ALTAANZ Conference 2023

Intersections, Crossings and Barriers

Online, 14–16 November 2023

Conference Program

Contents

Welcome

We are very excited to welcome conference attendees from the Australian and New Zealand region, but also from across the world for the 2023 conference. We have people joining us from Auckland, Ames, Brisbane, Christchurch, Canberra, Dubai, Kuala Lumpur, Londrina, London, Hamilton, Jyväskylä, Melbourne, Manoa, Okinawa, Osaka, Sydney, Singapore, Shanghai, St Andrews, St Gallen, Tokyo… the list goes on!

The decision to hold an online conference rather than return to in-person mode was not an easy one, but in the end, we felt an online conference has multiple benefits. ALTAANZ has always strived to be an inclusive and accessible association, with free membership and a free, open-access journal. An online event also enables more people to participate at less cost to them, and to the environment. In this spirit, this conference offers a relatively low-cost forum to share ideas and nurture emerging research and researchers. And, given the audience numbers and submissions, it seems there are plenty of people in the language assessment community who appreciate the chance to get together online. We are especially excited to be able to include so many student papers in this conference.

This year, the field of Applied Linguistics lost a giant. It is impossible to be in the field of language testing and assessment, and not to have encountered the work of Tim McNamara. His rich legacy includes substantial contributions on the topics of validity theory, measurement methods, social theory, fairness and justice, identity, citizenship testing, *lingua franca* testing, LSP and EAP assessment, classroom assessment, proficiency testing, theories of language ability, refugee and asylum seeker language assessment, social policy, gender, subjectivity, poststructuralism and the boundaries of Applied Linguistics itself. In May this year, we approached Tim with the idea of an award in the form of a lecture in his name. He was delighted with our suggestion that Suresh Canagarajah give the inaugural *Tim McNamara Lecture*.

Our theme – *Intersections, Crossings and Barriers* – is intended to capture the breadth of language assessment activity. The 2023 conference showcases the work of practitioners and researchers whose assessment practices or research are at the intersection of language and content, or whose assessment work straddles professional, disciplinary, social or linguistic boundaries that arise in workplaces, schools, institutions, organisations and jurisdictions. Our two plenary speakers, Suresh Canagarajah and Ingrid Piller, are excellent scholarly models who are able to see the critical role of language assessment from beyond the field, bringing us views informed by sociolinguistics, education, multilingualism, migration studies and disability studies.

We wish you many good ideas and fruitful connections arising from this conference. Please join us in the online social gatherings or the in-person hubs and do not be afraid to reach out to each other after the conference.

The ALTAANZ 2023 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, download a membership form from the website and email it to altaanz@gmail.com.

With thanks to…

Conference Organising Committee

Susy Macqueen

Xiaoxiao Kong

Tracey Millin

Maria Treadaway

Viola Lan Wei

Sharon Yahalom

Megan Yucel

Niles Zhao

Leila Zohali

Programming

Ute Knoch

Artwork

Rachel Rogan

Technology

Morena Dias Botelho de Magalhães

Ivy Chen

Best Student Presentation Award Committee

Maria Treadaway (Chair)

Ivy Chen

John Read

Abstract Reviewers

Denise Angelo

Ivy Chen

David Wei Dai

Mark Dawson-Smith

Catherine Elder

Rosemary Erlam

Jason Fan

Catherine Hudson

Naoki Ikeda

Noriko Iwashita

Xiaoxiao Kong

Ute Knoch

Julie Luxton

Susy Macqueen

Morena Dias Botelho de Magalhães

Tracey Millin

Johanna Motteram

Sally O'Hagan

Aek Phakiti

John Read

Carsten Roever

Maria Treadaway

Stephen Walker

Viola Lan Wei

Sharon Yahalom

Megan Yucel

Will Tiancheng Zhang

Niles Zhao

Leila Zohali

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.

Demonstrations

Demonstrations are an opportunity to find out about innovative tools, practices, methods or products for developing, scoring, analysing, delivering or researching language assessments. Demonstrations 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.

Workshops

A workshop is a 60-minute activity-based interactive session aimed at sharing practical/professional skills with the ALTAANZ community. A workshop offers participants hands-on experience of a particular area of *assessment practice* e.g., item writing, portfolio design, or *research practice*, e.g., an analytical method.

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.

TIP! Open the sidebar to see a clickable, alphabetical list of the abstracts.

THEMATIC BLOCKS

Classroom-based assessment
Rating and rubrics
Policy
Writing
Reading
Listening
Vocabulary
Language Assessment Literacy
Technology
Language for Specific Purposes
Other

DAY ONE: Tuesday 14 November

Australian Eastern Daylight Time (AEDT)	SESSION A	SESSION B
9.00am-10.00am	**Suresh Canagarajah Plenary Session: Assessing Crip Translingual Practices** *The Inaugural Tim McNamara Lecture*	
10.00am-10.45am	**Remembering Tim McNamara**	
10.45am-11.00am	Short Break	
11.00am-11.30am	**EFL classes as thinking labs. An approach toward comprehensive assessment** (Student Research Paper) Yomaira Angélica Herreño-Contreras	**Diagnostic writing assessment: Investigating the effects of time limits and understanding students' academic writing self-efficacy beliefs** (Research Paper) Viola Wei
11.30am-12.00pm	**Exploration of assessment practices in community languages schools** (Research Paper) Anna Mikhaylova	**Rethinking assessment for inclusivity. Considerations on academic writing in multilingual contexts** (Work-in-Progress) Ana Maria Benton Z
12.00pm-1.00pm	ALTAANZ AGM All members welcome	
1.00pm-2.00pm	**1.15-1.45pm Online social gathering**	
2.00pm-2.30pm	**Dynamic Assessment of Research Writing of Adult ESL Learners: Designing Rubrics for Content Analysis** (Work-in-Progress) Satuluri Sahana, Lina Mukhopadhyay	**Policy intentions and realities: Assessment of non-common foreign languages in China's secondary education** (Work-in-Progress) Chenyang Zhang
2.30pm-3.00pm	**Analysis of IELTS raters' commentary features of integrated (read-to-write) essays across three success groups of ESL international students** (Research Paper) Aek Phakiti	**Studies informing test equivalency tables – how well are they serving test users?** (Research Paper) Ute Knoch, Jason Fan
3.00pm-3.30pm	**What linguistic features do raters rely on when rating borderline cases?: Empirical examination of Complexity, Accuracy, Fluency of OPI tests** (Research Paper) Myoyoung Kim, Jee Eun Gaetz, Sunkwang Bae	**Linking translation tasks to the Common European Framework of Reference (CEFR): The case of the General English Proficiency Test (GEPT)** (Research Paper) Jason Fan, Ute Knoch, Ivy Chen, Jessica Wu
3.30pm-4.00pm	**Exploring the Sensitivity of the Multifaceted Receptive Vocabulary Assessment Test (MRVAT) in Detecting Changes among Lower-Level English Proficiency Learners** (Research Paper) Hosam Elmetaher	**Examining the relationship between language assessment and institutional policy: The case of high-stakes examinations in Singapore** (Research Paper) Azrifah Zakaria, Vahid Aryadoust
4.00pm-5.00pm	Break	
5.00pm-6.00pm	**Roundtable: The impact of AI on English language assessment** Peter Davidson, Christine Coombe, Barry O'Sullivan	

