

ALTAANZ Conference 2025



Balancing innovation and tradition in language assessment

Online, 11-13 November 2025



Contents

WELCOME	3
ABOUT ALTAANZ	4
WITH THANKS TO	5
PRESENTATION TYPES	6
CONFERENCE SCHEDULE	7
DAY ONE: TUESDAY 11 NOVEMBER	8
DAY TWO: WEDNESDAY 12 NOVEMBER	10
DAY THREE: THURSDAY 13 NOVEMBER.....	12
PLENARY ONE: TIM MCNAMARA LECTURE	14
PLENARY TWO	15
ROUNDTABLE ONE	16
ROUNDTABLE TWO	17
ROUNDTABLE THREE	18
STUDENT NETWORKING SESSION	19
MENTOR-MENTEE PROGRAM	20
ABSTRACTS (ALPHABETICAL LISTING BY FIRST AUTHOR'S SURNAME)	21

Welcome

We are excited to welcome you to the 2025 ALTAANZ online conference. With nearly 190 people from around 30 countries registered for the conference (at the time of writing this), the conference will have attracted nearly twice as many attendees as the 2023 event. Attendees and presenters are spread across many time zones and looking at the program, we are in for a stimulating event.

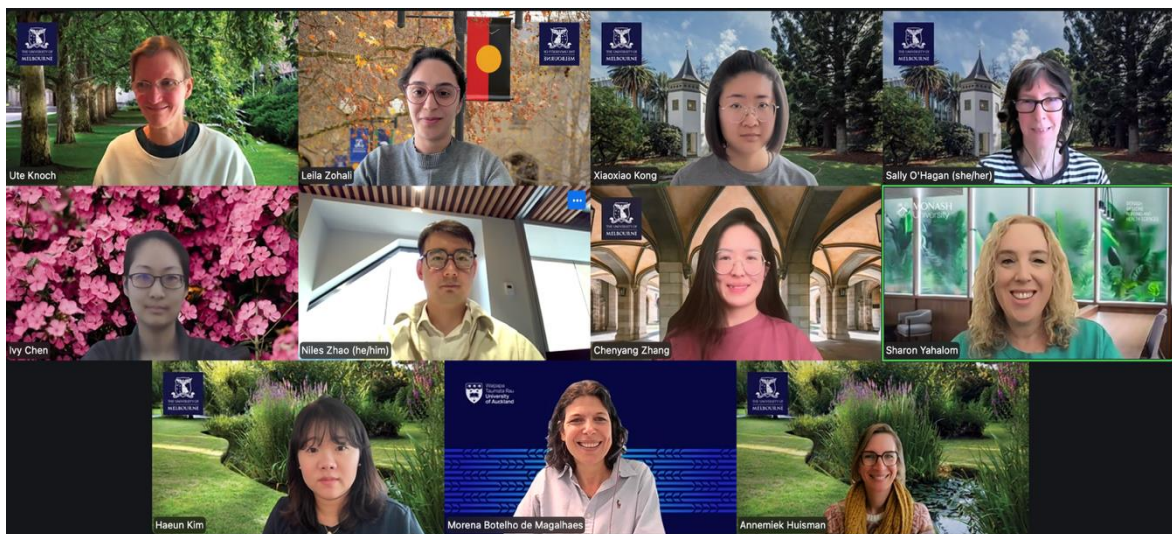
After the success of the 2023 online event, ALTAANZ made the decision to continue with online only conferences held every two years. We are hoping that this allows more people to participate in the conference. To support this, the conference offers relatively low registration fees, particularly to students (and low-waged or non-waged colleagues). Coupled with free membership and the only diamond open access journal in our field, ALTAANZ continues to strive to be accessible to as many people as possible. We are encouraged to see the number of students taking part in the conference – the future of language assessment in our region is bright.

Our theme — *Balancing innovation and tradition in language assessment* — is a reflection of the realities many language assessment researchers and practitioners are currently engaging with. Due to advances in technology, our theoretical models and assessment constructs are being challenged. The 2025 conference showcases the work of practitioners and researchers who are dealing with this reality in their daily work. Our two plenary speakers, Xiaoming Xi (who is presenting the Tim McNamara keynote lecture) and Luke Harding both focus on this theme, as do several of our roundtables. We are particularly delighted by our new collaboration with University English Centres Australia (UECA). Our partners from UECA will be presenting their excellent teacher practitioner research on Thursday afternoon, starting at 1.30pm AEDT. You can find more information about UECA here: <https://ueca.edu.au/>

Our student committee has also been active in the lead-up to the conference and is organising two exciting events for attendees— a student-led session (scheduled on Wednesday at 3pm AEDT) as well as a mentor-mentee scheme. The mentoring scheme will be held for the first time during the conference, and we encourage you to sign up as either a mentor or mentee. You can sign up to this scheme until 5 November at the following link: https://docs.google.com/forms/d/e/1FAIpQLSeLY2dpzmqWTmWy-CB1aXLbVHbY_EeG6T9FM7gMWJ5JY4vDBw/viewform

We hope that the three days will be stimulating and enjoyable for everyone. Please use our social media to share your thoughts during the conference: [Facebook](#) or [LinkedIn](#); Conference hashtag: **#altaanz2025**

The ALTAANZ 2025 Conference Committee



About ALTAANZ

The purpose of the *Association for Language Testing and Assessment of Australia and New Zealand* (ALTAANZ) is to promote best practice in language assessment in educational and professional settings in these two countries and to foster collaboration between academia, schools and other agencies responsible for language testing or assessment. Its goals are listed under three broad headings below:

Training

Stimulate professional growth and best practice in language testing and assessment through workshops and conferences.

Research

Promote research in language testing and assessment through seminars, conferences and/or publications (ALTAANZ publishes a web-based journal and a newsletter).

Policy formation/advice

Provide advice on assessment to public and other relevant agencies on assessment-related issues, and advocate on behalf of test-takers, students and other stakeholders whose life chances may be affected by assessment-related decisions.

For further information about the organisation, please visit the website at: <http://www.altaanz.org>

Membership

ALTAANZ aims to be inclusive and membership of the association is free. To become a member of ALTAANZ, please complete the online membership form on our website: <https://www.altaanz.org/join-altaanz.html>.

With thanks to...

Conference Organising Committee

Ute Knoch (Co-chair)

Jason Fan (Co-chair)

Ivy Chen

Morena Dias Botelho de Magalhães

Rena Gao

Annemiek Huisman

Haeun Kim

Xiaoxiao Kong

Susy Macqueen

Sally O'Hagan

Jet Tonogbanua

Diep Tran

Sharon Yahalom

Lu Yu

Chenyang Zhang

Niles Zhao

Leila Zohali

Programming

Ute Knoch

Jason Fan

Artwork

Ivy Chen

Jet Tonogbanua

Best Student Presentation Award Committee

John Read (Chair), Jason Fan, Catherine Hudson

Abstract Reviewers

Karen Ashton

Junghyun Baik

Morena Dias Botelho de Magalhães

Mark Dawson-Smith

Martin East

Rosemary Erlam

Jason Fan

Peter Gu

Catherine Hudson

Naoki Ikeda

Noriko Iwashita

Hannah Kim

Ute Knoch

Vincent Liang

Julie Luxton

Lyn May

Tracey Millin

Johanna Motteram

Aek Phakiti

John Read

Carsten Roever

Angela Scarino

Kyle Smith

Maria Treadaway

Xingcheng (Alex) Wang

Sharon Yahalom

Lu Yu

Megan Yucel

Chenyang Zhang

Presentation types

Research Papers

Research papers are for sharing developed empirical research or theoretical work (i.e., conceptual papers). Research papers are 20 minutes followed by 5 minutes of discussion. Research papers that are authored and presented solely by students are eligible for the Best Student Paper Award.

Works-in-Progress (WIPs)

A WIP session is an opportunity to share research and seek feedback on research projects or assessment practices that are in development. WiP sessions give the ALTAANZ community a chance to find out about emerging research trends and findings. Sessions are 20 minutes followed by 5 minutes of discussion.

Roundtables

A roundtable is a 60-minute structured discussion on a critical issue to the language assessment community, such as a specific policy area, a particular research concern or an assessment practice.

Conference schedule

All dates and times are in **Australian Eastern Daylight Time (AEDT)**, i.e., Melbourne/Canberra/Sydney time.

Sessions are delivered online live and are not recorded or stored.

Questions and comments are encouraged from the audience using the 'hand up' function or the 'chat' function in zoom.

Thematic blocks
Classroom-based assessment
Scoring: Human or with technology
Policy
Writing
Reading
Listening
Speaking
Vocabulary
Technology
Language for Specific Purposes
Feedback
Tasks
Dynamic assessment
Other

DAY ONE: Tuesday 11 November

Australian Eastern Daylight Time (AEDT)	SESSION A	SESSION B	SESSION C
9.00am-9.25am	Human versus machine: The effectiveness of ChatGPT in automated essay scoring (Research Paper) Jennifer Manning, Jeffrey Baldwin, Natasha Powell	Validation of high-stakes tests and in-house placement tests for assessing international teaching assistants' in-class presentation ability (Research Paper) Okim Kang, Masha Kostromitina, Yuna Bae	From cognitive to affective: Assessing feedback quality and scoring accuracy with a local LLM in a Canadian writing context (Research Paper) Johanathan Woodworth
9.30am-9.55am	Educator perspectives on automated writing scoring for young language learners: Applying a fairness and justice lens (Research Paper) Mark Chapman, Jieun Kim	Analyzing high-quality writing in healthcare: An explainable AI approach (Research Paper) Peter Kim	The development and validation of a diagnostic academic writing assessment for learners of Turkish as a second language (Student Research Paper) Özgü Güntekin
10.00am-10.30am	Break		
10.30am-11.30am	Tim McNamara Keynote Lecture by Xiaoming Xi: Reimagining what we measure: How AI is challenging the tradition of language assessment		
11.30am-12.30pm	Break		
12.30pm-12.55pm	The interplay of test methods, the eyes, and the mind: The curious case of listening comprehension (Research Paper) Tingting Liu, Vahid Aryadoust	Large language models as zero-shot evaluators of English-Chinese interpreting: A comparison of GPT-4o and DeepSeek-R1 (Research Paper) Chao Han	
1.00pm-1.25pm	Developing an AI Persona Perception Inventory (AIPPI) for multimedia-based lecture listening assessment (Work-in-Progress) Ziteng Wang, Vahid Aryadoust	Literacy in the wild: A study of the construct relevance of a health literacy test (Research Paper) Susy Macqueen, Rosalyn Thyer	
1.30pm-1.55pm	How does the listening construct transform under metacognition instruction? An experimental study (Work-in-Progress) Yanyan Huang, Vahid Aryadoust	Examining the relevance of three TOEFL Essentials Writing tasks to the accounting profession: the role of domain experts (Research Paper) Ute Knoch, Jason Fan, Michael Davey, Sally O'Hagan, Ivy Chen, Annemiek Huisman, Rohan Chandran	
2.00pm-3.30pm	Break		
3.30pm-3.55pm	Bridging technological innovation and assessment rigor: Computational validation of LLM-Generated CEFR-aligned reading passages for high-stakes testing (Work-in-progress) Norazha Paiman	Exploring teachers' communication demands in the Australian context: Implications for language assessment for professional registration (Student Research Paper) Xiaoxiao Kong	
4.00pm-4.25pm	Redressing an imbalance - assessing higher-order reading processes in a test of academic reading ability for university entry purposes (Work-in-progress) Stephen Walker	Interrogating the minimum English language standards required for teacher registration in Australia (Research Paper) Ute Knoch, Xiaoxiao Kong, Sally O'Hagan, Ivy Chen	
4.00pm-5.00pm	Break		

5.00pm-5.25pm	Between screens and stress: Investigating affective barriers to online formative assessment in EFL contexts (Student Research Paper) Alireza Maleki	Beyond scoring: DeepSeek-R1 for criterion-specific feedback generation in EFL writing evaluation (Research Paper) Tiancheng Zhang, Shiqi Li	
5.30pm-5.55pm	Teacher's perspectives on ELF assessment: ELF assessment scale construction and validation (Student Research Paper) Freshteh Tadayon	Profiles of university faculty for Generative Artificial Intelligence feedback: a comparative Europe-Australia study (Research Paper) Maria Teresa Mateo Girona, Ana Maria Ducasse, Carmen Lopez Ferrero	
6.00pm-7.00pm	Roundtable: Reconceptualizing the traditional provision of student feedback with AI innovation Peter Davidson, Dan Zhao, Barry O'Sullivan		

DAY TWO: Wednesday 12 November

Australian Eastern Daylight Time (AEDT)	SESSION A	SESSION B	SESSION C
8.30am-9.30am	Roundtable: Assessing oral communication using GenAI Haeun (Hannah) Kim, Ute Knoch, Gary Ockey, Inyoung Na, Gi Jung Kim, Rena Gao, Carsten Roever		
9.30am-9.55am		Demystifying AI-assisted writing process: How do L2 learners with varied levels of AI literacy engage with GenAI tools? (Research Paper) Carrie Peng	Cloze-elide as a formative test of reading (Research Paper) Trevor Holster
10.00am-10.25am	Predicting reading proficiency with vocabulary knowledge: Modalities, frequency, and test lengths (Research Paper) Ji-young Shin, Pablo Robles-García, Jeff Stewart	Investigating the effects of teacher feedback, peer feedback and AI-generated feedback on students' feedback literacy: A classroom-based study (Research Paper) Huijun Zhao, Tianmin Jiang	Wisdom in the counsel of many: Refining cognitive reading attributes through multi-criteria Fuzzy Delphi (Research Paper) Muhamad Firdaus Mohd Noh, Mohd Effendi Ewan Mohd Matore, Nur Ainil Sulaiman
10.30am-10.55am	Predicting multi-word expression density and diversity in a speaking test: Differences across test-taker L1 and English proficiency (Research Paper) Ivy Chen	Promoting ESL students' feedback literacy in an examination-oriented context (Research Paper) Boon Sier (Jeanette) Lim	Unraveling the skill integration in integrated reading-to-speak tasks: The case of L2 Chinese learners (Student Research Paper) Xiaozhu Wang
11.00am-12.00pm	Break		
12.00pm-12.25pm	Assesment of L2 Japanese vocabulary: Building assessment expertise and resources with JFL teachers in Australian secondary schools (Work-in-progress) Fusae Nojima	Exploring test impact in policy space: The impact of Languages Other Than English subjects in the National Matriculation Test in China (Student Research Paper) Chenyang Zhang	Exploring the possibilities of integrating communicative AI into the IELTS test preparation process: The new horizon of human-computer communication (Research Paper) Carlo Perrotta, Sima Mohammadi
12.30pm-12.55pm	The process of revalidating published L2 vocabulary tests for a specific population of learners (Research Paper) John Read, Thi Ngoc Yen Dang, Thi Phuong Dung Cao, Thi My Hang Nguyen	Fitting in at work: Problematizing English testing for skilled migration in Australia (Research Paper) Kellie Frost, Michael Davey	Re-examining test taker agency in AI-mediated language assessment: An ecological approach (Research Paper) Jason Fan, Niles Zhao
1.00pm-1.30pm	Break		
1.30pm-2.30pm	ALTAANZ AGM All members welcome		
2.30pm-3.00pm	Break		
3.00pm-4.00pm	ALTAANZ student event		
4.00pm-5.00pm	Break		

5.00pm-5.25pm	The value of using writing scales in IELTS writing test preparation: An ecological perspective (Work-in-Progress) Yuzhu Su	Towards a novel dual-AI-in-the-loop framework for efficient automatic item generation (Research Paper) Yichen Jia, Vahid Aryadoust	
5.30pm-5.55pm	A collaborative approach to test validation: Using a rating scale to evaluate evidence and guide discussion (Research Paper) Michelle Czajkowski, Bram de Jong	Exploring the importance of academic integrity and cheating prevention in post-pandemic language assessment: Extending the concept of language assessment literacy (Work-in-Progress) Anna Soltyska	Assessment of PFL proficiency: Factors influencing learner performance in different cloze test formats (Work-in-Progress) Clara Setas
6.00pm-7.00pm	Plenary Address by Luke Harding: Imagined interlocutors and 'authentic' chatbots: What can corpus linguistics reveal about constructs in computer-mediated assessment?		

