/Media/Research/pdf/RM-24-04.pdf**

Davis, L., & Norris, J. M. (2023). *A comparison of two TOEFL® writing tasks* (Research Memorandum No. RM-23-06). ETS. **https://www.ets.org/Media/Research/pdf/RM-23-06.pdf**

Gu, L., Li, S., Li, T., & Norris, J. M. (2023). *Maintaining score quality on the enhanced TOEFL iBT® test* (Research Memorandum No. RM-23-05). ETS. **https://www.ets.org/Media/Research/pdf/RM-23-05.pdf**

Wolf, M. K., Suhan, M., Ginsburgh, M., Futagi, Y., & Li, F. (2024). *Design framework for the TOEFL Primary® Writing test* (Research Memorandum No. RM-24-02). ETS. **https://www.ets.org/Media/Research/pdf/RM-24-02.pdf**

https://www.ets.org/toefl/research/

ETS TOEFL®