DAY TWO: Wednesday 15 November

Australian Eastern Daylight Time (AEDT)	SESSION A	SESSION B	SESSION C
8.00am-8.30am	**Nurturing Assessment for Learning through Assessment of teachers: Praxis-pedagogy for teacher/student development** (Student Research Paper) Harsha Dulari Wijessekera	**To what extent does an instructional rubric affect Japanese junior high school students' Eiken writing performance and assessment literacy?** (Research Paper) Chiho Young-Johnson, Noh Kawase, Jerami L. Vanderholm	
8.30am-9.00am	**Who Will Teach the Teacher Educator? Findings and Implications of Promoting Brazilian Teacher Educators' Language Assessment Literacy** (Research Paper) Isadora Teixeira Moraes	**Assessing Digital Multimodal Production: A Mixed-Method Descriptive Rubric** (Student Research Paper) Chia-Hsin Yin	
9.00am-9.30am	**Assessment for Engagement: Designing for and analysing levels of engagement in assessment activities** (Research Paper) Susy Macqueen, Jinxiao Xie, Neha Jagannath	**Blooket used for formatively assessing students' understanding of the mark scheme for IGCSE 0500 English Descriptive Writing Paper 2** (Demonstration) Louise Finucane	
9.30am-10.00am	Short Break		
10.00am-11.00am	**Student networking session**		
11.00am-11.30am	**Developing the Little Kids Word List app – a MacArthur Bates Communicative Development Inventory (CDI) for Aboriginal languages in central Australia** (Research Paper) Carmel O'Shannessy, Vanessa Davis, Alice Nelson	**21 years supporting students with their academic English language needs - Is DELNA still fit for purpose?** (Work-in-Progress) Morena Botelho de Magalhaes, Rosemary Erlam	**A Reliability Generalization Meta-analysis of L2 Reading Comprehension Assessments** (Research Paper) Huijun Zhao, Vahid Aryadoust
11.30am-12.00pm	**What are the Barriers to Developing English-as-a-Second/Foreign-Language (L2) Learners' Metaphor Awareness? Evidence from the Development and Validation of Metaphor Awareness Instruments** (Research Paper) Ting Ma, Lawrence Jun Zhang, Judy Parr	**Standardised tests of English proficiency of international students: Evaluating the state of play** (Research Paper) John Read	**Comparing register variations in ChatGPT-generated texts, natural texts, and reading assessment texts** (Work-in-Progress) Zhang Wenxin, Vahid Aryadoust, Azrifah Zakaria
12.00pm-1.00pm	Long Break		

DAY TWO: Wednesday 15 November continued

1.00pm-1.30pm	**Student ethical considerations on the use of language assistance tools for assessed academic writing** (Research Paper) Elpida Petraki, Averil Grieve, Amir Rouhshad, Alan Bechaz, David Wei Dai	**Examining lexical sophistication, diversity features and business vocabulary usage in business English learners' writing performance** (Student research paper) Yuhu Zou	**A Corpus-Based Frame Semantic Analysis of Commercialized Listening Tests: Implications for Content Validity** (Research Paper) Zhao Yufan, Vahid Aryadoust
1.30pm-2.00pm	**Effect of Peer Feedback on the Accuracy of Peer Assessment of ESL Argumentative Writing** (Student Research Paper) Xiao Xie	**Investigating test constructs for assessing EMI-readiness of content lecturers in Thai international medical program**s (Work-in-Progress) Teaka Sowaprux, Jirada Wudthayagorn, Thanakorn Jirasevijinda	**A reliability generalization meta-analysis of metacognitive awareness measures in second language listening** (Research Paper) Jiayu Zhai, Vahid Aryadoust
2.00pm-2.30pm	**ChatGPT (3.5-turbo) and Writing papers of Cambridge English examinations** (Student Research Paper) Daniil M. Ozernyi	**Validation of language self-assessment descriptors for use with the Defence Force** (Work-in-Progress) Ksenia Zhao	
2.30pm-3.00pm	**Unpacking the Drivers of Citation Counts in Language Assessment Research: A Bibliometric Study** (Research Paper) Zhang Sai, Vahid Aryadoust	**Plain or precise: How writers do "readability" in health information for multiple audiences** (Student Research Paper) Jeanie Henchman	
3.00pm-5.00pm	Long Break (possible local gathering) **3.30-4.30 Online Social Gathering**		
5.00pm-5.30pm	**Language assessment literacy: Teachers' attitudes in ensuring fair classroom assessment practices** (Student Research Paper) Karim Rezagah	**Indigenous criteria and differential rater behaviour: On the challenges of assessing the complex LSP construct of Teacher Language Competence** (Research Paper) Olivia Rütti-Joy	
5.30pm-6.00pm	**Opening the black box of experience in language assessment literacy (LAL) research: A sociocultural perspective** (Research Paper) Xuan Minh Ngo	**Exploring the language demands of early childhood and secondary teachers in Australia: Implications for language assessment for teacher registration** (Work-in-Progress) Xiaoxiao Kong	

DAY THREE: Thursday 16 November

Australian Eastern Daylight Time (AEDT)	SESSION A	SESSION B
8.00am-8.30am	**Test takers' attitudes towards at-home testing during and post pandemic** (Work-in-Progress) Jieun Kim	**The relationship between writing tasks and second language writers' written stance** (Work-in-Progress) Giang Tran
8.30am-9.00am	**Leveraging Generative AI for Enhanced Content Development in Language Testing: Implications for Item Writer Recruitment, Training, and Engagement** (Research Paper) Michelle Y. Chen, Jennifer J. Flasko	**The power of testers and their tests: A sociological analysis of assessment practices in Australian Direct Entry Programs** (Student Research Paper) Kyle Smith
9.00am-9.30am	**Redefining Language Assessment in the wake of AI invasion and technological innovations in language testing** (Student Research Paper) Mohammad Haseen Ahmed	**Oral Recall as Assessment of Reading Comprehension for Intermediate-level Chinese as Second Language Learners** (Research Paper) Shuyi Yang
9.30am-10.00am	**Interaction matters: Automated assessment of interactive features in paired speaking tasks** (Student Research Paper) Rena (Wei) Gao	**Measuring L2 reading assessment processes via user experience** (Demonstration) Sarah Goodwin
10.00am-10.30am	**Developing a screening test of transcription ability to support the creation of reliable transcripts of indistinct forensic audio** (Research Paper) Ute Knoch, Helen Fraser, Debbie Loakes Jason Fan, Ivy Chen	**Demonstrating the utility of the Japanese-Vocabulary Levels Test to support course-integrated extensive reading in Japanese as a foreign language** (Demonstration) Kimberley Rothville, Michiyo Mori
10.30am-11.00am	Short Break	
11.00am-12.00pm	**Ingrid Piller Plenary Session: Language Testing for University Admission**	
12.00pm-2.00pm	Long Break (possible local gathering)	
2.00pm-2.30pm	**Exploring the relationship between spoken lexical diversity scores and human rater vocabulary scores (Work-in-Progress)** Philip Head, Thwin Myint Maw	**A Corpus Linguistic Study of High-Stakes English Listening Tests: A Multidimensional Analysis of Gaokao in China** (Student Research Paper) Tao Xuelian
2.30pm-3.00pm	**Collocation use in a Japanese university-wide test of spoken English: Differences across proficiency levels** (Research Paper) Ivy Chen, Katsunori Kanzawa	**A micro-level analysis of L2 test-takers' construct-relevant and irrelevant strategic processes in their IELTS listening performance** (Research Paper) Nick Zhiwei Bi, Yue Wang
3.00pm-3.30pm	**The State of Classroom-based Assessment in Japan** (Work-in-Progress) Adam Murray, Taiko Tsuchihira	**A Computer-mediated Dynamic Assessment Approach to Feedback and Revisions in EAP Academic Writing** (Work-in-Progress) Adam Steinhoff
3.30pm-4.00pm	**An ecological view of the assessment of L2 Japanese in Australian secondary schools: Through the lenses of teachers and learners** (Work-in-Progress) Fusae Nojima	**Language proficiency and wages in Korea** (Work-in-Progress) Junghyun Baik
4.00pm-5.10pm	**Workshop: Collaboration and Craft: Insights into Writing Items for English Language Tests** Vicki Bos, Sophia Walker, Megan Yucel -- *Student Best Paper Award Presentation*	

ALTAANZ ALTAANZ Conference 2023

Plenary One: The Inaugural Tim McNamara Lecture

Assessing Crip Translingual Practices

Suresh Canagarajah, Pennsylvania State University

9.00 am-10.00am (AEDT) Tuesday 14 November

Abstract: "Crip" from disability studies draws from connotations of fracture, immobility, and constraints to highlight how such conditions of vulnerability can be resourceful and generative of communication and knowledge. In this talk, I present interactions from both disabled and "non-disabled" contexts to demonstrate how vulnerability is at the heart of all communication. Drawing from the expansive communicative strategies and semiotic resources adopted for meaning-making in these contexts, I argue for reorienting language analyses and pedagogies along constructs introduced by decolonization and disability studies. The talk will illustrate the challenges applied linguists will face in assessing meanings as emergent, distributed, and nonnormative.

Suresh Canagarajah is the Edwin Erle Sparks Professor of English, Applied Linguistics, and Asian Studies at Pennsylvania State University. He teaches courses in Global Englishes, Multilingual Academic Writing, Sociolinguistics, and Decolonization Studies. He taught earlier in the University of Jaffna, Sri Lanka, and the City University of New York. He was formerly the editor of the TESOL Quarterly and the President of the American Association of Applied Linguistics. His Routledge Handbook on Language and Migration (Routledge, 2017) won the 2020 best book award from the American Association of Applied Linguistics. His latest publication is Language Incompetence: Learning to Communicate through Cancer, Disability, and Anomalous Embodiment (Routledge, 2022).