DAY THREE: Thursday 13 November

Australian Eastern Daylight Time (AEDT)	SESSION A	SESSION B	SESSION C
8.30am-8.55am		Writing placement tests for an AI era: A multi-site cross-curricular domain analysis of AI expectations on university writing assignments (Research Paper) Rebecca Yeager, Rurik Tywoniw, Melissa Meisterheim, Ha Ram Kim	Online discourse analysis on comparison between PTE and IELTS among Korean learners (Work-in-progress) Yeachan Choi
9.00am-9.25am	AI-enhanced dynamic assessment of L2 argumentative writing: Designing responsive mediation to diagnose and promote learner writing development (Research Paper) Lu Yu, Matthew E. Poehner, Xiaozheng Dai, Xiaofei Lu, Jingyuan Zhuang	Examining the relevance of the PTE writing tasks to academic writing at university: Insights from test takers (Research Paper) Yangyang Li, Jason Fan, Ivy Chen	Benchmarking scores across tests: Stakeholders perceived test score equivalence for professions and universities across the world (Research Paper) Amanda Muller, Andrew Brenner
9.30am-9.55am	Reconceptualising dynamic assessment validity: A sociocultural theory-based framework with cognitive problem-solving scaffolding (Work-in-progress) Haenga Kim	L2 learner engagement with GenAI in IELTS argumentative writing practice under individual versus collaborative languaging conditions (Student Research Paper) Patrick Guo	When need doesn't equal use: Evaluating the impact of access to extended time in high-stakes testing (Research Paper) Ping-Lin Chuang, Ramsey Cardwell, Will Belzak, Jill Burstein
10.00am-10.30am	Break		
10.30am-10.55am	From ZPD to GPT: Designing AI-supported formative assessments without losing construct validity (Work-in-Progress) JaeYoon Park	Devising a plan for an assessment of spoken language in the university context (Work-in-Progress) Morena Botelho de Magalhaes, Rosemary Erlam	
11.00am-11.25am	Teacher assessment literacy: The relationship between language awareness and writing assessment behaviour (Research Paper) Susanne Stanyer *Winner of Penny McKay Award	Integrating ChatGPT into ESL university students' IELTS speaking practice (Student Research Paper) Mao Sasaki	
11.30am-11.55am	From postgraduate education and beyond: Exploring the TESOL teacher's journey of becoming language assessment literate (Work-in-Progress) Queenie Mak	Evaluating the test construct of functional adequacy in the monologic speech of Japanese learners of English (Research Paper) Rie Koizumi, Masakazu Ueno, Maki Imazawa, Mariko Abe	
12.00pm-1.30pm	Break		

1.30pm-1.55pm	Academic integrity, innovation and university demands: Finding the balance in our new integrated assessments (Teacher practitioner session) Irma Basu, Mohammed Sameer	Enhancing fairness in English proficiency testing: Exploring the impact of IELTS One Skill Retake (OSR) (Research Paper) Hye-won Lee, Reza Tasviri	
2.00pm-2.25pm	From many to one: Balancing innovation with established practice in curriculum-wide rubric design (Teacher practitioner session) John Gardiner	Continuity and change in standardized and situated language assessment practices: Revising the ILTA Guidelines for Practice (Research Paper) Susy Macqueen	
2.30pm-2.55pm	Making meaning visible: Rubrics that capture the full picture (Teacher practitioner session) Kate Randazzo, Amelia Mercieca	The power of testers and their tests: A sociological analysis of assessment practices in Australian Direct Entry Programs (Student Research Paper) Kyle Smith	
3.00pm-5.00pm	Break		
5.00pm-5.25pm	Contingent responses and interaction patterns among students of different English proficiency levels in oral test group discussions (Research Paper) Li Liu, Yin Jiamin	Assessing reading comprehension of young Indian learners through a multilingual multimodal design: A study (Work-in-Progress) Lina Mukhopadhyay	Assessing the economic value of official language proficiency: Evidence from South Korea's labor market (Student Research Paper) Junghyun Baik
5.30pm-5.55pm	Building the foundations for automatic assessment of verbal and nonverbal aspects of spoken interaction in Finnish as a second language (Work-in-Progress) Riikka Ullakonoja, Ilona Lähteenmäki, Nora Raud, Nhan Phan, Tamás Grósz, Henna Suuronen, Raili Hilden, Mikko Kurimo, Mikko Kuronen, Anna von Zansen & Maria Kautonen	Exploring the diagnostic potential of AI-based chatbot interactions as part of L2 English reading computerised dynamic assessment (Research Paper) Dmitri Leontjev, Ari Huhta	Collaboration with policymakers (Student Research Paper) Laura Schildt
6.00pm-7.00pm	Roundtable: Airing and sharing: Responses to challenges for English language learner assessment in school policy in Australia, New Zealand, United States and England/UK Catherine Hudson, Denise Angelo, Julie Luxton, Sue Creagh, Rosalie Grant, Susy Macqueen		
7.00pm-7.10pm	Closing		

Plenary One: Tim McNamara Lecture

Reimagining what we measure: How AI is challenging the tradition of language assessment

Xiaoming Xi, Hong Kong Examinations and Assessment Authority

10.30 am-11.30am (AEDT), Tuesday 11 November

Abstract: For decades, technology has reshaped "how" we measure language proficiency, from computer-based testing to AI systems that simulate human written or speech interaction. Yet the widespread use of generative AI does more than impact assessment methods; it exposes a mismatch between "what" traditional language assessments measure and "what" matters in real-world communication. Today, language users increasingly develop, edit and refine writing and speaking outputs with the help of AI, blurring the traditional boundaries of language proficiency. In this talk, I argue that AI is not merely changing "how" we assess, but fundamentally challenging "what" we value as communicative competence. I will explore new, emerging dimensions of communicative competence that reflect how communication occurs in an AI-mediated world. I will also discuss key conceptual and practical challenges associated with giving language users access to generative AI tools during assessment and a potential path forward.



Xiaoming Xi is Director at the Hong Kong Examinations and Assessment Authority, leading the Assessment Technology and Research Division, the Education Assessment Services Division and the International and Professional Examinations Division. Previously she was Executive Director of New Product Development and Senior Director in R&D at ETS. Her research leadership has impacted global large-scale tests such as ETS's TOEFL, TOEIC and higher education tests as well as Hong Kong's tests for college admissions, teacher certification and students' progress monitoring. A strong contributor to the educational assessment community, Xiaoming has been a Council Member of the International Test Commission (ITC) since September 2023. She has also served on the Executive Board of the International Language Testing Association (ILTA), chaired various ILTA award committees,

and currently chairs the ILTA By-Laws Committee. Xiaoming has been on the Editorial Board of several leading assessment journals, and has won multiple awards, including the 2015 Top 25 Women in Higher Education and Beyond, the Sage/ILTA Best Book Award, and the ILTA Best Language Testing Paper Award. Xiaoming has published widely in assessment theories and practices including validity, fairness, construct definition, assessment design, human and automated scoring, and AI technology. She has guest edited two AI-related special issues "Automated Scoring and Feedback Systems" and "Advancing Language Assessment with AI and ML" for Language Testing and Language Assessment Quarterly respectively and has multiple patents in applications of AI.

Plenary Two

Imagined interlocutors and “authentic” chatbots: What can corpus linguistics reveal about constructs in speaking assessment?

Luke Harding, Lancaster University

06.00 pm-07.00pm (AEDT), Wednesday 12 November

Abstract: The field of language testing and assessment is at a critical juncture. While technology-driven innovations seek to improve efficiencies in development, administration and scoring processes, the constraints of that same technology might limit our ability to capture rich and complex constructs, particularly in speaking assessment. There remains, however, much to understand about the nature of speaking across a wide variety of computer-mediated assessment formats, both with respect to how humans interact with computers, but also the extent to which AI-powered conversational agents can effectively mimic human communicative performance. In this talk, I will draw on work conducted collaboratively with colleagues at Lancaster University to demonstrate how corpus linguistics can provide insights into the speaking construct captured in computer-mediated speaking assessment. I will argue that corpus-based approaches are uniquely positioned to identify stable patterns of recurring features across large datasets, leading to robust inferences about what aspects of spoken performance can (and cannot) be elicited in different speaking assessment environments. I will focus on two specific research projects that illustrate different applications of corpus-based methods: (1) an exploration of pragmatic and interactional features in the British Council-Lancaster Aptis Corpus, a dataset of over one-million words drawn from test-taker responses on a semi-direct computer mediated speaking assessment; and (2) an evaluation of the authenticity of a ChatGPT-powered chatbot, in comparison with spoken production in a general target language use domain, in the context of a low-stakes, formative assessment tool. In the final part of the talk, I will argue that familiarity with corpus methods will be an important component within a wider repertoire of skills needed by professional language assessment developers and researchers as digital technology, and generative AI in particular, continue to influence practices.



*Luke Harding is a Professor of Applied Linguistics at Lancaster University (UK). His research interests are in language assessment and applied linguistics more broadly, particularly assessing listening and speaking, World Englishes and English as a Lingua Franca, language assessment literacy, and the use of digital technology in language assessment. Luke served as co-editor of the journal *Language Testing* from 2017-2022 and is co-editor (with Glenn Fulcher) of the *Routledge Handbook of Language Testing (2nd Edition)*. Luke is currently President of the International Language Testing Association (ILTA).*

Roundtable One

Reconceptualizing the traditional provision of student feedback with AI innovation

Peter Davidson, Dan Zhao, Barry O'Sullivan

6.00pm- 7.00pm (AEDT), Tuesday 11 November

Abstract: This roundtable explores how AI is reshaping EFL academic writing assessment through personalized, corrective feedback. Speakers will examine the benefits and challenges of integrating AI into feedback processes, from prompt design and student engagement to learner trust and large-scale assessment. The session highlights how AI is transforming teacher roles, feedback strategies, and assessment constructs. Practical and ethical concerns will be addressed, with insights from educators and researchers on how to balance AI-driven innovation with meaningful human involvement in language learning.

Roundtable Two

Assessing oral communication using GenAI

Haeun (Hannah) Kim, Gary Ockey, Inyoung Na, Gi Jung Kim, Rena Gao, Carsten Roever

8.30 am- 9.30 am (AEDT), Wednesday 12 November

Abstract: Pragmatic and interactional competence are integral components of L2 communicative competence, yet they remain difficult to assess systematically (Roever, 2022). Assessment of these competencies requires interactive tasks (e.g., role plays, paired discussions), which provide the necessary context and input to elicit ratable discourse. However, such tasks pose challenges for validity due to interlocutor variability (e.g., personality, proficiency) and the co-constructed nature of interaction, which make it difficult to isolate and score an individual's performance reliably. With these challenges in mind, the series of talks in this session will begin by proposing directions for using generative AI to assess pragmatic and interactional competence. The opening talk will be followed by three studies examining various applications of generative AI as a standardized interlocutor, feedback provider, and rater in oral communication assessments.

Roundtable Three

Airing and sharing: Responses to challenges for English language learner assessment in school policy in Australia, New Zealand, United States and England/UK

Catherine Hudson, Denise Angelo, Julie Luxton, Sue Creagh, Rosalie Grant, Susy Macqueen

6.00 pm- 7.00 pm (AEDT), Thursday 13 November

Abstract: This (Roundtable discussion) explores current concerns at the forefront of assessment policy for school-aged English language learners (ELLs) in four English speaking majority countries. A climate of continual educational policy change affects each jurisdiction. However, assessment of and for this cohort plays out differentially according to previous policy/legislative moves along with educational agendas of the day. Each speaker will address current processes and tools for identifying the full ELL cohort and for classifying their level of English L2 proficiency along with sticking points. In addition, speakers will describe systemic accountability measures that follow from ELL assessment, such as data collection, funding allocation and cut off, requisite teaching interventions, performance disaggregations, and student accommodations. Presentations will address research evidence for the usefulness of ELL assessment tools in disaggregating national and local achievement data and guiding policy responses to support successful ELL student learning (e.g. among others Creagh, 2014; Strand & Lindorff, 2020). Through this transnational exchange, we will realise global synergies that can inform interventions in the “discursive web” of policy making (Goldberg, 2006; Elder, 2021).

Student Networking Session

ALTAANZ student session: Research, careers, and connections

Organisers: Xuehua Fu, Xiaoxiao Kong, Chenyang Zhang, Dan Zhou

3.00 pm- 4.00 pm (AEDT), Wednesday 12 November

This session is designed to provide a space for students to connect with one another, share their research interests, and explore career pathways. The format will be interactive and student-led, ensuring everyone has the chance to participate and contribute.



Activities

The session will consist of three parallel Zoom breakout discussions.

Students are free to join whichever breakout room best suits their interests or move between them.

Research Interest Sharing: A space to present your current projects and research areas.

1. **Professional Career Exploration:** Conversations about academic and non-academic career pathways and opportunities.
2. **General Interests and Connections:** An open space to share broader personal interests and build connections beyond research.

Organisation

- Each breakout room will be facilitated by a student representative.
- Representatives will help guide the discussion, ensure everyone has the chance to contribute, and keep time.
- Participants are encouraged to bring questions, ideas, or experiences to share.

Mentor-Mentee Program

Mentor-Mentee program: Language testing & career development

Organisers: Xuehua Fu, Xiaoxiao Kong, Chenyang Zhang, Dan Zhou

The Mentor-Mentee Program, organised by the students, connects researchers in language testing to exchange insights and career guidance. This initiative aims to facilitate one-on-one mentoring sessions close to or during the conference, in an informal meeting.

Activities

All registered attendees of the ALTAANZ 2025 Conference are welcome to participate. We encourage colleagues at all career stages to join the program — including graduate students, early career researchers (ECRs), and mid- to later-career professionals. The session will consist of one-on-one mentor-mentee sessions through Zoom during or around the conference date

Organisation

- Mentors and mentees are paired in advance based on shared research or career interests.
- Initial introductions and expectations are established through email communication.
- Following this, pairs arrange a one-on-one mentoring session, scheduled close to or during the conference.
- Each session lasts no more than 30 minutes and is conducted via Zoom.



Abstracts (alphabetical listing by first author's surname)

Assessing the Economic Value of Official Language Proficiency: Evidence from South Korea's Labor Market

Author(s): Junghyun Baik

Key words: language assessment, practical validity, curriculum design, literacy development, language economics

Abstract: This study investigates the wage returns to official language proficiency in South Korea's labor market, focusing on both linear and nonlinear patterns and their implications for language assessment and policy. Using nationally representative data from the Korean Education and Employment Panel Survey 1 (KEEP1; n = 1,159; female = 619, 53.4%), language proficiency is measured by levels 1 - 9 from the Korean as a national language section of the College Scholastic Ability Test (CSAT). The CSAT offers strong content validity as a language assessment instrument, as it evaluates a range of cognitive-linguistic skills, including reading comprehension, critical thinking, vocabulary knowledge, and syntactic competence.

To estimate wage effects, the study employs Weighted Least Squares (WLS) based on the Mincerian wage equation (Mincer, 1974), along with spline functions and semi-parametric kernel regression to detect nonlinearity. A differencing method controls for key confounders such as education and work experience (Yatchew, 2003).

Findings show that Korean proficiency significantly predicts wages. A one-level increase yields a 2.5% premium, and categorized results reveal significant wage gains from Level 4 upward (9.8 - 22.0%). These effects are pronounced among bachelor's degree holders, suggesting that language proficiency serves as a key signal during recruitment processes requiring advanced communication. Nonparametric analysis confirms a nonlinear pattern: wage gains remain minimal at low proficiency, modest at mid-levels, and accelerate sharply at higher levels, indicating the central role of language proficiency as both human capital and a screening mechanism, particularly in early-career wage determination.

The study offers implications for language assessment and curriculum design. It highlights the need to develop tools and programs that strengthen higher-order cognitive language skills, including critical reading and reasoning. It also emphasizes the importance of practical validity through education-industry collaboration. Finally, the results support lifelong literacy policies and resource allocation that foster advanced language development for individuals seeking re-employment.

Devising a plan for an assessment of spoken language in the university context

Author(s): Morena Botelho de Magalhaes, Rosemary Erlam

Key words: PELA, speaking, diagnostic, assessment, university

Abstract: DELNA (Diagnostic English Language Needs Assessment) is the University of Auckland's post-entry language assessment program. Students first take a screening assessment (Elder & von Randow, 2008), and those scoring below a minimum satisfactory standard on this measure complete a diagnostic assessment comprising listening, reading and writing tasks. An assessment of spoken language has never been included as a survey of staff and students found "that students needed to use speaking less in their studies than the other three skills and it was rated the least critical skill in determining academic success" (Read, 2015, pp. 50-51). The costs around developing a speaking assessment were also determinant in the decision not to incorporate speaking into DELNA. However, recent anecdotal feedback from faculty representatives suggests that spoken communication skills are negatively impacting on students' engagement with academic content. With the widespread use of Generative AI and various writing enhancement software, students may demonstrate satisfactory levels of academic English writing in their work, but not necessarily engage meaningfully with content if spoken communication hinders participation. Assessing students' oral language skills as they begin tertiary study is now a pressing matter. DELNA is therefore planning to investigate the feasibility of incorporating a speaking component into the assessment program. As a first step, faculties will be consulted to establish what aspects of spoken communication should be targeted for the assessment information to be useful. The information obtained will be important in establishing a way forward, and in determining whether an assessment that is already available might be suitable or whether developing one or more (faculty specific) assessments might be necessary. At the time this project is presented, it is expected that such consultation will have occurred and that a more detailed plan will have been devised. Progress to date will then be discussed.