The **Tim McNamara Lecture** honours the distinguished contribution Tim McNamara made to the fields of Language Testing and Assessment and to Applied Linguistics more generally. The awardee is someone who can make a significant contribution to the conference program, and need not be from within the field of language testing and assessment. The ALTAANZ 2023 Conference hosts the inaugural *Tim McNamara Lecture*.

Plenary Two

Language Testing for University Admission

Ingrid Piller, Macquarie University

11.00 am-12.00pm (AEDT), Thursday 16 November

Abstract: English language proficiency (ELP) is central to the academic achievement of the 1.5 million students enrolled in Australian universities each year. Yet, students are highly linguistically diverse, with a mix of domestic students from English- and non-English-speaking backgrounds and international students from national contexts where English may be the main language, an official language in a multilingual context, or a foreign language with limited communicative functions. How do universities manage students' linguistic diversity through their admission requirements and set students up for success? In this talk, I examine ELP requirements for university admission in Go8 universities to answer this question. Our language ideological analysis found two categorically different constructs of ELP: inherent ELP based on citizenship, linguistic heritage, and prior education, and tested ELP. These two different conceptualizations of ELP map onto two dichotomous student groups. One of these is deemed to naturally speak English while the other is constructed as deficient and subject to perpetual scrutiny. These language ideological constructs frame ELP as a matter of individual responsibility rather than part of embedded in learning processes. Conversely, they obscure the need for continuous language development of all students and the need for pedagogical innovation in linguistically diverse educational institutions.

Ingrid Piller is Distinguished Professor of Applied Linguistics at Macquarie University, Sydney, where she previously served as Executive Director of the Adult Migrant English Program Research Centre (AMEP RC). Over the course of her international career, she has also held appointments at universities in Germany, Switzerland, United Arab Emirates and USA. She is a Fellow of the Australian Academy of the Humanities and recipient of a 2018 Anneliese Maier Research Award. Ingrid Piller is an applied sociolinguist with research expertise in intercultural communication, language learning, multilingualism, and bilingual education. She has published, lectured and consulted widely in these areas. Ingrid Piller is the author of Linguistic Diversity and Social Justice (Oxford University Press, 2016), which won the 2017 Prose Award in the Language and Linguistics category and the 2017 BAAL Book Prize. She is also the author of the bestselling Intercultural Communication *(Edinburgh University Press, 2nd ed., 2017) and over 400 other publications. Ingrid Piller is a member of the Australian Research Council (ARC) College of Experts, served as editor-in-chief of the international sociolinguistics journal* Multilingua *(De Gruyter Mouton; 2013-2022) and edits the sociolinguistics portal Language on the Move, through which many of her publications and those of her team, including their research blog, can be accessed. She tweets about linguistic diversity @lg_on_the_move.*

References

Piller, I. (2023). How do universities decide whose English needs to be tested for admission? *Language on the Move*. https://www.languageonthemove.com/how-do-universities-decide-whose-english-needs-to-be-tested-for-admission/

Piller, I., & Bodis, A. (2022). Marking and unmarking the (non)native speaker through English language proficiency requirements for university admission. *Language in Society*, 1-23. https://doi.org/10.1017/S0047404522000689

Roundtable

The impact of AI on English language assessment

Peter Davidson, Christine Coombe, Barry O'Sullivan

5.00pm- 6.00pm (AEDT), Tuesday 14 November

Abstract: Generative AI-powered conversational interfaces such as ChatGPT have the potential to revolutionize teaching, learning, and assessment. Recent developments in AI, and in particular AI-gerarative tools, have resulted in much hype and hysteria around the use and abuse of AI by students in educational settings. While some countries, educational institutions, and teachers, have sought to ban the use of GAI, others recognized the incredible potential that it offers to students, teachers, and test writers. In this roundtable discussion we would like to discuss the impact that GAI is having upon English language assessment, from the classroom tests, to national tests, to international language tests.

GAI is having a significant impact on assessment, in the classroom and at international testing body level. It is timely, therefore, to have a discussion on how teachers, educational institutions, and testing bodies are reacting to GAI, and to discuss how it is impacting on assessment practices now, and how it may shape assessment practices in the future. The first speaker, Peter Davidson, will outline the concerns that teachers have expressed about GAI in educational settings. He will then discuss how GAI is used by students and teachers in the classroom, and how this has impacted on classroom assessment. He will also suggest how assessments can be devised that make it difficult for students to simply rely on GAI tools to generate a complete response. The second speaker, Christine Coombe, will outline how GAI can be leveraged by classroom teachers to develop assessments by identifying and adapting source texts, writing source texts, generating questions, writing rubrics, delivering tests, rating tests, analysing test items, and by providing students with feedback. The third speaker, Barry O'Sullivan, will outline how GAI has impacted the development of an English language assessment system, touching on how it contributes to areas as diverse as machine-assisted text and task generation, task design and delivery, and automated scoring and reporting. Benefits to be addressed include efficiency, personalization, and agility of design and development, while limitations include originality and copyright issues, system hallucination, transparency of decisions, appropriateness of scoring and feedback, and bias. The final area to be discussed is that of construct definition and representation.

Collaboration and Craft: Insights into Writing Items for English Language Tests

Vicki Bos, Sophia Walker, Megan Yucel

4.00 pm- 5.00 pm (AEDT), Thursday 16 November

Abstract: In this workshop, participants will delve into the art and science of item writing for English language tests, focusing on assessing reading and listening skills. The aim is to provide teachers with practical tips that they can readily apply to their item development processes. Through interactive activities and collaborative discussions, participants will gain a deeper understanding of the crucial aspects of item writing for receptive skills assessment. This workshop is tailored for English language teachers, curriculum developers, and assessment practitioners seeking practical tips for item writing in English language tests, particularly those assessing reading and listening skills. Participants should have a basic understanding of language assessment principles and familiarity with English language testing.

Workshop Objectives:

- Introduce participants to the fundamental principles and guidelines for item writing in receptive skills assessment.
- Explore common item types used in assessing reading and listening skills.
- Provide practical tips and strategies for constructing effective test items within a limited timeframe.
- Encourage collaborative learning through group activities and discussions.

Image by StartupStockPhotos from Pixabay

The presenters, Vicki Bos, Sophia Walker and Megan Yucel**,** bring a wealth of test development experience to the workshop, having all written items for a range of large-scale national and international examinations, assessing a variety of skills, including listening, reading, writing, speaking, and use of English.

Student Networking Session

Unlocking Opportunities: A Graduate Student’s Guide to Research Award Application

Organisers: Niles Zhao, Tiancheng (Will) Zhang

10.00 am- 11.00 am (AEDT), Wednesday 15 November

Contact: niles.zhao@unimelb.edu.au

Abstract: Research awards and grants are pivotal instruments for acquiring essential funding, resources, and academic recognition. However, the process of applying for such funding is not always straightforward and researchers who are interested in obtaining this funding can be easily intimidated by the requirements and the procedures. This is especially true for graduate students who have less experience.

Image by Gordon Johnson from Pixabay

In this informative session, two guest speakers, who are graduate students themselves, will demystify the grant application process and empower participants to secure the support they need. Participants will gain valuable insights into the world of research awards, learn how to identify the right opportunities, and discover the strategies to craft compelling applications that stand out. Following the talks, there will be a Q & A session and opportunities to networks with other graduate students.

Information about speakers

Rena Gao *is a PhD candidate from the University of Melbourne supervised by Prof. Carsten Rover and Dr. Jey Han Lau. Her research focuses on using an automated way to score the interactive features in paired speaking assessment. Her prior work includes working on scoring rubric for interactional features in paired speaking assessment and using Natural Language Processing techniques to improve the realizability of automated scoring in speaking assessment for the English language Rena is a recipient of the 2022 Duolingo English Test Doctoral Dissertation Award.*

Shengkai Yin *is a joint PhD candidate from Shanghai Jiao Tong University and The University of Melbourne, supervised by Prof. Yan Jin, A/prof. Jason Fan and Prof. Ute Knoch. He has published research related to peer assessment, speaking assessment, and critical thinking ability. His current research focuses on rating scale development and validation. Shengkai is a recipient of the 2022 Duolingo English Test Doctoral Dissertation Award and 2023 British Council Assessment Research Award.*

Social opportunities

Join us in-person or online for some social networking!