Educator perspectives on automated writing scoring for young language learners: Applying a fairness and justice lens

Author(s): Mark Chapman, Jieun Kim

Key words: writing, technology, K-12, educators, fairness

Abstract: Language educators in K-12 education contexts are on the front lines of the battle between tradition and innovation in both classroom and assessment practices. While technologies like automated writing scoring and feedback (AWSF) encourage adoption, their implementation in K-12 settings has been gradual, as some educators remain hesitant. Accordingly, research on AWSF has largely focused on postsecondary contexts, with limited exploration of use for younger learners (Huawei & Aryadoust, 2023). Addressing this gap, this two-phase study investigates the perceptions of educators in K-12 public schools in the United States.

First, focus group interviews with 14 educators explored their perceptions of the benefits and risks of AWSF for young language learners and identified 15 emerging themes. These themes were reviewed based on Kunnan's (2018) fairness and justice framework, establishing three categories: consistency and meaningfulness, bias and accessibility, and washback. Educators expressed optimism about AWSF's reliability in addressing scoring criteria and ability to differentiate between writing performance levels. However, some raised concerns that AWSF might overly penalize mechanical errors. Additional concerns emerged regarding potential biases based on learners' linguistic backgrounds and the accessibility of AWSF for students with disabilities. Educators were hopeful about washback, particularly the potential to provide immediate feedback in students' first language.

Subsequently, 739 educators from 32 U.S. states responded to a survey developed based on the focus group findings. The survey results indicated that language educators generally viewed AWSF positively for its ability to enhance scoring efficiency, score student writing reliably, and provide feedback. However, they expressed reservations about appropriateness for students with disabilities. Many educators were also concerned about reduced teacher-student interaction and emphasized the need to combine AWSF with human evaluation.

We conclude with educators' suggestions for addressing AWSF challenges, including the need for thorough piloting, transparency about how scores are generated, and maintaining human interaction.

Predicting multi-word expression density and diversity in a speaking test: Differences across test-taker L1 and English proficiency

Author(s): Ivy Chen

Key words: multi-word phrases, collocations, speaking test, corpus-based theoretical models

Abstract: Multi-word expressions (MWEs) are ubiquitous and naturally occurring in both written and spoken language. It is well-documented that they aid processing, yet learners find MWE acquisition difficult. Current research on MWEs has mostly focused solely on MWE frequency, with no systematic comparison across MWE groups or between learner L1s, while mixed results have been found for the effect of proficiency on MWE use. To explain these mixed results, Chen and Kanzawa (2025) proposed a model predicting MWE group difficulty and a model of MWE phrase density (changes due to level of acquisition), which, taken together, accurately predicted phrase density differences across four proficiency levels of a speaking test of English (test-taker L1 Japanese) for 12 MWE groups, six lexical (e.g., verb-noun) and six grammatical (e.g., adjective-preposition).

In this paper, the aim was to validate the use of these two models for predicting MWE phrase density (i.e., counting all instances of MWEs produced, even when the same MWEs are repeated) and phrase diversity (i.e., focusing on range and counting each different MWE once, ignoring repetitions). Ten MWE groups were included, again half lexical and half grammatical. A different speaking test of English was used (Aptis), with three different test-taker L1s (Arabic, Chinese, Spanish) and a different and wider range of English proficiency levels. MWEs were extracted using collocations with a span of +/-3 (while the previous study used bigrams). Individual differences were included in this study to check the usefulness of the predictions for rating purposes.

Findings showed that phrase density and diversity patterns across proficiency and L1 were similar, and that the models predicted these patterns quite well. This is most obvious with MWE groups that were predicted to differ in difficulty across L1s. Implications are discussed in relation to MWE acquisition in general and for rating Aptis speaking performances.

Online Discourse Analysis on comparison between PTE and IELTS among Korean learners

Author(s): Yeachan Choi

Key words: online discourse analysis, Korean learners of English, PTE, IELTS

Abstract: PTE and IELTS are two most representative English examinations that are required for many people outside of Australia and New Zealand seeking them as their destination for study-abroad or immigration. In the digital age, learners often first turn to online space to explore their options between the two examinations before any serious investment into preparing for either one. In this regard, analyzing what is shared in online venues such as YouTube and blogs would provide valuable insights as to how these exams are perceived among particular population. Given the author's background as Korean, this study aims to present a rich description of the online discourse which compares the PTE and IELTS among Korean learners of English. Specifically, the study collects thirty YouTube videos, five user comments per video, and thirty blog posts in Naver (a popular Korean search engine) to derive findings using thematic analysis. In addition to an exploratory, basic description of those who are involved in the online discourse, the emerging themes are expected to contribute to an in-depth understanding of the concerns that Korean learners bear in their initial stages of preparing for PTE and IELTS. Discussions of the findings in light of some principles of language assessments, including practicality, reliability, validity, authenticity, and washback, will follow.

When Need Doesn't Equal Use: Evaluating the Impact of Access to Extended Time in High-Stakes Testing

Author(s): Ping-Lin Chuang, Ramsey Cardwell, Will Belzak, Jill Burstein

Key words: accessibility and accommodations, high-stakes, English testing, extended time

Abstract: Accommodations in high-stakes standardized tests enhance fairness by reducing disability-related construct-irrelevant variance, supporting a test's cognitive validity (Weir, 2005). Among various accommodations, extended time (ET) is the most common due to its ease of implementation and benefits in addressing various test-taker needs. However, research on accommodations in language testing, particularly large-scale quantitative studies, remains limited (Taylor & Banerjee, 2023). This study examines the impact of ET on performance for a digital, computer-adaptive high-stakes English test. It focuses on (1) differences between test takers with and without disabilities and (2) performance patterns across the practice and official tests.

The experimental design included random assignment to either a control group (standard timed conditions) or an experimental group (50% more time, except for speaking tasks) on an online practice test that simulates the official test. Participants were also asked (optionally) to report their disability conditions. Of the 3966 participants, approximately 17% reported having a disability. Participants who reported having disabilities in the experimental group scored significantly higher than those in the control group. Additionally, the score difference between test takers with and without reported disabilities was slightly smaller in the experimental group compared to the control group. These findings suggest that ET may provide a marginal benefit to test takers with disabilities.

We further tracked participants' performance in the official tests. In contrast to the practice test findings, the score difference between test takers with and without reported disabilities was larger for those who had been in the experimental group than those in the control group. Notably, almost none of the test takers requested ET for the official test. This study underscores the potential of ET in reducing disparities and improving fairness, while highlighting the importance of ensuring access to accommodations in operational testing environments.

A Collaborative Approach to Test Validation: Using a Rating Scale to Evaluate Evidence and Guide Discussion

Author(s): Michelle Czajkowski, Bram de Jong

Key words: stakeholder collaboration, argument-based validation, research methodology, post-entry language tests, rating scale development

Abstract: The argument-based approach to validation (Kane, 2006) is now a standard framework in language assessment (Dursun & Li, 2021; Chapelle & Li, 2021). It provides a procedurally objective process designed to ensure discipline, transparency, and consistency in validity judgments. However, while the framework offers structure, it cannot eliminate subjectivity entirely. Evaluating warrants, assumptions, and supporting evidence inevitably involves interpretation. Different validators (researchers, test developers, other stakeholders) bring different levels of expertise and access to evidence, and disagreements between them can reflect legitimate and informed differences in

perspective. The question, then, is not whether subjectivity exists, but how it can be brought to the surface, managed, and integrated into the validation process.

This small-scale study explored this question as part of the validation of a post-entry language test. After constructing the argument (following Knoch & Elder, 2013), the project lead, an external researcher, and a second expert stakeholder, the test's lead developer, collaborated in evaluating the argument. Using a specially designed rating scale with two criteria (amount of evidence, level of support) both parties independently rated the collected evidence for each warrant. This was followed by a structured discussion focusing on areas of disagreement. In some cases, new evidence or alternative perspectives led to a more complete evaluation. In others, disagreement was rooted in broader understandings of test use or purpose; these differences were acknowledged and documented.

We propose to share this approach, along with reflections on its value and limitations. Our talk will outline the motivation for the study, the development and use of the rating scale, the level and nature of agreement observed, and the negotiation process used to explore differing perspectives. We believe this approach offers a structured yet flexible method that reflects both the discipline and the dialogue at the heart of argument-based validation.

Re-examining test taker agency in AI-mediated language assessment: An ecological approach

Author(s): Jason Fan, Niles Zhao

Key words: test taker agency, AI-mediated assessment, test validity, test fairness

Abstract: AI technology is increasingly being used in language learning and assessment. AI-mediated language assessment, which uses AI in task design, delivery, scoring, and feedback, raises new questions about test taker engagement (Fan & Jin, 2023). Agency is a complex and contested notion which has been conceptualised in various ways across different intellectual traditions (e.g., Ahearn, 2001; Biesta & Tedder, 2007; Priestley et al., 2015). Although agency has been a recurring topic in language teaching/learning and teacher education, it has rarely been mentioned or researched in the field of language assessment. Contemporary validity frameworks, such as the argument-based validation framework, also largely fail to account for test takers' dynamic agentic engagement with their contexts of action and its implications for test validity (Fan, Frost & Zhang, 2023).

In this talk, we draw on agency theory, specifically adopting an ecological perspective to agency (Emirbayer & Mische, 1998), to reexamine test taker agency in the context of AI-mediated language learning and assessment. According to this perspective, agency is an emergent phenomenon that is achieved through the interplay between personal capacities and the environment; it is understood as temporal and relational, consisting of three key components: the iterative component (actors' orientation to the past), the projective component (orientation to imagined futures), and the practical-evaluative component (engagement with the present).

We use test impact as an example to demonstrate the relevance and usefulness of this theoretical lens for understanding and elucidating test taker agency in AI-mediated language assessment. In accordance with the three key components of the ecological agency model, we propose guiding questions that should be considered when investigating test taker agency in such contexts. Moreover, we propose a research agenda and offer practical guidelines aimed at fostering a nuanced understanding of test taker agency and its broader implications for test validity, fairness and justice in AI-mediated language assessment.

The Development and Validation of a Diagnostic Academic Writing Assessment for Learners of Turkish as a Second Language

Author(s): Özgü G Güntekin

Key words: diagnostic writing assessment, academic writing assessment, argument-based validation, mixed-methods research construct validity

Abstract: This study presents the development and validation of a Diagnostic Academic Writing Assessment (DAWA) designed for learners of Turkish as a second language (TSL). Motivated by the need for formative, learning-oriented writing assessments, this research reconceptualises writing assessment through a diagnostic lens. The DAWA was developed to address the absence of standardised academic writing assessments in Turkish and to define, for the first time, the features of academic writing in Turkish-medium higher education.

Drawing on literature in English and Turkish academic writing, the study established a construct model of academic writing in Turkish. This informed the creation of an integrated writing task, a multidimensional diagnostic checklist for scoring, and an automated feedback report. The tool was piloted through an online platform with 92 TSL students and 13 trained raters.

Validation employed an argument-based approach, supported by data from many-facet Rasch measurement (MFRM), qualitative feedback, and perception studies. The results provided evidence for the DAWA's construct validity, scale functionality, and formative potential while highlighting challenges in rater consistency and dimensional clarity.

This paper engages directly with the conference theme by showing how a construct-driven, learning-focused model can be operationalised in a robust empirical framework. The study bridges innovation in diagnostic assessment with established validity theory and offers implications for construct design, scale development, and formative assessment in multilingual and under-researched contexts.

L2 Learner Engagement with GenAI in IELTS Argumentative Writing Practice under Individual versus Collaborative Linguaging Conditions

Author(s): Patrick Guo

Key words: GenAI-assisted L2 writing, learner engagement, depth of processing, collaborative languaging, real-time writing process

Abstract: While previous research has explored student engagement with AI-generated feedback during the composition revision stage, little is known about how second language (L2) learners engage with generative artificial intelligence (GenAI) tools throughout the real-time writing processes, especially under different languaging conditions. This study explores the behavioural and cognitive engagement of Chinese learners of English with ChatGPT to support their IELTS argumentative writing practice under individual versus collaborative languaging conditions. Sixteen university students were tasked with completing an IELTS argumentative essay either independently or in dyads, during which they were instructed to interact with ChatGPT and process its output via individual (thinking aloud) or collaborative (dyadic discussions) oral languaging. Students' behavioural engagement was operationalised as prompt foci and sources of mediation, while the analysis of cognitive engagement employed the concept of depth of processing (DoP). Findings revealed that students in the individual languaging condition exhibited a greater reliance on ChatGPT across various writing issues, while peer discussion was found to supplement interactions with GenAI in the collaborative group. The analysis of participants' DoP suggested that collaborative oral languaging enabled learners to demonstrate deeper cognitive processing in significantly more languaging episodes compared to those who individually engaged with ChatGPT. This study offers insights for the joint employment of GenAI chatbots and collaborative languaging in L2 writing classrooms.

Large language models as zero-shot evaluators of English-Chinese interpreting: A comparison of GPT-4o and DeepSeek-R1

Author(s): Chao Han

Key words: large language model, zero-shot evaluator, interpreting assessment, automatic assessment, spoken-language interpreting

Abstract: Assessing translation and interpreting (T&I) is essential in tertiary-level T&I education, professional certification, and foreign language testing. Recently, researchers have explored automating T&I assessment, with large language models (LLMs) emerging as a promising agent for automatic scoring.

This study presents one of the first large-scale empirical investigations into the scoring reliability, severity, and validity of GPT-4o and DeepSeek-R1 in English-Chinese consecutive and simultaneous interpreting assessment.

Using over 500 pre-scored samples from the Interpreting Quality Evaluation Corpus (IQEC), the study configured eight e-raters per LLM, systematically varying three scoring parameters: reference availability (zero vs four references), scoring granularity (segment vs document-level scoring), and model randomness (temperature 0 vs 1).

A series of statistical analyses, including intraclass correlation coefficients, linear mixed modeling, and many-facet Rasch measurement, yielded three main findings

- Regarding scoring reliability: LLM-based e-raters exhibited high internal self-consistency and outperformed human raters in generating more reliable scores overall.
- As for scoring severity: Compared to human raters, GPT-4o displayed similar levels of severity, whereas DeepSeek-R1 tended to underestimate interpreting quality. Overall, GPT-4o was significantly more lenient than DeepSeek-R1, with scoring severity influenced by interpreting direction, mode, reference availability, scoring granularity, and assessment criteria.
- In terms of scoring validity: Both LLMs demonstrated moderately strong correlations with human raters, though variations were observed across interpreting directions and assessment criteria. GPT-4o achieved higher scoring accuracy than DeepSeek-R1, with interpreting direction, mode, reference availability, scoring granularity, and assessment criteria affecting LLMs' scoring accuracy.

We believe that these findings offer valuable insights into optimizing LLM-based automatic scoring and have theoretical and practical implications for the automatic assessment of interlingual interpreting and other language performance assessments.

Cloze-elide as a Formative Test of Rauding

Author(s): Trevor Holster

Key words: formative assessment, cloze-elide, Rasch analysis, reading, listening

Abstract: Reading ability is heavily dependent on aural processing, a process termed 'rauding' (Carver, 1993). However, grammar translation is still heavily emphasized in Japanese classrooms (Maruo, 2012), resulting in extremely slow reading and lack of awareness of the importance of expeditious reading. Shadowing of reading texts, where students must listen to an audio recording of a written text, is one technique to encourage expeditious reading and rauding (Nakanishi & Ueda, 2011), but this raises the question of classroom assessment of shadowing and measurement of learning gains.

Traditional reading comprehension tests emphasize recall of information after reading, conflating the reading process with long-term memory. Process-oriented tests assess reading processes during decoding, rewarding automatization and working memory (Koda, 2004). The cloze-elide (CE) format (Davies, 1975), where students must elide redundant words added to a text, provides a process-oriented test of expeditious reading.