Image by Engin Akyurt from Pixabay

Online networking sessions

Tuesday 14th November 1.15-1.45pm AEDT

Wednesday 15th November 3.30-4.30pm AEDT

In person gatherings

If you are in one of the following cities, please get in touch with the contact person to find out what is going on.

Auckland contact Morena Dias Botelho de Magalhaes m.magalhaes@auckland.ac.nz

Christchurch contact Tracey Millin tracey.millin@canterbury.ac.nz

Melbourne contact Annemiek Huisman ltrc-info@unimelb.edu.au

Brisbane contact Megan Yucel meganyucel@gmail.com

Canberra contact Susy Macqueen susy.macqueen@anu.edu.au

Image by StockSnap from Pixabay

Abstracts (Open sidebar for clickable, alphabetical listing by title)

A reliability generalization meta-analysis of metacognitive awareness measures in second language listening

Author(s): Jiayu Zhai, Vahid Aryadoust

Key words: L2 listening, metacognitive awareness, reliability generalization (RG)

Abstract: Second language (L2) metacognitive awareness in listening has been predominantly assessed by a multidimensional instrument named the metacognitive awareness listening questionnaire (MALQ). Nevertheless, there is yet no study examining the reliability of the MALQ measures from a meta-analytical perspective. The purpose of the present study was to examine variability in the reliability of MALQ measures in previous research in the field of L2 listening. A reliability generalization (RG) meta-analysis was conducted using Cronbach's alpha estimates obtained from 35 studies using MALQ. The results showed that the estimated average reliability was 0.80 (95% CI: 0.78 to 0.82) for the MALQ measures. Specifically, the cumulative reliability coefficients of the five dimensions were: 0.64 for directed attention, 0.72 for mental translation, 0.73 for planning and evaluation, 0.71 for person knowledge, and 0.77 for problem solving. We further found evidence for reliability induction, heterogeneity, and publication bias. To investigate the possible causes of the observed heterogeneity, a meta-regression was performed to explore the effect of mediators on the reliability coefficients. The results revealed that studies conducted in the pre-tertiary educational setting, or the studies in which the first language (L1) of the participants were east Asian languages, or those that were published more recently tended to report higher reliability coefficients. Based on these findings, suggestions for future research are provided.

Comparing register variations in ChatGPT-generated texts, natural texts, and reading assessment texts

Author(s): Zhang Wenxin, Vahid Aryadoust, Azrifah Zakaria

Key words: generative AI, language assessment, multidimensional analysis

Abstract: This ongoing study aims to compare the register features of three corpora of texts to provide a comprehensive understanding of the strengths and limitations of AI-generated texts in the context of language assessment. The first corpus consists of texts written by ChatGPT, a generative Artificial Intelligence (AI). Generative AI systems can create original and coherent content based on learned patterns and examples from extensive training data. In this study, we utilized ChatGPT's language modeling abilities to generate a sample corpus of reading comprehension texts for analysis. Another sample corpus of reading comprehension assessment texts was constructed from standardized proficiency tests of English. A third corpus was derived from a standard large corpus of naturally-occurring texts. We adopted a Multi-Dimensional Analysis (MDA) approach to compare and contrast the linguistic and content aspects of the texts from the three corpora. The texts were first tagged by using a tagger and the tags were submitted to MDA analysis, which applies exploratory factor analysis to reduce the linguistic features into a parsimonious set of factors or latent variables. The use of MDA allowed us to conduct a systematic comparison and examination of multiple dimensions across the three corpora to determine whether the texts in the three corpora demonstrate significant register variations. Through the exploration of linguistic features and content in the text sources, the findings of this research will contribute to suggestions for utilizing generative AI in reading assessment.

21 years supporting students with their academic English language needs - Is DELNA still fit for purpose?

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

Key words: PELA, validity, academic literacy

Abstract: At the University of Auckland, first-year undergraduate students, doctoral candidates, and some postgraduate cohorts are required to complete DELNA (Diagnostic English Language Needs Assessment) as they begin their studies. DELNA has been in operation since 2002, with between 8,000 and 10,000 assessments administered each year. The development of DELNA and how its delivery evolved to adapt to stakeholders' needs has been well documented (e.g., Elder, 2003; Elder & von Randow, 2008; Erlam, von Randow, & Read, 2013; Doe, 2014; Read, 2015, 2017; Read & von Randow, 2013, 2016). Yet, the need for evidence that DELNA remains fit for purpose continues. Knoch and Elder's (2013) framework against which post-entry language assessments can be evaluated provides the basis for building an argument for the validity of DELNA. A study investigating the assessment programme in its current context is ongoing, with the initial phase already reported on (Erlam & Botelho de Magalhães, 2021). In the study's second phase, as well as analysing students' scores in the DELNA assessments versus students' academic performance (evidenced by passing grades and cumulative GPAs), the DELNA team wishes to include students' perspectives in the research. Interviews were conducted with some students at the end of their first year of study with a second and a third interview envisaged to take place in subsequent years. While the first interviews focused on students' experiences of taking DELNA and accessing support, the next interviews will aim to gain more understanding of students' academic literacy development as they progress in their degree. However, moving forward with the project has been challenging. There are practical difficulties (e.g., recruiting participants), an eminent external review of DELNA, and questions arising around the obligatory assessment of all students, including those who choose to complete their degree in Te Reo Māori, for example.

A Computer-mediated Dynamic Assessment Approach to Feedback and Revisions in EAP Academic Writing

Author(s): Adam Steinhoff

Key words: Dynamic assessment, Computer-mediated assessment, Revisions to academic writing

Abstract: At the core of research into Dynamic Assessment (DA) is Vygotsky's (1978) Sociocultural Theory and the use of working within learners' Zones of Proximal Development (ZPD). By working within the ZPD, teachers not only provide feedback to learners, but can also gain deeper insights into their learners' potential language abilities. Within the area of DA are calls for research into how DA approaches can be incorporated into classroom practice. The current study explored the use of online DA, using Zoom, to provide feedback and allow for immediate revisions to EAP learners' academic writing. Five students from an EAP programme at a university in Australia volunteered to participate in the study. The study addresses the following research questions in relation to the academic writing of ESL international EAP students: (1) What are the sociocultural processes during one-to-one computer-mediated DA feedback sessions? (2) What revisions are made during one-to-one computer-mediated DA feedback sessions? (3) How do the sociocultural processes influence the revisions that are made during one-to-one computer-mediated feedback sessions? Qualitative data was collected to answer the three research questions. It is believed that the results of the study will shed light on how teachers can use DA in the EAP classroom to complement summative assessment.

A Corpus Linguistic Study of High-Stakes English Listening Tests: A Multidimensional Analysis of Gaokao in China

Author(s): Tao Xuelian

Key words: content validity, Chinese Gaokao listening tests, multidimensional analysis

Abstract: Gao Kao is a high-stakes exam for all Chinese students that can have a life-changing impact. Since listening has been a part of the Gao Kao English exam's curriculum for more than 20 years, numerous research have looked at the listening test's content validity. Topics, readability, duration, pace, etc. were all thoroughly investigated. With the help of multidimensional analysis (MDA), the current study aims to investigate the Gao Kao English listening tests from a novel angle: the linguistic variation found in listening transcripts from various text genres, geographic locations, and time periods. Based on factor analysis, MDA is a methodological approach that identifies co-occurrence patterns of linguistic features. A pattern where a text with a high frequency of private verbs is likely to also have a high frequency of the "that" deletion and a low frequency of nouns and prepositions is one example of a dimension. Some studies of multidimensional analysis of the corpus of spoken texts in listening tests in China such as CET 4 indicate that there are similarities and differences among short conversations, long dialogues, passages, and dictations. The current study investigates a corpus of 160 listening tests comprising the national unified test papers and autonomous local test papers developed in different provinces and cities in China. We compare these tests based on their text types and across regions and time periods. The preliminary results indicated that there were significant and meaningful differences across short conversations, long dialogues, and monologues in six main dimensions that emerged in MDA. In addition, region and text type were associated with dimension scores of Gaokao listening texts, while time or any interactions between these three factors were not. Implications of using corpus linguistics and MDA for content validity are discussed.