This research investigated the suitability of cloze-elide as a process-oriented classroom assessment of shadowed reading in reading classes in a Japanese university. A machine-readable pencil-and-paper CE format was developed, making CE simple and practical to construct, administer, and score. Weekly review tests were administered to 108 students across the 15-week semester. Each test was repeated across two weeks, first as a shadowed test, where students listened and had to identify extraneous words that were not on the audio recording, then as a pure reading test without an audio recording.

Rasch analysis showed that shadowed reading was substantively easier than unshadowed reading, as expected, with a statistically significant difference of 0.78 logits. Person reliability from the 243 test items was .95, but unshadowed items were found to somewhat misfit compared to the shadowed items. Overall, the shadowed CE format was found to be practical and effective as a formative classroom assessment to encourage rauding.

How does the listening construct transform under metacognition instruction? An experimental study

Author(s): Huang Yanyan, Vahid Aryadoust

Key words: listening construct, metacognition instruction, experimental study

Abstract: It is recognized in language assessment that the listening construct can evolve through targeted instruction and the development of interconnected latent traits, such as metacognition. This conceptualization positions the L2 listening construct as dynamic rather than static, with growth in listening proficiency influenced by the advancement of related constructs within the nomothetic span of L2 listening. However, most listening assessment studies conceptualize the listening construct as static and there is a dearth of research on the dynamic nature of it, particularly with regards to the effect of growth in other interconnected processes such as metacognition on the vicissitude of the listening construct. Thus, in an ongoing study we investigated how the listening construct shifts over time under two conditions:

(1) when learners receive metacognitive instruction, and (2) when learners receive no instruction (control group). Additionally, since listening comprehension poses significant challenges for lower-proficiency language learners, an issue especially evident among Chinese primary school English learners, we focused on this population.

We employed an explanatory sequential mixed-methods design, with participants assigned into experimental ($n = 32$) and control groups ($n = 32$). The experimental group received metacognitive instruction in 8 sessions, one session per week over a period of 2 months, while the control group followed traditional listening instruction methods. Data collection utilized multiple instruments including standardized English listening comprehension tests and the metacognitive awareness listening questionnaire (MALQ). Follow-up semi-structured interviews provided more profound insights into participants' experiences and perceptions. To analyze the data, we are using several quantitative techniques. First, a Rasch model analysis will be conducted separately on each scale to ensure the instruments' psychometric validity. The second analysis stage involves quantitative data analysis such as the general linear model and the analysis of the interviews.

This study will contribute to the understanding of the L2 listening construct in young, low-proficiency language learners and its change due to metacognitive instruction as a pedagogical practice. The findings have practical implications for the nature of the L2 listening construct and its dynamic relationships with metacognitive strategies.

Towards A Novel Dual-AI-in-the-Loop Framework for Efficient Automatic Item Generation

Author(s): Jia Yichen, Vahid Aryadoust

Key words: AIG, large language model, item evaluator, finetuning

Abstract: Conventional test development remains resource-intensive, which fails to meet growing demands for internet-based assessments that require large item banks while maintaining test security. Automatic item generation (AIG) offers a promising solution but faces a quality-efficiency tradeoff.

Previous research has shown some success in using AI for AIG, but significant challenges remain. While some studies have demonstrated improvements in item quality through human-AI interaction, the need for constant human involvement limits scalability and may introduce bias. In addition, the proportion of AI-generated items that pass human evaluation has generally been low, highlighting the need for more effective and autonomous approaches.

The present research adopts a novel approach called the dual-AI-in-the-loop, which is introduced by Aryadoust (2025) in an upcoming Routledge book titled 'Assessing Listening in the Age of Artificial Intelligence.' In this approach, AI serves dual roles: generating items and evaluating those generated items. Leveraging this approach, we first go beyond traditional prompting that requires constant human interaction by finetuning ChatGPT-4o using listening test items as samples. Our findings demonstrate that the finetuned model complete script and MCQ generation for one test set within minutes, with characteristics similar to high-quality listening tests across multiple linguistic features. Second, we implement a 'secondary-AI-in-the-loop' system by finetuning an AI model specifically for item evaluation. This evaluator enhances AIG efficiency by filtering items that do not meet quality standards before human review.

This integrated dual-AI system reduces resource requirements while maintaining quality standards. Our method may offer a cost-effective solution for producing in-class and/or large-scale assessment items with minimal post-editing requirements, which addresses the fundamental quality-efficiency tradeoff in AIG. We will report on the performance of both models and discuss implications for test development in the generative AI era.

Validation of High-Stakes Tests and In-House Placement Tests for Assessing International Teaching Assistants' In-Class Presentation Ability

Author(s): Okim Kang, Masha Kostromitina, Yuna Bae

Key words: construct-relevant validity, international teaching assistant, placement test, high-stakes proficiency test, in-house proficiency test

Abstract: Several pieces of evidence need to be considered to validate a language proficiency test (Chapelle, 2020). One essential indication is the extent to which test scores are associated with test-takers' construct-relevant ability in the target language use domain. Given that speaking is of utmost importance for International Teaching Assistants (ITAs) admission and appointment (Dalman & Kang, 2023), ITAs in the U.S. oftentimes undergo oral proficiency scrutiny. While some institutions have created their own in-house assessments, others employed scores in the speaking sections of

high-stakes language tests (Farnsworth, 2013). However, it is largely unknown how ITAs' performance on these tests relates to their classroom performance. Therefore, the current study investigated the extent to which Duolingo English Test (DET) scores related to ITAs' in-class presentation scores as well as ITA programs' placement test scores. Ninety-two ITAs (44 female, 48 male) were recruited from four U.S. research universities with established ITA programs. They represented diverse L1 backgrounds, including Mandarin Chinese, Korean, Hindi, and Persian. The ITAs took an official version of the DET and completed a background survey providing their placement test scores. Nine ITA program instructors from aforementioned institutions released students' in-class presentation evaluations. Teachers also provided ratings using a synthesized rubric, which incorporated aspects of program-specific rubrics. Calculated correlations corrected using the Thorndike's case II method suggested that DET speaking scores were strongly related to the two sets of in-class presentation scores ($r = .78$ for in-house rubrics and $r = .82$ for synthesized rubrics) across ITA groups from all four institutions. Yet, ITAs' in-house placement test scores had varying correlation ranges from $r = .31$ to $r = .71$, implying large variability among institutions. Findings support institutionally standardized approaches for ITA's speaking assessment and offer important implications for general language testing practices as well as stakeholder decision-making and admission policies.

Reconceptualising Dynamic Assessment Validity: A Sociocultural Theory-based Framework with Cognitive Problem-Solving Scaffolding

Author(s): Haenga Kim

Key words: dynamic assessment, validity framework, cognitive problem-solving, sociocultural theory, second language acquisition

Abstract: Dynamic Assessment (DA), grounded in Vygotsky's (1978) Sociocultural Theory (SCT), underscores mediated learning through cooperation within the Zone of Proximal Development (ZPD). While existing DA research has demonstrated its positive impact on Second Language Acquisition (SLA), its systematic validation in relation to the learning process remains underdeveloped, and traditional graduated prompting lacks a structured learning sequence. Therefore, in this study, I propose a twofold theoretical contribution: (1) a novel Cognitive Problem-Solving (CPS)-integrated graduated prompting system and (2) an SCT-based DA validity framework.

The CPS model integrates Zimmerman's (2002) self-regulated learning framework, cognitive skill acquisition theory (VanLehn, 1996) and Cognitive Load Theory (Sweller, 1988) to structure mediation into four graduated steps: problem identification, cognitive processing strategies for resolution, guided problem-solving, and self-regulation for transcendence. Unlike traditional DA scaffolding, which moves from implicit to explicit prompting without structured cognitive development, this model introduces a systematic sequence of cognitive engagement, promoting internalisation and transcendence. This aligns with Krashen's (1985) Input Hypothesis (Interaction), DeKeyser's (2007) Skill Acquisition Theory (Internalisation), and Feuerstein et al.'s (1988) concept of Transcendence, linking scaffolding directly to learning generalisation and long-term skill retention.

To validate the CPS-based DA framework, an SCT-aligned DA validity framework is introduced, extending Kane's (1992, 2006) argument-based validity and Bachman and Palmer's (1996) assessment use argument. The SCT-based DA framework consists of five interconnected validity components: (1) Intention Validity (Bachman and Palmer's construct validity), (2) Interaction Validity (Bachman and Palmer's interactiveness and practicality), (3) Internalisation Validity (Kane's extrapolation inference), (4) Transcendence Validity (Kane's generalization inference), and (5) Independence Validity (Bachman and Palmer's assessment consequences).

By integrating CPS mechanisms to DA framework as pedagogy and assessment, I aim to provide a feasible DA framework for teachers to use in classroom instruction and assessment, supporting the intersection of learning and assessment.

Analyzing High-Quality Writing in Healthcare: An Explainable AI Approach

Author(s): Peter Kim

Key words: explainable AI, healthcare, writing English for specific purposes, automated assessment, linguistic features

Abstract: The rapid advancement of AI is transforming how we understand complex constructs and validate assessment tools. In language testing, AI offers new opportunities to analyze writing quality in a systematic and interpretable way. This study develops an explainable AI framework to identify key linguistic patterns that characterize high-quality writing

among healthcare professionals. Unlike black-box models, this approach integrates linguistic features including lexical diversity, syntactic complexity, readability, and discourse cohesion to enhance the assessment of writing proficiency.

The study examines writing responses from the Occupational English Test (OET), an English language test for healthcare professionals. By analyzing linguistic features such as type-token ratio, sentence length, readability scores and discourse markers, this research contributes to language assessment by refining our understanding of what constitutes high-quality writing in professional healthcare contexts.

Key findings include:

1. Sentence Structure & Complexity ' High-quality writing is more concise, averaging 21.5 words per sentence compared to 30.8 in low-quality writing, suggesting clarity enhances quality.
2. Cohesion & Readability ' Higher cohesion scores (0.64 vs. 0.37) indicate better logical flow, while readability scores were slightly lower.
3. Lexical Diversity & Vocabulary Use ' High-quality writing had slightly lower indices (0.63 vs. 0.66), indicating that lexical diversity is not a strong factor.
4. Transition Words & Connectivity ' Transition words were rare in both types, with high-quality responses using at most one, suggesting healthcare writing differs from academic writing.
5. Grammar & Word Choice ' Greater use of nouns and adjectives enhances descriptions and precision, while more pronoun references improve continuity.
6. Overall Structure ' High-quality writing is longer (239 vs. 206 words) with more sentences (11.9 vs. 8.9), reflecting well-developed arguments.

This research advances AI-driven assessment, offering insights into linguistic expectations in professional healthcare communication. Future work will refine the framework for broader applications in automated evaluation and assessment.

Examining the relevance of three TOEFL Essentials Writing tasks to the accounting profession

Author(s): Ute Knoch, Jason Fan, Michael Davey, Sally O'Hagan, Ivy Chen, Annemiek Huisman

Key words: validation, task relevance, domain experts

Abstract: Large-scale English language proficiency tests are increasingly used to make decisions about professional registration despite not originally being developed to make predictions about language use in the workplace. Domain experts can play a valuable role as informants in establishing the relevance of test tasks to a specific TLU domain. However, this practice has seldom been critically examined.

The current study was designed to (1) explore the relevance of three writing tasks (i.e., Build a Sentence, Write for an Academic Discussion, Write an Email) from the TOEFL Essentials test to the accounting profession and (2) evaluate the judgements of the domain expert participants. Twenty accountants from non-English speaking backgrounds (all recent migrants to Australia or New Zealand) as well as three accounting educators were interviewed for the study. All participants were asked to take the TOEFL Essentials practice test to ensure familiarity with the writing tasks. They were then asked in the interviews whether they thought the tasks resembled writing tasks at work, and to justify their answers. The data was analysed qualitatively to identify (a) to what extent the participants considered the three tasks relevant and (b) what task features they attended to when commenting on the relevance of the tasks.

The findings showed that the participants generally found the Build a Sentence the least relevant of the three writing tasks, and the Write an Email task the most relevant. When reviewing the task aspects the participants judged as relevant to writing demands in their workplace, it was shown the participants focussed on a small/narrow range of task aspects. They also made 'misaligned connections', comparing aspects of test tasks and workplace tasks that did not align (e.g., comparing a writing task to events in a spoken meeting). The findings are discussed in terms of domain expert involvement in test validation research.

Interrogating the minimum English language standards required for teacher registration in Australia

Author(s): Ute Knoch, Xiaoxiao Kong, Sally O'Hagan, Ivy Chen

Key words: standard setting, professional registration, teacher language proficiency, policy, test validation

Abstract: Overseas-trained teachers and international graduates of initial teacher education programs in Australia applying for registration in Australia are required to demonstrate that they have the required level of English language proficiency to effectively communicate in various school settings in Australia. To meet this requirement, applicants can currently take one of two approved tests ' IELTS (Academic) or ISLPR. The IELTS Academic test is designed to test general academic English language proficiency (not specific to any particular domain), while the ISLPR draws on topics related to teaching. The minimum standards currently considered sufficient for teacher registration in Australia are as follows: for the IELTS Academic, a minimum of 8.0 for Speaking and Listening and 7.0 for Reading and Writing, with an overall score of no lower than 7.5; and for the ISLPR, a minimum of 4 for each skill (i.e., speaking, listening, reading, writing) is required. It is not clear how these minimum standards were initially set.

This study set out to establish, empirically, the minimum standard required by involving teachers in a standard-setting study drawing on the analytic judgement method (Plake & Hambleton, 2001). Due to funding constraints, we were only able to focus on speaking and writing in the IELTS Academic test. Fifty-one teacher participants drawn from a range of school sectors and states and territories in Australia were invited to review previously scored IELTS Academic writing and speaking performances. The data was collected and analysed in two phases: range-finding and pin-pointing.

The results of the statistical analysis showed that the current IELTS Academic writing standards are in line with the teacher judgements we collected, however, the speaking requirement of IELTS 8 was considered too high. The results are discussed in relation to the current policy-environment in Australia.

Evaluating the test construct of functional adequacy in the monologic speech of Japanese learners of English

Author(s): Rie Koizumi, Masakazu Ueno, Maki Imazawa, Mariko Abe

Key words: speaking assessment, assessing functional adequacy, Japanese learners of English, monologic speech, rater training

Abstract: This presentation evaluates the measured test construct, focusing on functional adequacy (FA) in the speech of L1 Japanese learners of L2 English. Kuiken and Vedder (2018) proposed that FA complements traditional constructs such as complexity, accuracy, and fluency in language assessment. They introduced an FA rubric assessing four dimensions: content, task requirements, comprehensibility, and coherence and cohesion. Measuring new constructs requires careful definition, administering, scoring, and evaluation of measurement quality and test utility (Chapelle & Voss, 2021). This study investigates the relationships between the FA dimensions and speaking proficiency to examine how the intended construct is measured.

We used monologic data from the Longitudinal Corpus of Spoken English, comprising 122 Japanese secondary students at CEFR proficiency levels Pre-A1 to B2. These students completed 10 tasks in the Telephone Standard Speaking Test (TSST) over eight sessions in 23 months. FA benchmarking scores were established for the learners' speeches using Kuiken and Vedder's (2018) rubric. After rater training, seven raters evaluated the speeches, with one session randomly selected for each learner. The TSST holistic scores served as an indicator of speaking proficiency. The FA ratings were analyzed using many-facet Rasch analysis (Linacre, 2025) and Pearson's product-moment correlations to examine the relationships between variables.

The results revealed that tasks, raters, and criteria functioned as expected. The four criteria fit the Rasch model (infit mean squares of 0.82 - 1.17), indicating their independent contribution to the FA measurement. Strong correlations were found among the FA dimensions' logit scores ($r = .94 - .98$) and between these scores and holistic scores ($r = .74 - .81$). These findings suggest that FA dimensions are relatively strongly associated with each other, with FA being a major component of L2 speaking. Future studies should examine the relationship between FA and other dimensions, such as fluency. Implications for assessment are discussed.

Exploring teachers' communication demands in the Australian context: Implications for language assessment for professional registration

Author(s): Xiaoxiao Kong

Key words: language for specific purposes, language assessment for professional purposes, test validation, professional communication, needs analysis

Abstract: Since 2011, overseas-trained teachers and international graduates in Australia must meet set scores in approved English proficiency tests to register and work within early childhood and school settings (Australian Institute for Teaching and School Leadership, 2011). Despite nationwide implementation of this requirement, the appropriateness of the English language tests for teacher registration remains largely unexplored.

The study investigates (1) the language and communication demands of early childhood, primary school, and secondary school teachers in Australia, and (2) the suitability of the IELTS Academic, the only English language proficiency test approved nationwide, for teacher registration. Guided by an argument-based validation framework (e.g., Knoch & Chapelle, 2018; Knoch & Macqueen, 2020) and needs analysis for language assessment for professional purposes (LAPP; Knoch & Macqueen, 2020; Long, 2005), the study employs a sequential exploratory design, incorporating document analysis, focus groups (12 groups, 37 participants), a survey (n = 123), and interviews (n = 18). These were conducted to examine teachers' language demands and challenges as well as views on the IELTS Academic as a measure of workplace language proficiency. Findings reveal differences in teachers' workplace communication tasks and language skills required across education settings. Compared with school teaching, early childhood contexts exhibited lower degrees of alignment with the IELTS Academic task characteristics, raising concerns over the test's appropriateness across education levels. Additionally, domain insiders highlighted communication challenges faced by all teachers, and by extension, fairness issues regarding administering LAPPs solely to non-English speaking background applicants.

This study represents validation efforts for a high-stakes language assessment for professional purposes and a needs analysis of teachers' workplace language demands in Australia. Findings provide implications for policy and assessment practices amidst the nationwide teacher shortage and increasing immigrant entry through teaching occupations, potentially impacting student outcomes.

Enhancing Fairness in English Proficiency Testing: Exploring the Impact of IELTS One Skill Retake (OSR)

Author(s): Hye-won Lee, Reza Tasviri

Key words: bias for best, test delivery, scoring validity, stakeholder perceptions, test fairness

Abstract: English proficiency tests play a pivotal role in shaping the academic and professional trajectories of test-takers, influencing their prospects in the migration and higher education sectors. As these scores can significantly impact individuals' lives, many test takers opt to retake language proficiency tests multiple times, often in pursuit of specific sub-scores. While recent research has begun to address the challenges around this, a comprehensive understanding of the factors contributing to repeat test-taking and the broader implications for the validity arguments associated with these tests remains limited.

This proposal aims to examine how a guided method, such as IELTS One Skill Retake (OSR), can address these challenges by offering test takers the option to retake only one section of the test within a short time frame. This "bias for best" strategy (Van Moere & Hanlon, 2020) seeks to alleviate construct-irrelevant variables such as performance anxiety and fatigue, while maintaining validity and fairness. The presentation will explore the implications of this approach, using a comprehensive mixed-methods design that integrates over 60,000 test-taker data points, a large-scale questionnaire (n=578), and interviews (n=29) with test-takers and receiving organisations.

The analysis will investigate global trends as well as local ones from select regions, specifically those with a high volume of test-takers. By comparing local and global patterns, the study will explore cultural and social factors that may influence test outcomes. The findings will discuss the effectiveness of OSR, its potential to improve test fairness, and how this innovative method can enhance traditional practices in language assessment. Ultimately, this presentation will offer recommendations for guiding best practices in the field, with a focus on enhancing the effectiveness and fairness of language assessments in an evolving global context.

Exploring the diagnostic potential of AI-based chatbot interactions as part of L2 English reading computerised dynamic assessment

Author(s): Dmitri Leontjev, Ari Huhta

Key words: dynamic assessment, artificial intelligence, mediation, L2 English, reading

Abstract: With the development of generative artificial intelligence, studies exploring the potential of AI to complement and enhance L2 assessment have proliferated. This is evidenced, for example, at the 2025 AAAL conference, where a number of papers reported on AI-based assessment of L2 writing: including scoring (e.g., Yamashita, 2025), feedback, dynamic assessment to develop lexical richness and argumentation (Dai et al., 2025; Xu & Xia 2025), as well as on teachers' perceptions of AI-generated feedback. We report on a part of an ongoing project in Finland, focusing on our exploration of L2 learners' engagement with an AI-based chatbot that yields extra diagnostic information about learners' challenges and changes in thinking processes during L2 computerised dynamic assessment of reading. Rooted in Vygotskian sociocultural theory, dynamic assessment allows for uncovering learner abilities in the process of development (known as Zone of Proximal Development) by including graduated and contingent support (mediation) to the assessment process, provided to learners when they encounter difficulties (Poehner, 2008). We developed a chatbot that is enabled after learners arrive at the correct response, encouraging them to ask questions. In its responses to learner queries, the chatbot takes account of the task, the text, learner responses, and mediation learners receive. We report the results of a study of chatbot functionality, particularly on additional insights into learner areas of struggle, which complements the information about learners' mediated performance that emerged from the learners' engagement with mediation. We focus on two L2 English learners in upper-secondary education and report on (a) their interactions with mediation and the chatbot and (b) their follow-up interviews to illustrate how additional diagnostic information emerged from learners' engagement with the chatbot. We will build an argument for the potential of chatbot as a part of mediation sequences in L2 reading computerised dynamic assessment.

Examining the Relevance of the PTE Writing Tasks to Academic Writing at University: Insights from Test Takers

Author(s): Yangyang Li, Jason Fan, Ivy Chen

Key words: argument-based validation framework, academic writing skills, international students' perspective

Abstract: This study investigated international students' perceptions of the relevance of the Pearson Test of English Academic (hereafter PTE) writing tasks to academic writing at an English-speaking university. It addressed a key concern in language assessment: whether standardized high-stakes language tests adequately represent the academic demands students face in real-world university contexts. The study draws on the argument-based validation framework (e.g., Kane, 2002). Specifically, it focused on three key inferences in the argument-based validation framework, including domain description, extrapolation, and utilization, to examine how well the PTE writing tasks align with academic writing requirements.

Ten international students from diverse disciplines, enrolled at a large Australian public university and who had taken the PTE test within past two years, participated in this study. Data were collected through semi-structured interviews (n = 5) and a focus group discussion (n = 5). Thematic analysis was conducted to analyse the interview and focus group data using NVivo 14. The data were coded into three themes, including a) academic writing tasks and abilities, b) PTE writing tasks' relevance to academic writing, and c) evaluation of PTE preparation strategies.

This study found that the participants identified 5 essential skills for academic writing, including critical thinking, logical structure, accuracy of language use, referencing and citation format, and information extraction and summary. They perceived that PTE writing tasks partially covered basic skills but lacked assessment of higher-order abilities such as critical thinking. Moreover, test preparation strategies, often centred around using templates and memorisation, were perceived to be highly effective in improving PTE scores efficiently but have limited transferability to academic writing at the university.

The findings highlighted a lack of alignment between PTE writing tasks and the writing skills required in their academic writing in the university. This gap suggests the need to reconsider how standardized assessments are designed and used for academic admissions. The study contributes to ongoing discussions on test validity and fairness, offering implications for higher education institutions, test preparation programs, and future test takers.

Promoting ESL Students' Feedback Literacy in an Examination-oriented Context

Author(s): Boon Sier Jeanette Lim

Key words: student feedback literacy, classroom-based assessment system, bio-ecological model

Abstract: In contexts where direct instruction and summative assessment are emphasised, formative assessment is less valued and more difficult to implement (Carless, 2012). This naturalistic study explores the extent to which formative assessment practices embedded in a classroom-based assessment system serve to build learners' feedback literacy. The assessment system is part of an English language bridging programme of a private Malaysian university. It incorporates formative and summative assessment practices like student self-assessment, standardised assessment tasks and assessment feedback. Part of a larger study, this presentation focuses on (1) how students perceived the purposes and processes of the assessment practices comprising an assessment cycle in the system, and (2) how they engaged with the assessment information generated. Data came from semi-structured interviews of six student participants, a questionnaire involving 38 respondents, classroom observations and analysis of assessment-related documents. The findings suggest that the assessment system promoted the student participants' feedback literacy. They perceived the intended purposes, task requirements and usefulness of student self-assessment, a writing assessment and feedback practices that constituted the assessment cycle. They also engaged with the feedback given cognitively, behaviourally and affectively, with evidence of feedback uptake. By using Bronfenbrenner and Morris' (2006) bio-ecological model to map the contextual, individual and time factors that helped shaped the student participants' feedback literacy, this study facilitates further research that seeks to apply or up-scale what has been done to similar or other contexts. (232 words)

The interplay of test methods, the eyes, and the mind: The curious case of listening comprehension

Author(s): Tingting Liu, Vahid Aryadoust

Key words: cognitive process, test method, test validity, item presentation, item format

Abstract: Test methods, specifically item presentation and item format, have been shown to influence test-takers' scores. However, less attention has been given to understanding the cognitive processes underlying test score differences and how listeners adapt their gaze behavior based on different test methods. The present study explored how two different item presentation methods impact the gaze and cognitive processes of listeners: (1) while-listening-performance (WLP), where test items are presented concurrently with the listening tasks, and (2) post-listening-performance (PLP), where items are presented after the listening task is completed. This study also examined the effects of two item formats: multiple-choice questions (MCQs) and open-ended questions (OEs). Using a Graeco-Latin square design, four psychometrically validated short-talk listening tests were administered to a sample of ten participants, and their gaze behaviors were recorded with a Tobii eye-tracker, allowing for a detailed verbal report of their thought processes during the test. The findings revealed notable differences in how test-takers navigate their attention during listening, depending on whether items were presented during or after listening. Additionally, the degree to which test-takers' responses accurately reflected their listening ability varied across item formats. The verbal reports indicated that test-takers often adopted alternative listening or test-wise strategies in response to the test methods, rather than engaging in the cognitive processes the test was designed to measure. These results suggest that test methods can significantly shape listeners' cognitive and gaze processes. Consequently, test scores "though considered psychometrically valid" may not accurately reflect the latent traits that listening tests are intended to measure. The study has implications for authenticity and test design, emphasizing the need to consider the cognitive processes elicited by different test methods when designing listening assessments.

Contingent Responses and Interaction Patterns Among Students of Different English Proficiency Levels in Oral Test Group Discussions

Author(s): Liu Li, Yin Jiamin

Key words: contingent response, interactional competence, language proficiency level, group discussion

Abstract: This study focuses on a key component of oral interaction competence, 'contingent response', defined as the ability of participants to respond appropriately based on the content of the previous speaker. While existing research has provided a theoretical foundation for defining contingent responses, there remains a lack of sufficient empirical evidence regarding students' specific performance in producing such responses in testing contexts. To address this gap,

this study investigates the differences in generating contingent responses among students of varying English proficiency levels during group discussions in a university-based English oral proficiency test. Employing a mixed-methods approach combining quantitative analysis and conversation analysis, the study examines students' performance in terms of response quantity, types, and interaction patterns.

The findings reveal that language proficiency significantly impacts the ability to produce targeted responses: higher-proficiency students generate more responses, exhibit a greater variety of response types, and tend to use collaborative responses such as support, expansion, elaboration, and supplementation, with interaction patterns predominantly characterized by collaboration. In contrast, lower-proficiency students rely more on challenge responses, with interaction patterns dominated by parallel modes and lower levels of mutual assistance. By exploring the relationship between language proficiency and interaction competence, this study provides empirical evidence to inform language teaching and assessment frameworks based on interactive competence. Future research may further examine the construct of interactive competence across different communicative contexts to refine assessment models and offer more targeted guidance for oral language instruction and assessment.

Continuity and change in standardized and situated language assessment practices: Revising the ILTA Guidelines for Practice

Author(s): Susy Macqueen

Key words: guidelines for practice, standardised testing, classroom-based assessment, language assessment literacy, stakeholders

Abstract: The International Language Testing Association (ILTA) Guidelines for Practice is a resource for anyone who designs, provides, evaluates, endorses, uses or undergoes language assessment, including policy makers, teachers, testing agencies and people who are assessed. The aim of the Guidelines is to promote fair, valid and transparent practices in language assessment. In 2024, the Guidelines were revised to keep pace with developments in technology, practice, theory and ethos. This talk gives a brief history of the ILTA Guidelines and provides the context and procedure for the current revisions. The large range of stakeholders, roles and audiences, and the broad scope of contemporary assessment practices necessitated a significant reworking of the existing Guidelines in three broad thematic areas: Trustworthiness, Communications and Fairness and Consistency. These themes apply across the two main sections in the Guidelines: (1) formal, standardized testing and (2) assessment in learning environments. The content of the 2024 Guidelines will be introduced and areas where the scope of practice has changed will be discussed. This significant reworking of the ILTA Guidelines provides a useful and accessible resource which may support best practice across diverse contexts and the development of Language Assessment Literacy.

Literacy in the wild: A study of the construct relevance of a health literacy test

Author(s): Susy Macqueen, Rosalyn Thyer

Key words: language for specific purposes, health literacy, health communication, speaking skills, layperson assessment

Abstract: Language for Specific Purposes (LSP) assessments typically focus on the language trajectories of subject matter experts, whether they are engaged in formal learning of specialist knowledge (such as nursing students) or fulfilling a language requirement for specialist work (e.g., for professional registration). Layperson assessments, on the other hand, are assessments that gauge the understanding of non-specialists (laypersons) who have cause to engage with specialist content (Knoch & Macqueen, 2020). Health literacy tests are layperson assessments that assess the extent to which patients can understand and apply health information (Berkman et al., 2010). As health care is mediated through communication, how health literacy is enacted communicatively in the health domain itself is an important validity question for these instruments. This paper examines the operationalised construct of a common health literacy test ('Newest Vital Sign') and compares it with the spoken interactions of patients undergoing treatment for heart failure. It takes a comparative case study approach, focusing on patients with different levels of experience with their chronic condition, ranging from newly-diagnosed to highly experienced. The study uses cognitive interviewing techniques (Miller et al., 2014) and theme-oriented discourse analytic methods (Roberts & Sarangi, 2005). Findings show that the operationalised construct of this widely used test does not capture the co-constructed, communicative nature of health literacy, which is critical in health care provision and in safety standards. It is argued that it is necessary to reconceptualise the assessment of health literacy in a way that reflects the ongoing learning, communication and collaboration that is necessary for the respectful management of chronic conditions.

From postgraduate education and beyond: Exploring the TESOL teacher's journey of becoming language assessment literate

Author(s): Queenie Mak

Key words: language assessment literacy, TESOL teachers, postgraduate TESOL education

Abstract: [Background] While the research literature on language assessment literacy (LAL) is continuously expanding, little is known about how TESOL (Teaching English to Speakers of Other Languages) teachers become language assessment literate over time. [Aims] This doctoral study, drawing on the LAL content model (Fulcher, 2012) and the LAL continuum model (Pill & Harding, 2013), explored the LAL developmental stages of TESOL teachers from postgraduate teacher education through to professional teaching practices. It also investigated the contributions of postgraduate education made to TESOL teachers' LAL development, along with other mediating factors that interacted with this development process. [Context and participants] The study was conducted within a postgraduate language assessment subject offered by a regional Australian university. The participants, selected through purposeful sampling, included 15 postgraduate student teachers undertaking the assessment subject, 10 postgraduate alumni teachers who had completed the same subject, and the subject lecturer. [Data collection] Following a mixed-methods case study design, data were collected in July 2023 to March 2024 using questionnaires and interviews. The questionnaires, a self-evaluation instrument containing 27 Likert-scale statements, measured the participants' self-perceived LAL, while the semi-structured interviews provided deeper insights into the participants' experiences, reflections, and insights gained throughout their LAL development journey. [Data analysis - in progress] The questionnaire and interview data are being analyzed using descriptive statistics and thematic analysis respectively. [Preliminary findings] Initial results indicate that the postgraduate language assessment subject contributed positively to the LAL development of both the student and alumni teachers. Other contributing factors, from the perspectives of the participants, included increasing teaching and assessment experience, intrinsic motivation to become better assessors, and empowerment from schools. These findings are expected to have implications for TESOL teachers' professional development in language assessment, and the enhancement in the design of postgraduate TESOL education.

Between Screens and Stress: Investigating Affective Barriers to Online Formative Assessment in EFL Contexts

Author(s): Alireza Maleki

Key words: affective barriers, online formative assessment, quality education, reduced inequalities, EFL teachers

Abstract: While extensive research has explored various challenges in online assessment, the affective barriers faced by teachers in the context of online formative assessment, particularly in English as a Foreign Language (EFL) settings, remain underexamined. This study seeks to bridge this gap by investigating EFL teachers' perspectives on the affective challenges associated with online formative assessments. Employing a mixed-methods approach, the study involved 30 participants from three distinct educational settings. In the qualitative phase, in-depth focus group discussions were conducted via Google Meet, allowing participants to articulate their experiences and concerns regarding online formative assessment. The qualitative analysis revealed three broad categories of affective barriers: psychological and emotional barriers, workload-related challenges, and student and system-related concerns. A ranking scale was then utilized to assess the perceived significance of these categories. The findings offer valuable insights for EFL teachers, policymakers, and educational assessment bodies, highlighting the need for teacher support mechanisms, professional development programs, and assessment policies that acknowledge and address these affective challenges. This study underscores the importance of fostering a more teacher-centered approach to online assessment design, ensuring that both technological and pedagogical frameworks align with educators' emotional and professional well-being.