A Corpus-Based Frame Semantic Analysis of Commercialized Listening Tests: Implications for Content Validity

Author(s): Zhao Yufan, Vahid Aryadoust

Key words: Frame semantics, Corpus analysis, Content Validity

Abstract: This study applied an automatized frame semantic method to investigate how knowledge has been organized in simulated lectures in the listening sections of the International English Language Testing System (IELTS) and the Test of English as a Foreign Language (TOEFL). Frame semantics is a linguistic theory to investigate how meaning is structured in language. We constructed two corpora from the lecture scripts of 68 simulated lectures in IELTS and 285 in TOEFL. In addition, the lectures selected from the Michigan Corpus of Academic Spoken English (MICASE) served as the reference corpus of the study. The data was submitted to automatized semantic tagging, which generated 488 semantic frames. Subsequently, we conducted three comparisons: IELTS vs. TOEFL, IELTS vs. MICASE lectures, and TOEFL vs. MICASE lectures. The resulting outputs were sorted by Bayes factor, and visualized using bar graphs and tables displaying the significant frequency differences of semantic categories across various discourse fields. The findings indicate significant differences in the use of semantic categories among IELTS, TOEFL, and MICASE lectures. Each lecture corpus exhibited distinct levels of conceptual use across different discourse fields. For example, IELTS and TOEFL display distinct category preferences, with IELTS featuring a significantly higher percentage of semantic categories related to sports, places, business, farming and horticulture, education, and mental actions and processes. On the other hand, the TOEFL lectures covered a greater range of semantic categories pertaining to arts, the universe, plants, substances and materials, and temperature. Additionally, the IELTS lectures include other scenarios such as program introduction and health-related lectures, while the TOEFL lectures emphasize semantic categories that are exclusive to academic target language use domains. This study is the first to apply frame semantics to language assessment, providing insights into the differences in semantic features between these widely used commercialized language tests and their content validity.

A micro-level analysis of L2 test-takers' construct-relevant and irrelevant strategic processes in their IELTS listening performance

Author(s): Nick Bi, Yue Wang

Key words: strategic processes, construct-relevant and irrelevant strategies, cognitive validity

Abstract: The strategic processes that contribute to construct-relevant variations in test results (such as metacognitive and cognitive strategies) and construct-irrelevant variances (such as test-wiseness strategies) have been identified in both practice and theories. Few studies, however, have attempted to incorporate construct-relevant and -irrelevant strategic processes into a more comprehensive conceptual model. This study probed into the complicated relationships between construct-relevant strategies (i.e., metacognitive and cognitive strategies), construct-irrelevant strategies (i.e., test-wiseness strategies), and listening performance from a micro perspective. A total of 473 Chinese EFL undergraduates with various majors were recruited to take an IELTS listening test and complete the metacognitive and cognitive strategy questionnaire (MCSQ) and test-wiseness strategy questionnaire (TWSQ). This study identified three metacognitive strategies (planning, monitoring and evaluating) and four cognitive strategies (memory, voice and imagery inference, retrieval and comprehending) as construct-relevant processes. By contrast, test-wiseness strategies, as a unitary factor, encompassed three dimensions (deductive reasoning, test format and cue-using). The structural equation model (SEM) results revealed a multi-directional relationship among metacognitive strategies and a uni-directional relationship among cognitive strategies. And the micro-level analyses show that (1) planning strategies directly regulated memory, comprehending and test-wiseness strategies; (2) monitoring strategies had an executive function on comprehending strategies, whereas evaluating strategies did not directly regulate any types of strategies; (3) memory, voice and retrieval strategies assisted test-takers' listening performance via comprehending strategies; (4) only comprehending strategies had a direct impact on listening performance, while all metacognitive strategies were not found to directly influence the listening score, and test-wiseness strategies had a negative but insignificant effect on the listening test performance. The findings corroborate that the IELTS listening test has high cognitive validity. The study also implies that language test preparation programmes should place less emphasis on teaching test-wiseness strategy and more on construct-relevant strategic processes and language knowledge.

A Reliability Generalization Meta-analysis of L2 Reading Comprehension Assessments

Author(s): Huijun Zhao, Vahid Aryadoust

Key words: Reliability Generalization, Meta-analysis, L2 Reading Comprehension Assessments

Abstract: Scoring reliability, among other facets, determines the internal validity of language assessment. However, not all language assessment researchers consistently apply or report reliability coefficients in their studies. To emphasize the need for more well-designed quantitative research, which incorporates factors affecting reliability and advocates for robust reliability reporting practices, this ongoing study conducted a reliability generalization meta-analysis of L2 reading comprehension tests. We examined 1883 individual studies from Scopus, the Web of Science, ERIC, and LLBA databases for possible inclusion and assessed 255 studies as eligible for our inclusion criteria. Within these 255 studies, certain malpractices in reliability reports were detected, including no reliability report for tests (38.4%), instances of inducted reliability (7.5%), and reporting reliability ranges for tests (3.1%). We further scrutinized 97 Cronbach's alpha estimates from 72 studies that reported Cronbach's alpha estimates properly and coded 25 potential predictors comprising of the characteristics of the study, the test, and test-takers. Our analysis showed an average reliability coefficient of 0.786 in the selected studies, which was significantly lower than that of the studies using inducted reliability coefficients (α=0.893). A Q-test showed significant heterogeneity among the selected studies. Subsequently, a moderator analysis revealed that participants' age mean, study design, number of test items, test purpose, text genre, and text length could explain shares of variance in the reliability coefficients across the studies. The recognition of these moderators in reliability warrants a more nuanced approach to reporting and interpreting reliability. Thus, rather than establishing rigid range criteria for reliability, it is important that we acknowledge the multifaceted nature and inherent complexity of reliability and revisit the practice of adopting rigid standards. Implications of this proposed approach for investigating validity of reading tests is further discussed.

An ecological view of the assessment of L2 Japanese in Australian secondary schools: Through the lenses of teachers and learners

Author(s): Fusae Nojima

Key words: L2 Japanese language learning, Learning-Oriented Assessment (LOA) , Language Assessment Literacy (LAL)

Abstract: The importance of classroom assessment as an integral part of teaching and learning has attracted significant attention in L2 language education. Built on the earlier conceptualisation of assessment for learning by Black and Wiliam (1998a), Learning-Oriented Assessment (LOA), has been theorised to provide a coherent system combining different uses of assessment to promote learning (Turner & Purpura, 2016; Leung, 2020; Saville, 2021). Efforts have often been made to understand LOA from teachers' perspectives. This has highlighted the importance of teachers' knowledge, skills and principles to use assessment to support learning ('Language Assessment Literacy' - LAL). However, further efforts to understand the interface between the curriculum from both teachers' and learners' perceptions of classroom assessment within the LOA framework may result in a greater understanding of such assessment. In Australia, the content of classroom assessment is guided by the standards laid out in the Australian Curriculum (ACARA). Teachers develop learning goals based on achievement standards and content descriptions in the curriculum; they develop tasks to achieve the goals; they support learners to achieve the goals; and they assess the outcomes of how much has been achieved. In the context of Japanese language classrooms, this study proposes to explore teachers' assessment design processes to glean insights on the nature of how 'assessment for learning' is operationalised. The implementation of these tasks will then be examined through both teachers' and learners' perspectives, paying particular attention to how assessment is enacted in local contexts. A participatory approach will provide a way for the researcher and Japanese language teachers in secondary schools to work together. Interviews will be conducted with teachers and students and assessment artifacts will be collected. The content analysis will be used to analyse the qualitative data. It is hoped that this research will contribute to more meaningful assessment, deeper learner engagement and professional development for teachers.

Analysis of IELTS raters' commentary features of integrated (read-to-write) essays across three success groups of ESL international students

Author(s): Aek Phakiti

Key words: Assessing integrated writing, Raters' comments, Features of integrated essays

Abstract: This presentation reports on part of a larger study that examines the validity and feasibility of including an integrated academic writing task as part of the IELTS test. 154 ESL international students participated in the current study by completing an online independent writing task (Task 2 in the current IELTS Writing), followed by an integrated writing task (reading-to-write). This presentation focuses on unique and distinctive features of written comments made by two accredited IELTS raters on essay responses of three success levels (Level 1 = low; Level 2 = Mediocre; Level 3 = high). Rater 1 provided all 154 comments (100%), whereas Rater 2 provided 124 comments (80.5%). The current study adopted the qualitative analysis approach taken by Miles et al. (2014). It was found that there were some overlapping features of essays across the three groups, as well as distinctive features. For example, Level 1 essays tended to be underdeveloped and under-length essays that stated unclear positions or arguments, either underused or overused the source texts, and did not acknowledge the source texts. Level 2 essays had similar features to Level 1 essays (e.g., over-reliant on the source texts in essays, underdeveloped essays). Positive features of Level 2 essays include clear positioning essays with relevant source texts. Although Level 3 essays exhibited a few similar issues to Levels 1 and 2 essays (e.g., under development, improper citations), they demonstrated the use of relevant and accurate source text use and integrated source texts with students' own knowledge). The study has allowed us to understand the similarities and differences between various features of essays produced by different groups of test takers. It has provided empirical evidence for understanding sources of score variations in an integrated essay task. Implications for assessing integrated writing will be discussed.

Assessing Digital Multimodal Production: A Mixed-Method Descriptive Rubric

Author(s): Chia-Hsin Yin

Key words: Multimodal Assessment, Digital Multimodal Composition, Digital Storytelling

Abstract: In this theoretical research, we propose a triangulated approach to evaluate digital multimodal compositions (DMC) projects in K-12 EFL and ESL contexts. As storytelling projects offer ELLs many opportunities to share their experiences through narratives and avenues for relating with others (Kovach, 2018), digital storytelling (DST) projects are a popular type of DMC production. There have been multiple efforts to apply different frameworks in the assessment of both the production processes and final products of DMC practices of ELLs (e.g., Hung, et al., 2022), however, the resulting assessment rubrics have been predominantly generated for researcher purposes rather than in-practice teachers. As such, given that current process-oriented DMC assessment rubrics are complicated for in-practice teachers, our mixed-method descriptive rubric provides teacher-accessible evaluation tools for DSTs. Specifically, we incorporate (1) Systematic Functional Linguistics' (SFL) Appraisal Framework (Ngo & Unsworth, 2015; Unsworth & Mills, 2020), (2) Rose's Sites of Visual Meaning-Making (Kendrick et al., 2022; Rose, 2016), and (3) Language Complexity (Bardovi-Harlig, 1992; Zheng & Warschauer, 2018). That is, SFL and its Appraisal Framework enable evaluators to understand DMCs through (1) the use of language within specific contexts and (2) the efficacy of expressing affect, judgment, and appreciation through the attitude system's metalanguage. Next, this study used two of Rose's (2016) sites for visual meaning-making, the sites of image (the DST itself) and audience (the preferred reading of the story). Lastly, this study employs the sentential coordination index, which measures language complexity and can be used to track linguistic development over time. The proposed assessment rubric suggests a proportional relationship between the aim of the DST (e.g., ELLs' intentions or teacher's instructions) and the three frameworks. While more evidence for generalizability is needed, this study seeks to take steps toward DMC assessment validation. Research and pedagogical implications will be discussed.

Assessment for Engagement: Designing for and analysing levels of engagement in assessment activities

Author(s): Susy Macqueen, Jinxiao Xie, Neha Jagannath

Key words: Classroom-based assessment, Assessment for Learning, L2 engagement

Abstract: It is assumed that assessing students will serve the dual purposes of informing teaching and driving learning. Over the years, various conceptualisations have emerged which connect learning with assessment, for example, "assessment for learning" (Broadfoot et al., 1999), "dynamic assessment" (Lantolf & Poehner, 2011) and "learning-oriented assessment" (Turner & Purpura, 2016). These approaches have drawn attention, not just to the integral role of assessment in effective teaching and learning, but to the centrality of learner action and self-regulation. At the same time, L2 motivation (Mercer & Dörnyei, 2020) and language awareness (Svalberg, 2009) researchers have turned their attention the nature of engagement in language learning: learners' "active participation and involvement" (Mercer & Dornyei, 2020, p. 2). The concept of Assessment for Engagement brings these theoretical approaches together in order to (1) build the potential for engagement into assessment tasks and (2) observe and reflect on the nature of learners' engagement, as a critical ingredient of learning, in their doing of assessment tasks. This paper sets out the notion of Assessment for Engagement and examines the potential and actual engagement in two disparate task contexts. The first is individual writers using L2 English in interaction with automated feedback from Grammarly, and the second is L2 Japanese learners carrying out group oral role-play tasks in-person and online. In both contexts, the L2 users' reflections on the nature of their engagement were gathered. These experiences show that levels of engagement are inherent in language learning activities and that engagement plays out in relatively predictable and unpredictable ways across its social, cognitive, behavioural and emotional dimensions (see Hiver et al., 2021; Svalberg, 2009). It is argued that by observing, analysing and reflecting on the nature of learners' engagement, both teachers and learners can facilitate self-regulated learning through assessment.

Blooket used for formatively assessing students' understanding of the mark scheme for IGCSE 0500 English Descriptive Writing Paper 2

Author(s): Louise Finucane

Key words: Technological innovation, Formative Assessment, Learning and Assessment

Abstract: Blooket will be used in the demonstration to show how to formatively assess students' understanding of the mark scheme for the International General Certificate of Secondary Education (IGCSE) 0500 English Descriptive Writing Examination Paper 2 Directed Writing and Composition. Year 11 students within New Zealand spend their time in class and at home preparing for this external examination paper. Blooket is an low-stakes interactive online learning tool that is used within the teaching and learning cycle. It offers students immediate feedback for each question answered. This Blooket activity follows on from a previous lesson where the set homework is given in a flipped classroom approach. The homework assigned is for the Year 11 students who were required to listen and read through the recorded PowerPoint slides of the set homework. They needed to have annotated the mark scheme and explained each area of the mark scheme in their own words. As a follow-up to this homework, Blooket is used as a formative assessment tool to measure the students' true understanding of the mark scheme, and homework. The data is then openly discussed and displayed with the whole class to address any specific areas of challenges and strengths. This in turn helps to foster the students' own awareness of components for descriptive writing and improve learner autonomy. Following the demonstration, time will be given to discuss the disadvantages and benefits of using Blooket, as well as ways this can be adapted into formatively assessing the students' understanding of other language assessments using the mark schemes.

ChatGPT (3.5-turbo) and Writing papers of Cambridge English examinations

Author(s): Daniil M. Ozernyi

Key words: ChatGPT, Cambridge English exams, Asessing Writing

Abstract: ChatGPT (OpenAI) has been noted as a powerful tool for cheating on standardized exams. There have been various investigations of ChatGPT's ability to pass AP tests in the US as well as other standardized tests for native speakers. However, ChatGPT's ability to succeed at well-established and validated ESOL tests like Cambridge line (B1 Preliminary, B2 First, C1 Advanced, C2 Proficiency) remains relatively uninvestigated. Our purpose here is to scrutinize ChatGPT's ability to create written responses for the Cambridge line. We inputted the prompts as they appeared in the openly available exam papers to ChatGPT and collected the responses. The responses were then given to four raters who scores them using the Cambridge scale. We report on these data. In terms of inter-rater reliability, the obtained scores were reliable: the average intraclass correlation coefficient was ~0.7 for all exams, and p = &lt;.000. The statistics were run with irr R package with model "oneway" and type "agreement". In terms of quantitative data, ChatGPT was able to obtain a passing grade for all exams. This carries significance for policy and safety decision-making, given several recent high-profile cheating cases. In terms of the qualitative data, the only pattern that emerged across exams (meaning that we saw them both in C2 and B1 exams) is that ChatGPT is frugal with respect to cohesive devices but good at organizing texts into paragraphs. Our findings revealed that sometimes ChatGPT flounders the lexical part, repeating the vocabulary from the prompt and otherwise. This only emerged when we asked for (relatively) large pieces of writing – about 280 (C2 Proficiency). The ability to effectively paraphrase using a variety of grammatical forms and synonymy seems to be a weak spot of the AI-generated text as well. However, there are indications that even this is likely to improve which we detail.