Human versus machine: The effectiveness of ChatGPT in automated essay scoring.

Author(s): Jennifer Manning, Jeffrey Baldwin, Natasha Powell

Key words: AI, ChatGPT, assessment, graduate students, South Korea

Abstract: As AI tools become more integrated into HE, they may improve teaching efficiency. ChatGPT's ability to support learning and instructional design suggests it could transform grading (Dai et al., 2023). However, limited

research exists on its use in assessment. This study examines GPTs' potential as instructor-assisting tools for evaluating student writing.

Fifty annotated bibliographies from graduate students at two STEM universities in South Korea were collected. A rubric assessing content, academic tone, structure, and grammar. Human raters (HRs), experienced ELT instructors, independently graded the texts and input the rubric and texts into GPT models (3.5 and 4). Statistical analyses using SPSS assessed the accuracy, reliability, and consistency of ratings.

The HRs showed strong consistency in scoring, while ChatGPT3.5 (GPT3) and ChatGPT4 (GPT 4) assigned higher, more variable scores. GPT4 exhibited greater inconsistency than GPT3. Normality tests indicated HR scores were normally distributed, whereas GPT scores were skewed. Mean Absolute Deviation (MAD) analysis showed good alignment among HRs, with GPT4B aligning most closely with them. Intraclass Correlation Coefficient (ICC) results confirmed HRs had high agreement, GPT3 had moderate agreement, and GPT4 had poor agreement. Spearman's correlation revealed strong agreement among HRs, but weaker correlations between GPTs and HRs, suggesting GPTs were inconsistent in replicating human assessments.

This suggests that AI could supplement, but not replace, human assessment, highlighting the need for instructor involvement, prompt engineering training, along with ethical considerations.

Profiles of university faculty for Generative Artificial Intelligence feedback: a comparative Europe-Australia study

Author(s): Maria Teresa Mateo-Girona, Ana Mari-a Ducasse, Carmen Lopez Ferrero

Key words: e-feedback literacy, GenAI, pre-service teacher training, digital competence, Master's thesis

Abstract: The emergence of Generative Artificial Intelligence (GenAI) in academia has generated a profound debate about its impact on the assessment processes and learning support; in particular, this research focuses on the context of Master's thesis supervision. This empirical study, descriptive and exploratory in nature, investigates how university teachers are responding to the emerging use of GenAI tools in the written feedback they provide to students during the writing of their dissertations. The study focuses on three axes: first, the uses, perceptions and needs of teachers are analysed. Secondly, the socio-demographic variables of the teaching staff (gender, age, teaching experience) that influence these uses, perceptions and needs are identified. Thirdly, the relationship is established between teachers' digital and writing competence profiles and their degree of integration of GenAI in their feedback practices. A validated questionnaire was administered to Master's thesis supervisors from European and Australian universities. The results show that the majority of teachers do not use GenAI tools to provide feedback on Master's thesis, although they do recognise their potential. The CEFR (Common European Framework of Reference) profiles B1 Integrator and B2 Expert, based on a digital competence model, predominate over the C1 Leader and C2 Pioneer profiles in written competence. With regard to training needs, the majority of teachers express an urgent need for specific training in the critical and ethical use of GenAI. These findings allow us to problematise the traditional constructs of writing assessment within a context of technological transformation and propose frameworks for GenAI-supported feedback. Consequently, proposals for critical literacy in GenAI are developed to promote formative, reflective and contextualised evaluative practices that integrate technological innovation without renouncing fundamental pedagogical principles.

Wisdom in the Counsel of Many: Refining Cognitive Reading Attributes through Multi-Criteria Fuzzy Delphi

Author(s): Muhamad Firdaus Mohd Noh, Mohd Effendi Ewan Mohd Matore, Nur Ainil Sulaiman

Key words: reading attribute, cognitive diagnosis, test construct, language assessment, cognitive diagnostic

Abstract: A fundamental component of cognitive diagnosis involves the specification of attributes. Attributes, defined as discrete, fine-grained skills, must fulfil four essential criteria: they should align with the curriculum of the intended population (relevance), contribute meaningfully to the overall construct being measured (importance), be capable of being operationalized into test items (measurability), and offer detailed diagnostic feedback to stakeholders (specificity). The present study aims to specify cognitive reading attributes by eliciting expert evaluations based on these four criteria. Utilizing a multi-criteria Fuzzy Delphi Method, the study engaged ten expert panellists, comprising university lecturers with established expertise in language instruction, assessment and testing, selected through purposive sampling. The instrument employed in this study was developed through a comprehensive review of existing literature, reading taxonomies, and established reading assessment frameworks, resulting in the identification of 15 key cognitive reading attributes for expert evaluation. Anchored on the established acceptance criteria, namely a threshold

value (d) of less than 0.2, expert consensus exceeding 75%, and a defuzzification value surpassing the α -cut of 0.5, the analysis revealed that seven attributes were deemed relevant, ten were identified as important, seven were considered measurable, and nine were recognized as specific. To determine the overall ranking of the attributes across these four evaluation criteria, expert consensus values were compared. The results indicated that the top five attributes, based on their consistent prominence across the dimensions of relevance, importance, measurability, and specificity, are: retrieving explicit information, recognizing word meaning, identifying main ideas, making inferences, and summarizing main ideas. These findings underscore the critical role of expert-driven validation in establishing cognitively diagnostic reading attributes that are theoretically sound and practically applicable in assessment design. However, the study is limited by the homogenous expert panel, suggesting that future research should incorporate a broader range of stakeholders, including classroom practitioners and curriculum developers.

Assessing reading comprehension of young Indian learners through a multilingual multimodal design: A study

Author(s): Lina Mukhopadhyay

Key words: reading comprehension, home language, multilingual-multimodal assessment, formative assessment, primary graders

Abstract: Reading comprehension skill requiring knowledge of academic language in print is a pre-requisite for success in school as it supports knowledge development and its application in language and content classes. However, in multilingual countries in the Global South like India, primary grades often have a language of instruction (like English) that is not available as a home language for the majority of learners who come from low SES families with little or no support for literacy practices at home (UNESCO GEM report, 2016). Though such children are advanced in their orality in multiple languages they use at home and surrounding contexts, their print comprehension in the language of instruction is not well-developed (Tsimpli et al., 2020). Added to this challenge is the fact that assessment of reading comprehension is largely monolingual and summative. In contrast, teacher-led formative and alternative assessments are either absent or do not tap into learners' home languages to assess comprehension skills.

To address the gap of under use of young learners' (fourth and fifth graders) home language(s) in formative reading assessments, the University of Cambridge in collaboration with a consortium of Indian universities have designed and piloted multilingual and multimodal reading assessments. These are examples of classroom-based assessments of early comprehension skills in English mediated by languages like Telugu, Hindi and Assamese. The assessments are on narrative and expository texts for reading sub-skills with parallel MCQ items presented in the target language and one home language. The findings reveal the positive impact of home language on expository text comprehension, which use academic language and are central to a child's understanding of all school subjects. The findings have implications for training teachers about the principles of need-based alternative reading assessments as a new direction to assess young multilingual learners in a responsive and ethical manner.

Benchmarking scores across tests: Stakeholders perceived test score equivalence for professions and universities across the world.

Author(s): Amanda Muller

Key words: test-score equivalence, professions, universities, benchmarking, globally

Abstract: This paper gives the results of a benchmarking activity to establish the range of minimum test scores that stakeholders use for professional registration and professional degrees across Australia, Canada, Ireland, New Zealand, United Kingdom, & United States. It will explain the trends and the surprising discrepancies found in score setting equivalence across the tests of IELTS, TOEFL, PTE, C1A/CAE, and DET. It compares the stakeholder score setting against the equivalence score tables provided by test makers.

The presentation will demonstrate, using graphing of the data, that stakeholders' wide variation in setting score equivalence across tests means that the candidates true proficiency cannot be guaranteed. This score equivalence variation found in the study will be demonstrated to undermine the use of English proficiency testing. The consequences of this disagreement among stakeholder equivalence scores are discussed, including how it encourages candidates to shop around for the easiest requirement and gives the wrong message about candidates having adequate proficiency to enter a degree or profession.

This study also will discuss variations between the scores required for professional registration (which has knock-on effects for migration) and the scores required for professional degrees. Ideally, due to student placement and practice on the public, university and profession scores should match. The data of this study will reveal the amount of mismatch between degrees and professions.

Finally, this study will outline any differences between countries, and explore whether factors such as international student ratio or university ranking affects choice of minimum proficiency scores.

Assessment of L2 Japanese vocabulary: Building assessment expertise and resources with JFL teachers in Australian secondary schools

Author(s): Fusae Nojima

Key words: L2 vocabulary assessment, teacher collaboration, classroom-based assessment, praxis-oriented research , Japanese as a foreign language (JFL)

Abstract: Vocabulary, or lexical knowledge, is essential for second language (L2) learning. L2 learners must acquire a significant amount of vocabulary to communicate effectively in the target language. Research has shown that vocabulary learning is crucial for all language skills (i.e., listening, speaking, reading, and writing). While both explicit teaching and implicit learning have been found effective, time constraints and learner diversity often limit vocabulary instruction and assessment in the classroom.

Although vocabulary is fundamental, it must be integrated across language skills. Learning words in isolation is not the goal; rather, learners need to use vocabulary meaningfully in communication. Given its importance, assessing vocabulary effectively is critical for guiding instruction and monitoring student progress. Yet vocabulary assessment is complex: How many words do students need to know, and how well do they need to know them for a given learning goal? Should vocabulary be assessed in isolation or embedded in broader language tasks?

This study explores the current vocabulary assessment practices and perspectives of Japanese as a foreign language (JFL) teachers in Australian secondary schools. It aims to collaboratively develop a context-sensitive assessment tool and item bank tailored to L2 Japanese learners. By working closely with teachers, the study seeks to identify which dimensions of vocabulary knowledge are valued and how they are assessed in practice. The research adopts a praxis-oriented framework, using a mixed-methods design, including teacher surveys, collaborative workshops, teacher interviews, and student interviews. Ultimately, this study aims to co-construct practical, classroom-relevant vocabulary assessment resources that align with both pedagogical values and learner needs.

Bridging Technological Innovation and Assessment Rigor: Computational Validation of LLM-Generated CEFR-Aligned Reading Passages for High-Stakes Testing

Author(s): Norazha Paiman

Key words: automated item generation, CEFR alignment, computational textual analysis, large language models, assessment validity

Abstract: This study confronts the challenge of reconciling technological innovation with psychometric rigor in language assessment through systematic validation of reading passages generated by Large Language Models (LLMs). As AI-generated content permeates high-stakes testing environments, this research establishes methodological frameworks to evaluate whether such outputs can satisfy the stringent validity requirements of traditional proficiency benchmarks. Employing a multi-method computational validation approach, the study analyses 30 discipline-specific reading passages (n=15 per model) from ChatGPT-4 and Claude 3.5 against CEFR proficiency standards. A layered analytical architecture combines: (1) multi-dimensional readability assessment, (2) corpus-based lexical sophistication analysis, (3) discourse complexity metrics, and (4) CEFR alignment through Text Inspector. Quantitative outcomes reveal robust statistically significant alignment between target and empirical proficiency levels (Pearson's $r=.78$, $p<.001$), with 93.3% of ChatGPT and 86.7% of Claude outputs exceeding CEFR C2 thresholds. Both models demonstrated C2-appropriate lexical distributions ($\chi^2=4.32$, $p=.115$), yet marked disparities emerged in linguistic patterning: ChatGPT-4 exhibited superior syntactic consistency (Cohen's $d=1.24$), while Claude 3.5 generated texts with greater pragmatic variation (entropy differential=0.67 bits/word). These differential capabilities suggest model-specific utility: ChatGPT for standardized structural precision versus Claude for contextually adaptive content. The findings advance assessment validity theory by operationalizing computational verification protocols for AI-generated test materials. This empirical

demonstration that LLMs can produce CEFR-compliant passages at scale reconfigures debates about automated item generation, proposing a hybrid paradigm where computational linguistics techniques safeguard construct validity while harnessing AI's productive capacity. For assessment practitioners, the study provides an actionable validation framework combining psychometric rigor with computational efficiency. By establishing statistically grounded thresholds for lexical complexity, discourse cohesion, and pragmatic adequacy, it enables systematic quality control of LLM outputs within existing validity paradigms. This study ultimately bridges the innovation-validity dichotomy, offering language testing institutions a replicable methodology to responsibly integrate generative AI into high-stakes assessment ecosystems while preserving measurement integrity.

From ZPD to GPT: Designing AI-Supported Formative Assessments without Losing Construct Validity

Author(s): JaeYoon Park

Key words: generative AI, formative assessment, differentiated instruction, classroom-based assessment, construct validity

Abstract: This work-in-progress investigates how generative AI tools such as ChatGPT can guide, support, or reshape traditional classroom-based language assessment practices. In an era where innovation and tradition often collide, the project explores how AI can be leveraged not to replace teacher expertise, but to extend it, particularly in formative, low-stakes assessment environments.

The study draws on three complementary frameworks. First, Lev Vygotsky's sociocultural theory and the concept of the Zone of Proximal Development (ZPD) position AI tools as mediators of learning, offering scaffolded feedback that supports learners in progressing toward independent performance. Second, Carol Ann Tomlinson's model of differentiated instruction underpins the exploration of how AI can offer adaptive, learner-specific assessment pathways, modifying complexity, format, or feedback style to suit varied learner needs. Third, Robert J. Mislevy's construct-centered design framework guides the evaluation of AI-generated assessment tasks, ensuring alignment with the targeted construct and avoiding irrelevant variance.

Currently in the planning and piloting phase, this project will share prototype assessment tasks and AI-generated feedback examples designed to support communicative competence and learner agency. No student data will be presented, but illustrative use-case scenarios will be included.

The presenter seeks feedback on refining the theoretical model, identifying risks of construct drift or over-scaffolding, and ensuring teacher judgment remains central to the assessment process. Ethical and practical considerations 'such as transparency, validity, and learner trust' will also be discussed.

This session invites dialogue about how generative AI can responsibly support a shift toward more responsive, differentiated, and theoretically grounded assessment practices in the language classroom.

Demystifying AI-assisted writing process: How do L2 learners with varied levels of AI literacy engage with GenAI tools?

Author(s): Carrie Peng

Key words: AI literacy, learner engagement, writing process, AI-assisted L2 writing

Abstract: Recent scholarship has revealed how generative artificial intelligence (GenAI) tools have the potential to assess second language (L2) learners' writing following provided rubrics and generate instant and personalised written feedback, for learners of varied L2 proficiency levels. In addition to language barriers faced by L2 writers, researchers noted that the levels of AI literacy also tend to influence how they can meaningfully engage with GenAI tools (Du & Yang, 2025; Wang et al., 2025). To better understand the implications of incorporating GenAI tools in L2 writing assessment and instruction, it is necessary to gain insights into how L2 writers with diverse AI literacy profiles employ various GenAI tools in authentic academic writing tasks and how they engage with GenAI output during their writing process. This multiple-case study presents preliminary findings on how varying levels of AI literacy shaped four postgraduate L2 learners' engagement with GenAI tools during their real-time academic writing processes. Using screen recordings, stimulated recall interviews, and a focus group interview, the study traced participants' distinctive patterns of behavioural, cognitive, metacognitive, and affective engagement across different writing stages. Findings suggest that critical AI literacy, or lack thereof, as a multifaceted construct, largely influences the ways L2 writers utilised different GenAI tools (e.g., durations, frequencies, and prompts) in drafting and revising their essays and how they

incorporated AI output with different levels of cognitive and metacognitive engagement. A range of individual and contextual factors, including individual writing self-efficacy, understanding of institutional AI policies, influences from peers and lecturers, and specific task requirements, also seem to influence L2 writers' engagement patterns with GenAI tools. Implications for L2 writing assessment and integrating GenAI tools into academic writing instruction will be discussed.

Exploring the possibilities of integrating communicative AI into the IELTS test preparation process: The new horizon of human-computer communication.