Collocation use in a Japanese university-wide test of spoken English: Differences across proficiency levels

Author(s): Dr Ivy Chen, Dr Katsunori Kanzawa

Key words: collocations, corpus linguistics, validation

Abstract: This paper is the first of a series of projects aimed at evaluating and improving the KIT Speaking Test, a localised computer-based semi-direct test of English spoken language proficiency developed at a Japanese university; it focuses on test-taker collocation use across proficiency levels. Collocation use is especially pertinent to this test, where English is conceptualized as a lingua franca and performances are scored on task achievement (80%) and task delivery (fluency, 20%) rather than on grammatical accuracy, because collocations are somewhat predictable, so their use reduces cognitive load and is thus correlated with faster processing speed and speaking fluency. The test consists of nine tasks (four integrated listening and speaking). The data comes from the KIT Speaking Test Corpus (KISTEC), which contains 574 examinees' responses, with the low proficiency group consisting of around 12% of the 222,777-word corpus, the intermediate group 37%, the high-intermediate group 33%, and the advanced group 18%. Chi-square tests show that collocation use significantly differs across proficiency levels overall, with the low proficiency group using significantly fewer collocations than the other groups (apart from the advanced group). Individual collocation types show differing usage trends across proficiency levels: use of grammatical collocations (e.g., phrasal verbs) tended to stay relatively constant or increase across proficiency levels, while use of lexical collocations (e.g., adjective-noun collocations) and compound words tended to decrease or decrease after increasing across proficiency levels. This paper presents much-needed findings on collocation use in speaking tasks, which (a) deepens our knowledge of how vocabulary diversity differs across proficiency levels and (b) will be used in the development of an updated rating scale for the KIT and for rater training. Further data analysis will explore the effect of other factors such as predictability of collocations, "correctness" of collocations, and frequency of collocations in native speaker corpora.

Demonstrating the utility of the Japanese-Vocabulary Levels Test to support course-integrated extensive reading in Japanese as a foreign language

Author(s): Kimberley Rothville Michiyo Mori

Key words: Japanese, vocabulary assessment, extensive reading

Abstract: Extensive reading (ER) in Japanese is still a developing field. A small but growing body of research has indicated there are benefits to reading speed, reading comprehension, reading strategy use, learner motivation and attitudes, and vocabulary development, among others. However, without better understanding of learners' vocabulary knowledge before they begin to read, the benefits of ER may be truncated. The identification of this significant barrier to learner development of independent reading skills in Japanese has led to the development of the first Japanese Vocabulary Levels Test (J-VLT) to assess learner vocabulary knowledge to support the integration of ER activities within a tertiary-level Japanese as a foreign language course. The J-VLT assesses learner vocabulary knowledge of the 5000 most frequent Japanese words. It currently exists as a multiple-choice test, with work underway to develop a translation version. In this presentation, we demonstrate how the J-VLT is used in the upper-level language courses at our university to guide students to an appropriate starting level as they begin to read independently for the first time in course-integrated extensive reading activities. The results of the real-world test administrations demonstrate that even if a particular learners' vocabulary size overall is large, they may not know enough words in each particular word band to read extensively at that level. We show how the test's design makes it easy for learners, teachers, and researchers to understand a learners' vocabulary knowledge not just at the 1000-word band level, but also in smaller bands of 500 or 250 words. This supports learners to match themselves with appropriate level graded reading materials. Finally, we show how the use of the test can be combined with the J-LEX lexical profiling tool to assist learners as they move independently beyond the level of graded readers.

Developing a screening test of transcription ability to support the creation of reliable transcripts of indistinct forensic audio

Author(s): Helen Fraser, Debbie Loakes, Jason Fan, Ivy Chen

Key words: transcription test, forensic audio, screening test

Abstract: Covert recordings provide powerful evidence in criminal trials. The problem is that they are often of extremely poor quality, to the extent that they cannot be understood without the assistance of a transcript (Fraser and Loakes 2020). Current Australian law allows transcripts to be provided by police investigating the case (French and Fraser 2018). This has been shown to be unacceptable, even when safeguards embodied in legal procedure are followed properly (Fraser 2018). This paper reports on one aspect of a wider project aiming to create an evidence-based process for providing demonstrably reliable transcripts. The aim of this study was to create a screening test of transcription ability designed to identify transcribers with high levels of transcription ability. In particular, the aim was to explore the best scoring mechanism for the screening test, and procedures for efficient scoring. Forty participants with high levels of experience in transcribing audio, transcribed a 2-minute section of an indistinct forensic-like audio recording, with no prior knowledge of the content or context. Each participant's transcript was divided into intonation phrases (IPs) and each IP was scored against a master transcript which was available for the recording. The results were analysed using Rasch measurement, as well ANOVA and path analysis. Statistical results were further followed up to identify particular trends in the transcription patterns. The analysis revealed that certain IPs were more productive to measurement, and that many IPs could be excluded for more efficient scoring. The paper concludes with implications for the development of transcription tests and next steps for developing the evidence-based process.

Developing the Little Kids Word List app – a MacArthur Bates Communicative Development Inventory (CDI) for Aboriginal languages in central Australia

Author(s): Carmel O'Shannessy, Vanessa Davis, Jessie Bartlett, Alice Nelson

Key words: First Nations languages, language acquisition, Communicative Development Inventory

Abstract: Assessment of the language development of First Nations children in central Australia is a major challenge, because little is known about what the children are learning. We need to know the vocabulary spoken to children, because that is what they learn. There are guides for health professionals in the form of Communicative Development Inventories (CDI) for about 90 languages world-wide, but until now only one for an Australian Indigenous language, Kriol, spoken in northern Australia (Jones et al 2020). A CDI is a list of the most common words that young children up to age 3 years are likely to know and say. In Central Australia many Indigenous children grow up hearing and learning more than one language, and, speakers often do not have opportunities to engage with reading and writing in the languages they speak. This context presents a specific challenge for describing what children are learning and developing a vocabulary assessment tool. A multilingual CDI for four of the languages spoken by young children in Central Australia, the Little Kids' Word List app, has been developed. It includes 4 languages to allow for a child's multilingual vocabulary development. In this paper we outline the processes of development of the Little Kids Word List. The Little Kids' Word List app has the potential to make language assessments of a diverse cohort of young Indigenous children in Central Australia appropriate to the children's developing knowledge of the languages they speak, by making visible the children's languages knowledge base. It will help practitioners to identify potential developmental needs, if present. Assessments that are not based on empirical documentation of the children's and families' everyday languages and multilingualism are likely to give misleading results, either under-reporting knowledge that is present, or under-reporting difficulties children' may have.

Diagnostic writing assessment: Investigating the effects of time limits and understanding students' academic writing self-efficacy beliefs

Author(s): Viola Wei

Key words: PELA, diagnostic assessment, self-efficacy

Abstract: The diagnostic writing assessment, as part of DELNA, a post-entry language assessment, needs to be a fair assessment of students' abilities so that appropriate language learning advice can be provided. The writing tasks are currently administered with an imposed time limit. However, time limits are less likely to be imposed when students complete their written coursework. Often more time is given, during which students can access external resources and also revise, rewrite or edit their draft essays. Therefore, investigating the extent to which the imposed time limit might influence students' performance contributes to the understanding of the validity of DELNA writing assessment. Additionally, the study aims to find out whether students' beliefs align with their writing abilities, as indicated by the DELNA writing assessment. Self-evaluation can help students to reflect on their past learning and identify areas that need further improvement. This allows the language advisers to gain a more thorough understanding of students' needs and address these needs in the advisory sessions, improving the experience for students. In this study, student participants were asked to complete two writing tasks with and without time pressure and two surveys asking about their self-efficacy beliefs and their experience in taking the two tasks. Interview data was also collected from a subset of the participants. Results show that some students did perform better when the time limit was removed, but overall, the difference was small. This indicates that even with the imposed time limit, the writing assessment is likely to provide a fair assessment at a coarse level of students' writing abilities. However, the interview data shows that students generally have a preference for an assessment without the time limit as they believe this resembles real-world situations more and would be a better representation of their abilities.

Dynamic Assessment of Research Writing of Adult ESL Learners: Designing Rubrics for Content Analysis

Author(s): Satuluri Sahana, Lina Mukhopadhyay

Key words: content-based rubrics, dynamic assessment, ESL research writing

Abstract: Dynamic assessment (DA) allows the integration of instruction and assessment through active engagement to understand learners' language abilities and facilitates their linguistic growth (Lantolf & Poehner, 2010). This study aims to investigate the efficacy of DA and feedback for evaluating and scaffolding research writing skills of adult ESL learners. In research writing, given the writers' proficiency in grammar and organizational skills, content takes paramount importance as it constitutes critical components called idea units (Sawaki, 2020), hinting at knowledge depth and analysis to convey complex argumentative content. However, conventional assessment criteria with triadic components — content, language and organization - often overlook higher-order synthesis skills essential for generating coherent, complex and integrated texts (Zhao, 2022). We propose the design of a rubric for content analysis on appraising idea units. So, in this study, abstracts written by thirteen adult Indian ESL learners in the final semester of their master's programme were assessed for content at (i) macro-level — conveying information through idea units in three parts - introduction, body and conclusion - using appropriate text mediation strategies (CEFR, 2020) and (ii) micro-level — lexical features such as breadth and depth to evaluate the thematic development and quality of writing, using lexical tutor. The potential of DA would be harnessed by designing the content rubric through the identification of common idea units and the lexical range of the thirteen participants. The rubric would offer a comprehensive framework for evaluating the macro and micro dimensions of content and providing constructive feedback. The presentation will demonstrate how such an approach can effectively preserve the authenticity and value of writing tasks, like abstracts, allowing assessors to retain the integrity of learners' research while providing meaningful insights for their language development by scaffolding research writing ability through classroom-based DA.