Author(s): Carlo Perrotta, Sima Mohammadi

Key words: test preparation, language education, prompting, prompt engineering, communicative AI

Abstract: This paper examines Generative AI (GenAI) from a perspective of language education and language test preparation. While there is already a growing body of research on GenAI and language learning, the role of prompting, understood as a novel form of human-computer communication, has been neglected. To address this gap, we carried out a study in two parts. In the first part, we conducted a scoping review that focused on how prompting GenAI language models can support English language learning and assessment. The review identified three application scenarios for prompts: text generation, test item generation and automated assessment. For each scenario, we examined how prompts were constructed and how they could be replicated. In addition to these scenarios, the review also highlighted the emerging importance of 'meta prompts' which are distinct from user-oriented prompts in that they operate in the back-end of AI models and are not visible or modifiable. We also found that effective prompts can certainly increase the sophistication and precision of human-AI interactions, but the outputs of these interactions still display limitations in terms of contextual awareness, bias, reliability and performance consistency. In the second part, we conducted qualitative fieldwork to understand how prompting as a communicative practice is emerging in a real context of language learning and language test preparation. To this end, we carried out interviews and observations in a language school located in a large Australian city. We found that both teachers and students engage in highly contextual forms of sensemaking that influence informal theories about GenAI use and prompting. We also found evidence of a form of prompting amongst students that we termed 'tactical': unsophisticated but reflecting pragmatic and subjective priorities. In the conclusion, we reflect on the significance of prompting as pragmatic communication and suggest some implications and future research directions.

The process of revalidating published L2 vocabulary tests for a specific population of learners

Author(s): John Read, Thi Ngoc Yen Dang, Thi Phuong Dung Cao, Thi My Hang Nguyen

Key words: test re-validation, academic vocabulary knowledge, L2 vocabulary tests

Abstract: There have been recent calls to adopt more comprehensive procedures for validating published L2 vocabulary tests for use with learners in particular educational contexts (Read, 2023; Schmitt et al., 2020). The argument is that researchers should first obtain evidence that the instruments they propose to use are suitable for the target population of learners and for the intended research objectives. The work reported here was the first stage of a longitudinal project to investigate the development of academic vocabulary knowledge among university students undertaking English-medium study in Vietnam. Three pairs of published tests were selected for the study the New Computer Adaptive Test of Size and Strength (Aviad-Levitsky et al., 2019), the Academic Vocabulary Test (Pecorari et al, 2019), and the Academic Collocations Test (Nguyen et al., 2024) to assess respectively general English words, academic words, and academic collocations. An additional design requirement was that each paired test should have both a form-meaning recognition version, using a selected-response format, and a form recall version, with gap-fill items. In this presentation we will discuss the three steps involved in re-validating the three tests for use in our project. Initially, we conducted an expert review of the instruments, resulting in a redesign of one test version and several other revisions to the tests. The second step was a tryout of the tests with 24 learners from the target population. A detailed analysis of their responses led to further revisions. Thirdly, a larger pilot study was conducted with 200 students, including interviews with a sub-sample of them, to confirm the technical quality and overall appropriateness of the revised tests for the target population and the overall research project. The presentation will illustrate the kinds of revisions made to the tests and reflect on the value of the revalidation process in L2 vocabulary research.

Integrating ChatGPT into ESL University Students' IELTS Speaking Practice

Author(s): Mao Sasaki

Key words: *ChatGPT, IELTS speaking, speaking skills, feedback, assessment, AI-assisted language learning*

Abstract: This study explored how ChatGPT can support IELTS Speaking practice, analysing its assessment and feedback capabilities and learners' experiences. 10 ESL learners participated in a five-day speaking practice intervention, from which data was collected, including chat histories, transcriptions of ChatGPT-learner interactions, learners' diaries, and post-intervention questionnaire and interview responses. Each day, participants practised speaking twice on the same topic while receiving feedback and scores based on the four IELTS assessment criteria from ChatGPT: fluency/coherence, vocabulary, grammar, and pronunciation.

ChatGPT generally provided higher scores for participants' second attempts than their first attempts each day. However, the degree of improvement varied across assessment criteria, showing relatively equal gains in grammar, vocabulary, and fluency/coherence yet smaller gains in pronunciation. Some participants, however, received lower scores on their second attempts. Over five days, while participants typically improved their overall speaking scores, some did not and experienced considerable day-to-day fluctuations in their scores. These results raised questions about whether ChatGPT consistently and reliably rated participants' performance. Nevertheless, these scoring patterns may also be influenced by ChatGPT's feedback and learners' engagement.

ChatGPT provided specific corrective feedback on grammar, vocabulary, and fluency/coherence, while pronunciation feedback was less targeted (e.g., mimicking native speakers). The uneven quality of feedback may explain the varied score improvements across criteria.

Participants appreciated ChatGPT's specific feedback and its accessible and low-stress nature. However, they noted emotional disconnection and occasional frustration due to ChatGPT's mechanical features (e.g., lengthy responses), which appeared to influence participants' engagement in speaking practice and score improvement/fluctuations. Overall, ChatGPT can be used for speaking practice, though some limitations remain.

Collaboration with Policymakers

Author(s): Laura Schildt

Key words: language assessment policy, policymakers, language testing specialists, policy formation, collaboration with test score users

Abstract: Language policy is the primary mechanism through which public institutions organize, manage and manipulate language use in society (Shohamy, 2006). In the area of language testing, there is a pressing need for language experts to collaborate with policymakers in the development of fair language policies and to promote the ethical use of test scores. There has been significant work considering the use of tests as instruments of policy (McNamara & Shohamy, 2009; Shohamy, 2001) and the issues this raises for validity theory (Mcnamara, 2010; McNamara & Ryan, 2011). However, there is less research on the formation of language testing policies and the role of language testing specialists, making it difficult for junior language testers to learn about policy work from their more experienced peers. This is in stark contrast to other disciplines where the role of experts in policy is supported by robust theoretical frameworks (Brans et al., 2022; Head, 2015; Jungblut et al., 2024; Pielke, 2007). Therefore, the present study addresses this research gap using Q-methodology to analyse 43 statements related to policy structures, communication, advisory methods, and ethical principles. A sample of 52 language experts from 23 countries participated in the study. Q-factor analysis in QMethod software and Python 3 were employed to identify and prioritize the factors that influence collaboration with policymakers. Five overarching groups related to different aspects of effective collaboration are considered. The resulting policy advising index underscores the importance of factors such as the Policymaking Environment, Networking, Communication Skills, Policy Analysis, and Policy Advising. This research highlights the importance of a clear understanding of the factors that influence policymaking to effectively improve collaboration.

Assessment of PFL proficiency: Factors influencing learner performance in different cloze test formats

Author(s): Clara Setas

Key words: cloze test, influence factors, Portuguese as a foreign language assessment, assessment complexity, item and test design

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

Predicting reading proficiency with vocabulary knowledge: Modalities, frequency, and test lengths

Author(s): Ji-young Shin, Pablo Robles-García, Jeff Stewart

Key words: vocabulary, meaning recall, meaning recognition, reading, L2 English

Abstract: While vocabulary is known as the foundation of reading (Schmitt et al., 2011), mixed results have been reported regarding how different modalities of vocabulary knowledge – meaning recognition and meaning recall – predict reading comprehension (Kremmel & Schmitt, 2016; Laufer & Aviad-Levitzky, 2017). This study compared the prediction of reading by two modalities of vocabulary knowledge, focusing on test lengths and frequency bands.

A total of 119 L2 English speakers took one reading test (TOEFL reading section) and two vocabulary tests: the Updated Vocabulary Levels Test (Webb et al., 2017), a multiple-choice meaning-recognition format, and its meaning-recall version, which required translation of the words in their L1s. Both included 150 items with 30 words selected from each 1,000-frequency band of the 5,000 most frequent words. The total and frequency band scores of vocabulary tests were correlated with reading to compare the two modalities (RQ1) and differences by test lengths and frequency bands (RQ2).

Results showed that the two vocabulary tests displayed high reliability (.94 for the meaning recall test and .97 for the meaning-recognition test) and were significantly correlated with reading ($r = .53$ and $.58$, respectively), with little difference between them ($ZH = .79$, $p = .22$). Interestingly, shorter meaning-recall tests generally maintained strong associations. In contrast, meaning recognition showed significant increases from the 1K-band to longer tests, particularly from the two-band test (1K to 2K) to the four-band (1K to 4K) and full-length tests (1K to 5K).

Additionally, three linear regressions were conducted to assess the benefits of using both vocabulary tests to predict reading (RQ3), revealing slight but significant increases. The findings support strong links between both modalities of vocabulary knowledge and reading, suggesting that the same cognitive mechanism might operate differently between receptive and productive modalities. Moreover, meaning recall might be particularly beneficial when a full-length test is unavailable or impractical.

The power of testers and their tests: A sociological analysis of assessment practices in Australian Direct Entry Programs

Author(s): Kyle Smith

Key words: IELTS, power, Bourdieu, Shohamy, direct entry programs

Abstract: Critical Language Testing scholars have argued for over 30 years that large-scale standardised language tests such as IELTS have an unrivalled power to alter test-takers' behaviour and dictate decisions made by educators and policymakers. However, by locating this power in the tests themselves, scholars have overlooked the power accumulated by social agents whose ideas about assessment practices are seen as legitimate.

This paper explores sociologically the role that institutions play in power relations within the Language Testing field, taking up the Bourdieusian thread in Shohamy's (2001) *The Power of Tests*. From a Bourdieusian perspective, an institution such as IELTS exists durably in two forms: objectively, as test items, rubrics, procedures, etc., and subjectively, as IELTS-related dispositions within individual social agents. The paper also emphasises the crucial Bourdieusian concept of reflexivity, which means that scholars must account not only for the behaviour of test-takers, educators and policymakers, but for themselves and those who design language tests.

Methodologically, this paper reports on a Bourdieusian field analysis of assessment practices in Direct Entry Programs based on documents and interviews with staff from two university English centres in Australia, and Language Testing historiographical texts published since the 1960s. This analysis indicates that the contemporary 'power of tests' can be understood as originating in relations among language testers established in the 1960s and the assessment practices that they constructed and legitimised. At the same time, analysis shows that these power relations are being resisted, and staff in Direct Entry Programs are developing alternatives to Language Testing orthodoxy.

The paper argues that challenging a powerful Language Testing institution such as IELTS requires challenging both its objective and subjective forms, and posing alternatives that are also strongly institutionalised. It concludes with practical recommendations for institutionalising alternative approaches to assessment.

Exploring the importance of academic integrity and cheating prevention in post-pandemic language assessment: extending the concept of language assessment literacy

Author(s): Anna Soltyska

Key words: academic integrity, cheating, language assessment literacy, mixed-methods

Abstract: The emergence of online support tools, including GenAI, coupled with an accelerated post-pandemic shift in language testing from analogue paper-based to online and computer-based assessments, has led to increased security concerns among various stakeholders. Test developers and providers of large-scale standardised assessments claim to implement multiple security measures such as online proctoring and secure exam browsers to reduce cheating and test-related malpractice. At the same time, students and teachers involved in low-stakes classroom-based assessment are literally left to their own devices and are often unaware of what constitutes cheating and how to prevent academic misconduct in the face of new tools and no rules. To date, the concept of (language) assessment literacy has not specifically addressed issues of academic integrity and cheating, although they are linked to the validity and reliability of assessment.

The purpose of this mixed-methods study is to explore the importance and understanding of academic integrity and cheating prevention among different stakeholders in language assessment in higher education. In particular, the study aims to shed light on which practices are perceived as unlawful by different stakeholder groups, and what preventive and educational measures are taken to maintain the integrity of test-takers.

A scoping review of articles published in selected language assessment journals and papers presented at major language assessment conferences around the world aims to identify key areas of concern related to cheating and test security. Data will be collected through an online questionnaire for language teachers and students (test-takers) based on earlier assessment literacy surveys and issues emerging from the scoping review. Follow-up semi-structured interviews will be conducted with selected teachers and representatives of standardised test providers. Case studies will offer further insight into academic integrity curricula and classroom procedures.

Teacher Assessment Literacy: The relationship between language awareness and writing assessment behaviour

Author(s): Susanne Stanyer

Key words: language awareness, writing assessment, teacher assessment literacy, EAL

Abstract: This paper reports on a PhD study that explored the nature of the relationship between the language awareness (LA) of primary school teachers in Victoria, Australia, and the behaviour of these teachers as they assessed writing produced by young learners of English as an additional language. The study was undertaken to address the need for more research into the knowledge and skills required of teachers in order to make trustworthy assessment decisions, particularly when using assessment tools. A mixed methods approach was used in this study. In the first of two phases, 35 teachers completed an online survey and an online language awareness test. Resulting data were analysed using a largely quantitative approach, to determine the extent of the participants' declarative LA. In the second phase, qualitative methods were adopted. In this phase, three case study participants 'thought aloud' as they assessed samples of writing using criteria sheets from the Tools to Enhance Assessment Literacy project, and also participated in semi-structured interviews. Data were analysed manually to identify ways in which the participants dealt with linguistic concepts or metalanguage that were familiar and less familiar to them and to ascertain their shared/differing views on the relationship at the heart of the study. Several interrelated personal, professional and contextual factors, including the teachers' LA, were found to influence the teachers' assessment behaviour. Knowing about language was associated with accurate decision-making and self-reported confidence. Not fully understanding points of language was associated with inaccurate commentary, minimal engagement with criteria, self-reported experiences of doubt; and for two teachers, the use of problem-solving strategies. With implications for teachers' professional development, these findings establish LA as a critical component of language teachers' assessment literacy.

The value of using writing scales in IELTS writing test preparation: An ecological perspective

Author(s): Yuzhu (Betty) Su

Key words: IELTS writing scales, writing test preparation, ecological perspective

Abstract: This case study investigates how a teacher at a private language institution in southern China helps learners prepare for the International English Language Testing System (IELTS) writing test. It focusses on how the teacher uses the IELTS writing scales (used by examiners to score test essays) to assist students in understanding test expectations. Writing assessment research has explored the use of writing scales to provide feedback, support students' writing processes, and aid test preparation. This study narrows the focus to explore the students' and teacher's reflections and perceptions before, during, and after IELTS writing preparation classes.

An ecological perspective is adopted to provide a holistic understanding of the target "case". Accordingly, the study employs qualitative methods including semi-structured interviews, classroom observations, reflective diaries and analyses of student assignments. The ecological lens allows the research to consider the complex interplay between the teacher's instructional choices, students' engagement with the writing scales, and the broader testing and institutional context.

Findings from this case study offer insights into how writing scales can be effectively repurposed for test preparation, with implications for writing assessment practices that support learners' development. The ecological approach provides a rich understanding of the factors shaping the integration of writing scales in this specific instructional setting.

Teacher's perspectives on ELF assessment: ELF assessment scale construction and validation

Author(s): Fereshte Tadayon

Key words: ELF, assessment, domain, dimension, rating rubric

Abstract: This study investigated quantitative data from 236 English as a Foreign Language (EFL) teachers, seeking to capture their perspectives on English as a Lingua Franca (ELF) assessment. The exploratory factor analysis of the questionnaire yielded five major factors representing EFL teachers' perceptions on ELF assessment: (a) ELF-informed assessment domain, (b) test accommodation, (c) task and text variety, (d) locally-informed features, and (e) alternative assessment criteria. The ELF-informed assessment domain indicates a strong emphasis on evaluating Nonnative English Speakers (NNESs) based on their comprehension and engagement with diverse English varieties and their intercultural competence. This factor underscores the need for test constructs that reflect real-world scenarios and local practices

for a more effective assessment. Test accommodation, including the adjustment of test tasks and rating rubrics, emerged as critical to accommodating NNEs' real performance in tests. The importance of task and text variety was highlighted, advocating for adaptation to various accents, communication styles, and cultural perspectives. Locally-informed features were identified as essential for capturing localized language use and sociocultural contexts in testing materials. Additionally, alternative assessment criteria were emphasized, such as using proficient NNEs as benchmarks and selecting raters from different linguistic backgrounds. This nuanced approach ensures greater contextual validity and effectiveness in language assessments, tailored to local and global audiences, and aligned with the variable nature of ELF communication.

Building the foundations for automatic assessment of verbal and nonverbal aspects of spoken interaction in Finnish as a second language

Author(s): Riikka Ullakonoja

Key words: automatic speaking assessment, nonverbal features, interaction, Finnish

Abstract: We describe the theoretical and methodological basis of a multidisciplinary research project, focusing on automatic assessment of spoken interaction in Finnish as a second language. Our goal is novel in three ways. Firstly, we aim to develop an ASA (Automatic Speaking Assessment) system to assess oral proficiency in dialogic, instead of monologic, speech. Secondly, our approach includes automatic assessment of nonverbal features in interaction (e.g., gazes, hand, and head movements), extending beyond the more conventionally used ASA tools focusing only on speech signal. Thirdly, the language to be assessed is Finnish, a language with scarce previous studies on automatic assessment, apart from our previous project. We also present preliminary statistics and observations from the data.

The construct of spoken interaction traditionally deployed in assessment entails linguistic competences, while nonverbal cues are rarely included in rating scales used by human raters of interactive tasks despite their crucial role in human encounters. Thus, we aim to improve the validity of dialogic speaking assessment by including nonverbal features in the construct of oral proficiency and in automatic assessment. This is important, because paired speaking tasks are commonplace in educational settings and should thus be increasingly applied in high-stakes contexts.

Redressing an imbalance - assessing higher-order reading processes in a test of academic reading ability for university entry purposes

Author(s): Stephen Walker

Key words: academic reading constructs, higher-order cognitive processes, multi-stage adaptive design

Abstract: This presentation outlines the a priori validation process followed to develop a multi-stage adaptive test of academic reading which targets constructs considered crucial to success in university study settings.

Researchers (Green and Hawkey 2012, Weir et al 2012a, 2012b, Bax 2015, Owen 2016) have argued that reading tasks requiring higher-order levels of cognitive processing should be prioritised in tests of academic reading.

However, research into existing large-scale tests suggests that higher-level processes are under-represented (Owen 2016) with a significant proportion of items relying on the lower-level processes associated with careful local reading.

Following Weir's (2005) call for validation at the a priori stage of test development we identified the following constructs as those of principal interest:

- Expeditious global reading
- Careful global reading

We have developed specifications, exemplar tasks and are currently engaged in field trialling and item analysis. The test is computer-delivered and uses a multi-stage adaptive design.

We have questions regarding:

- Setting appropriate time limits to encourage the use of expeditious reading strategies in tasks designed to measure them
- Scoring the performance if students perform differentially on certain test stages

- Establishing the concurrent validity of the test if we are measuring constructs which differ significantly from those measured in other well-known tests
- Designing an eye-tracking study to look for evidence of the reading strategies being targeted

We hope that by presenting a working version of our assessment, and the rationale for what we have done so far, that we can get feedback to help us continue the test development process.

Unraveling the skill integration in integrated reading-to-speak tasks: The case of L2 Chinese learners

Author(s): Xiaozhu Wang

Key words: integrated assessment, reading-to-speak, strategy use, Chinese as a second language, assessment construct

Abstract: Integrated speaking tasks involve the integration of comprehension and speaking simultaneously. As a way to assess an L2 speaker's language use in an academic context, the integrated speaking task is used in large-scale language proficiency tests due to its authenticity and positive impact on language learners (Xi et al., 2021). Studies on the mechanism of skill integration in integrated speaking tasks are relatively scarce. This study is situated in the context of L2 Chinese assessment for advanced learners, and it extends the research scope by investigating the interactive relationship among independent speaking ability, general language proficiency, strategy use, and the performance in integrated reading-to-speak tasks.

The integrated reading-to-speak tasks we designed are of two types differing in visual resources, including texts and concept maps. 104 L2 Chinese learners with a proficiency of intermediate to high level participated in this study. Each participant completed 4 tasks (2 of each type) in a counter-balanced order and strategy-use questionnaires after each task type. Structural equation models were employed to explore the role of independent speaking ability, general language proficiency, and strategy use in integrated task performance.

Results showed that integrated speaking tasks assessed both oral and general language skills. Compared to text-based integrated speaking tasks, concept map-based tasks relied more heavily on independent speaking ability, while the former depended more on general language ability. In concept-map-based tasks, the "goal-setting" strategy had a positive effect on performance in integrated speaking tasks, whereas the "organizing content and addressing requirements" strategy had a negative effect. In text-based tasks, the "planning" strategy demonstrated a positive effect on performance in integrated speaking tasks. The study provided evidence to validate the construct of integrated speaking tasks, contributing to more accurate scoring dimensions and pedagogical practices in integrated language tasks.

Developing an AI Persona Perception Inventory (AIPPI) for Multimedia-Based Lecture Listening Assessment

Author(s): Ziteng Wang, Vahid Aryadoust

Key words: AI personas, questionnaire validation, multimedia listening

Abstract: Emerging Generative AI technologies offer innovative opportunities for multimedia-enhanced language assessments, yet learners' perceptions of AI-generated personas in listening tests remain underexplored. This study develops and validates an AI persona perception inventory (AIPPI), a tool grounded in Mayer's cognitive theory of multimedia learning (CTML), specifically the image and redundancy principles. The AIPPI is designed to measure university-level EFL learners' attitudes toward AI-generated lecturer personas in video-based listening assessments, using three constructs: affect, behavior, and cognition.

The questionnaire was designed based on multimedia learning and language assessment research, pilot-tested, and revised with expert feedback. It was administered to 142 Chinese EFL students alongside listening tasks featuring AI-generated lecture videos. Rasch analysis confirmed its psychometric validity through item-person fit statistics, dimensionality checks, and ordered response thresholds. Item reliability ranged from .60 to .96, and person reliability exceeded .89 across subscales. These results indicate the scale's stability and its capacity to differentiate learner perceptions effectively.

This work balances innovation and tradition by integrating AI personas—an emerging technological tool—into traditional listening assessment frameworks while employing established questionnaire validation methods. It addresses a critical gap in understanding how learners perceive AI lecturers, exploring whether these perceptions

enhance engagement or introduce challenges in assessment contexts. Preliminary findings suggest positive attitudes toward AI personas, with implications for their design and use in language testing.

The study contributes to multimedia-assisted listening theory by empirically applying CTML's image and redundancy principles within an assessment context and supports valid interpretation of AIPPI scores for future research and practice. By ensuring AI-enhanced assessments align with learners' needs and maintain validity, this research bridges innovative technology with traditional assessment principles, informing language testing policies and classroom practices.

From Cognitive to Affective: Assessing Feedback Quality and Scoring Accuracy with a Local LLM in a Canadian Writing Context

Author(s): Johanathan Woodworth

Key words: GenAI assessment, Large Language Models (LLMs), automated grading, automated feedback, educational technology

Abstract: This study explores the integration of local Large Language Models (LLMs) as generative AI tools for assessing open-ended writing assignments at a Canadian university, aligning with the ALTAANZ 2025 theme of balancing innovation and tradition in language assessment. The research addresses three objectives: (1) comparing the alignment between AI-generated and human-assigned grades, (2) evaluating the quality and relevance of summative feedback produced by LLMs, and (3) assessing the reliability of grading outcomes across multiple local models. Graduate-level humanities papers were graded by both human markers and three locally deployed LLMs, with all processes guided by an analytic rubric to ensure consistency and transparency. Statistical analyses (paired t-tests, ANOVA) revealed strong alignment between AI and human grades, though LLM feedback was often generic, less tailored to individual student needs, and overly positive. Inter-model comparisons showed consistent grading but limited variance, highlighting current limitations in AI's ability to capture nuanced or creative responses. Ethical considerations were central, with all data processed on university-secured hardware to safeguard privacy and institutional control. The study also critically examines algorithmic bias and the need for ongoing human oversight, echoing UNESCO's guidelines for ethical AI in education. While local LLMs can streamline grading and support formative feedback, their integration raises important questions about validity, individualization, and the preservation of pedagogical values. The findings inform institutional policy and professional development, offering a framework for responsibly leveraging AI in language assessment while maintaining academic integrity and addressing the evolving landscape of educational technology.

Writing placement tests for an AI era: A multi-site cross-curricular domain analysis of AI expectations on university writing assignments

Author(s): Rebecca Yeager, Rurik Tywoniw, Melissa Meisterheim, Ha Ram Kim

Key words: construct validity, construct evolution, local testing, domain analysis, AI-assisted writing

Abstract: In local testing contexts, domain analysis is essential to ensure that tests sample from language tasks which students will encounter in their future coursework. Knoch and Macqueen (2020) outline a framework for domain analysis by investigating task, language, processing, criteria, and standards specific to a target language use (TLU) domain. In an era of generative AI, we expand upon Knoch and Macqueen's framework by introducing the concept of domain resource analysis to understand the tools that test-takers may access in the TLU domain. We point to earlier work by Seguis and Lim (2020), who justified changes to writing assessment standards for medical workers on the basis of technological changes in the workplace. Our study seeks to apply this framework to better understand the technological resources available to student writers in university contexts.

We report on a survey administered in Fall 2024 across three U.S. universities, asking faculty across disciplines to describe their AI expectations for one or more writing tasks on a scale of 1-5 using the AI Assessment Scale (AIAS; Perkins et al., 2023) or via alternative text. Respondents described AI expectations and contextual information for a total of 64 writing tasks. Descriptive analysis revealed that most writing assignments did not allow AI (55%), with the remainder split across AIAS levels 2-5 (36%) or described textually (9%). Context variables were then entered into a nominal logistic regression predicting no AI, some AI, or alternative text.

Results indicate substantial heterogeneity in AI expectations, with significant differences across institutions, disciplines, and class sizes. Our findings caution that international writers cannot assume unrestricted access to generative AI for

all academic writing tasks, warranting continued assessment of unassisted writing ability on language placement tests. However, ongoing longitudinal and discipline-specific domain resource analysis is necessary to monitor the changing landscape of academic writing.

AI-enhanced dynamic assessment of L2 argumentative writing: Designing responsive mediation to diagnose and promote learner writing development

Author(s): Lu Yu, Matthew E. Poehner, Xiaozheng Dai, Xiaofei Lu, Jingyuan Zhuang

Key words: dynamic assessment, ChatGPT, L2 argumentative writing, Vygotskian sociocultural theory, mediation

Abstract: Grounded in Vygotskian Sociocultural Theory (Vygotsky, 1978), Dynamic Assessment (DA) integrates mediation (e.g., hints, suggestions, modeling, etc.) into the assessment procedures to identify learners' emerging abilities, interpreted according to their responsiveness (Poehner & Wang, 2021). Computerized DA (C-DA) has been pursued as an option to expand the scale of DA beyond one-to-one interactions but faces the challenge of limited flexibility in computerized procedures relative to individualized dialoging (e.g., Randall & Urbanski, 2023). This presentation reports an initial study of the potential for AI-enhanced DA (AI-DA). ChatGPT was trained to 1) identify argumentative elements and assess the quality of arguments according to scoring criteria adapted from prior research (e.g., Kushki et al., 2022; Qin & Karabacak, 2010) and 2) provide graduated mediational moves contingent upon learners' responsiveness (Nassaji et al., 2020). Specifically, for each argumentative element, ChatGPT starts with the most implicit form of mediation by inviting learners to self-assess the quality of their writing. It then progresses to increasingly explicit assistance until finally revealing the issue and offering explanations. We then present analysis of data collected from four sources: 1) drafts independently composed by L2 English learners recruited from a university academic writing program prior to and following DA intervention, 2) two raters' independent scoring of learners' essays, 3) transcripts of learner-ChatGPT interaction, and 4) semi-structured interviews conducted with each learner at the outset and conclusion of the study. Data analysis found that the AI tool was overall successful in identifying argumentative elements, evaluating the quality of argumentation, and providing graduated mediation to guide learners through self-reflection and revisions. Nevertheless, the system still lacks responsiveness to learner moves that characterizes person-to-person interaction, an area of ongoing effort to refine the AI-DA model. The presentation concludes by discussing the implications of administering AI-DA to promote L2 learners' writing development.

Exploring Test Impact in Policy Space: The Impact of Languages Other Than English Subjects in the National Matriculation Test in China

Author(s): Chenyang Zhang

Key words: agency, test impact, languages other than English, National Matriculation Test, language policy

Abstract: Since the promulgation of the Belt & Road Initiative in 2013, proficiency in foreign languages has become a critical factor in fostering China's international collaboration. While the popularity of English in China is undeniable, the importance of languages other than English (LOTE) has gained increasing recognition in recent years. The national policy now encourages secondary students to consider a LOTE subject, alongside English, in the National Matriculation Test (NMT), the most significant high-stakes test in China. However, the impact of LOTE subjects within the NMT has received limited attention.

Adopting a poststructuralist perspective, this study aims to explore how policy actors at various levels exercise their agency, "a discursively mobilised capacity to act" (Miller, 2010, p. 495) - in navigating the impact of LOTE subjects within the NMT. This approach is rooted in the understanding of language policy as a dynamic process of "creation, interpretation, and appropriation" (Johnson, 2013, p. 224) that unfolds across multiple policy levels.

Data were gathered through one-to-one semi-structured interviews with a local testing project coordinator, school administrators (n = 2), LOTE teachers (n = 4), and students (n = 30). The constructivist grounded theory (Tweed & Charmaz, 2011) was utilised to analyse the data. The findings reveal that language tests can yield dynamic outcomes by offering opportunities and limitations for stakeholders to exercise their agency within their sociopolitical space. On one hand, their agency is restricted when a test, as a governance tool, privileges language use. On the other hand, a test and its associated policies can be interpreted in diverse ways, enabling stakeholders to negotiate their agency. Based on the findings, a test impact model was developed, positing that test impact should be understood within the discursive

attribute of the sociopolitical context. Its dynamics and indeterminacy are shaped by stakeholders' agency at different policy levels.

Beyond Scoring: DeepSeek-R1 for Criterion-Specific Feedback Generation in EFL Writing Evaluation

Author(s): Tiancheng Zhang, Shiqi Li

Key words: GenAI, automated writing evaluation, criterion-specific feedback, academic writing, LLM fine-tuning

Abstract: Fine-tuning of large language models (LLMs) refers to the process of further training a pre-trained large language model on a dataset specific to a particular task or domain, in order to optimize its performance for that specific application. Fine-tuning of LLMs has gained significant attention in automated writing evaluation due to their exceptional natural language processing capabilities, efficient text generation, and sophisticated semantic understanding. While empirical studies have demonstrated strong alignment between LLM-generated scores and human raters, existing research predominantly focuses on GPT-series models and lacks in-depth exploration of feedback comment generation. This study addresses this gap by fine-tuning DeepSeek-R1, a high-performance LLM with superior reasoning capabilities, using academic English writing samples and corresponding raters' feedback from undergraduate students at a leading Chinese university. The model subsequently generated both numerical scores and evaluative comments for over a thousand academic essays. Quantitative analysis using percentage agreement, root mean square error, and quadratic weighted kappa, revealed remarkable consistency between DeepSeek-R1's scoring and human experts' ratings. A qualitative evaluation by five domain experts assessed 300 AI-generated comments across five dimensions proposed by Steiss et al. (2024): (a) criterion-based evaluation, (b) clear improvement guidance, (c) linguistic accuracy, (d) identification of key textual features, and (e) supportive tone. Results demonstrated the model's capacity to provide precise, actionable feedback that effectively addresses critical aspects of academic writing. This research extends the application scope of LLM fine-tuning in educational assessment while establishing a multidimensional framework for evaluating automated feedback systems. It has been experimentally applied to essay grading in the university, helping to alleviate teachers' workload and improve the efficiency of providing detailed feedback to students.

Investigating the effects of teacher feedback, peer feedback and AI-generated feedback on students' feedback literacy: A classroom-based study

Author(s): Huijun Zhao, Tianmin Jiang

Key words: students' feedback literacy, teacher feedback, peer feedback, AI-generated feedback, linear mixed-effects model

Abstract: Recent growth in research on feedback has focused on feedback processes, where learners actively generate, interpret, and utilize information for improvement. To maximize benefits from feedback processes, students' feedback literacy, the capabilities to understand, process and use feedback, holds significant potential. To date, there have been few studies on exploring how feedback from different sources can influence students' feedback literacy. This ongoing classroom-based study investigated the effects of teacher feedback, peer feedback and AI-generated feedback on students' feedback literacy. 162 participants in a language university were recruited and divided into three groups, each receiving feedback from teacher, peer and AI, respectively. We utilized a validated student feedback literacy scale to assess students' feedback literacy of participants from the three groups (Teacher, peer and AI) at four time points (Time 0 to Time 3) across 18 weeks in one academic semester. A linear mixed-effects model (LMEM) was conducted to investigate the effects of feedback group, time and the interaction of feedback group and time. Preliminary findings indicated statistically significant differences in students' feedback literacy over time and a significant interaction between time and group, whereas no significant main effect of group was observed. While post hoc comparisons showed no significant group differences before receiving feedback, within-group comparisons across the four time points revealed statistically significant increases in feedback literacy scores over time for all three groups. Scores in the AI-generated feedback group increased significantly at each time point. The peer feedback group demonstrated the most substantial improvement, whereas the teacher feedback group showed limited gains beyond Time 2. The study's limitations and pedagogical implications for feedback practices in the EFL writing classroom are further discussed.