Effect of Peer Feedback on the Accuracy of Peer Assessment of ESL Argumentative Writing

Author(s): Xiao Xie

Key words: evaluative judgement, peer feedback, peer assessment

Abstract: Evaluative judgement is the ability to make decisions about the quality of one's own work and that of others, is necessary not only in the student's current course but also for learning throughout life. Despite this, current feedback and assessment practices have been criticized for being unidirectional and overly focused on content and tasks, as well as positioning students as passive recipients. In order to change this, some assessment-related activities, including peer feedback and peer assessment, must be revisited and redesigned. In the past three decades, a number of empirical studies have examined how giving and receiving feedback affects students' writing performance and their attitudes towards engaging in peer feedback. However, different roles in peer feedback have been understudied in terms of their impact on ESL learners' evaluative judgements, especially regarding peer assessment accuracy of argumentative writing in English. This mixed-methods experimental study intends to analyse the accuracy of peer assessment of 24 Malaysian undergraduate students enrolled in the expository writing course. In the five weeks of peer feedback and peer assessment training, researchers aim to determine the impact of different feedback roles (feedback providers, feedback receivers, and feedback outsiders) on the accuracy of their peer assessment of multiple writing tasks based in terms of five analytical rating criteria (relevance and adequacy of content, compositional organization, cohesion, vocabulary and grammar). Using the Rasch Partial Credit Model (PCM) and thematic analysis, this study will have significant contributions to theory and practice in the area of learning-teaching-assessing of ESL argumentative writing and produce a module using the procedures and materials. The quantitative findings will help the researchers test the related theories in the new peer feedback and peer assessment hybrid settings while the qualitative data may open avenues for further research and new findings especially those related to evaluative judgement.

EFL classes as thinking labs. An approach toward comprehensive assessment

Author(s): Yomaira Angélica Herreño-Contreras

Key words: self-assessment, higher-thinking, EFL

Abstract: This presentation displays students' insights into three areas: the development of higher-order thinking skills (HOTS), learning English as a Foreign Language (EFL), and self-assessment. The study was implemented with sixth-semester Law students at a private Colombian university as part of research aimed at the development of HOTS along with communicative skills in EFL. The project was carried out using a qualitative research approach and followed an action-research cycle (Burns, 2010). This presentation shows the criteria used to assess students' performances and aspects they considered relevant to assess their EFL learning and HOTS development. The data collection instruments were a survey, a speaking rubric, and a semi-structured interview. In regard to HOTS, the results suggest that the students strengthened their analysis, evaluation, and creation skills, which means that they could trace connections between elements, justify and defend their position, produce original work, propose alternatives to solve legal cases, and discuss current issues such as right to health, or death penalty, among others. However, they did not fully correlate daily thinking actions (e.g., planning, collaborating or classifying) with HOTS. With respect to communicative competence, students were able to convey ideas related to their specific field of study (Law). In relation to self-assessment, students' engagement and participation in EFL classes based on HOTS helped them to have a critical perspective on their own learning, being capable of detecting their weaknesses and strengths. In this sense, EFL classes may be considered thinking labs where learners are exposed to activities that usually serve a twofold purpose – to foster communication in English and to develop HOTS – that ultimately results in a more comprehensive approach toward assessment in EFL.

Examining lexical sophistication, diversity features and business vocabulary usage in business English learners' writing performance

Author(s): Yuhu Zou

Key words: lexical proficiency, lexical features, Business English Writing

Abstract: The relationship between lexical features and the quality of L2 writing has received significant attention. The adept utilization of diverse and sophisticated vocabulary is commonly regarded as an indicator of exceptional writing quality. However, for non-native English learners, acquiring a wide array of vocabulary and employing them effectively within specific writing genres, such as academic and business writing, can pose a formidable challenge. This study focuses on examining lexical proficiency and investigates the lexical features and their predication ability of human judgement of lexical proficiency in Business English letters written by Chinese EFL learners. A total of 260 Business English letters written by students (N = 65, Major in Management Science and Economics) will be analyzed and rated by 3 raters for their lexical proficiency. The body part of those letters will be preserved and spelling errors will be corrected before converting them to plain text format using Python. Then, all files will be renamed in the same pattern. Subsequently, TAALES will be employed to capture lexical sophistication features, TAALED to measure lexical diversity features and Range to gauge the usage of business vocabulary. Following a pilot rating, three raters will evaluate the texts using a self-developed rubric. The data will be imported to STATA for descriptive, correlation and stepwise regression analysis. We find that several features are strong predictors, such as age of acquisition (AoA), average number of senses for nouns. For AoA, we assume that those students are novice learners of English, and their usage of words is still in the early stages of development or not yet fully proficient, which makes age of acquisition is a strong predictor. For average number of senses for nouns, since the number of senses a noun has is determined by its semantic range and the context in which it is used, we think, with a higher lexical proficiency, one might have a deeper understanding of the word and be able to use a variety of senses for it.

Examining the relationship between language assessment and institutional policy: The case of high-stakes examinations in Singapore

Author(s): Azrifah Zakaria, Vahid Aryadoust

Key words: language assessment policy, test administration, longitudinal

Abstract: This study examines how public policy and administration shaped language testing through a historical perspective. While previous studies have centred on the United States and the United Kingdom, this study proffers insight into how testing has been administered in a southeast Asian country, Singapore. Through a study of Singapore’s Ministry of Education’s archived documents, we examined how different factors affected the way examinations were planned and administered. We demonstrate how language testing as a sociocultural practice is enacted in a context of planned governance and discuss the implications for construct definitions and validity. The data were analysed using a method called historical narrative inquiry, focusing on three research questions: 1) How have the institutions acted to form policies pertaining to language testing in Singapore? 2) How have the institutions responded to the individual actions? 3) How have these institutional actions affected definition of the construct of language assessments? Two themes emerged from our analysis. Starting with Singapore’s colonial era, the first theme From Free Enterprise to Governance describes how multiple assessments co-existed, in absence of an established system, before the gradual move towards centralized planning. The second theme, Individual Agency and Institutional Change, occurred during Singapore’s transition from colony to independent state. The late colonial era is particularly illuminative in demonstrating how the needs of the individual can both dovetail and conflict with the institution. Responses by the ministry and examination board provide important insights to the processes of test development and how the different stakeholders affect language assessment. A gradual shift in the construct of the assessments - from functional literacy to communicative competence – is also observed and situated within these historical developments. We discuss the implications of this intertwining of public policy and test administration for the language assessment community, and other contexts beyond Singapore.

Exploration of assessment practices in community languages schools

Author(s): Anna Mikhaylova

Key words: community schools, approaches to language assessment, teacher training

Abstract: Community languages (CL) schools remain largely invisible to the larger English-speaking community and peripheral to the state education system (Cruickshank, Jung & Li, 2020; Nordstrom, 2020), in spite of being a major source of support for ethnolinguistic vitality and community building for many minority language speakers in Australia. Community language teacher training, developing evidence-based assessment practices and increasing research-practitioner collaboration have all been identified by Carreira & Kagan (2018) as priorities for the next 50 years of heritage language studies. The aim of this project is to make practical steps for reaching the challenging synergy between research and practice in language teaching (Sato & Leowen, 2018) in the context of Russian community schools in Queensland, Australia. Russian is not taught in either public or private primary or secondary schools in Queensland, and the existing Russian community/ethnic schools are not supported in any way through the formal education system, which presents a number of challenges to the teachers but also affords a lot of agency in how they approach language instruction in general and assessment practices in particular. These schools place children into classes based on biological age rather than language proficiency and are not required to conduct any formative or summative assessments. The data for this project were collected within a larger research partnership project with three Russian community schools in Queensland and include qualitative data from observations and professional development workshops held by the researchers at the school, semi-structured interviews with teachers, as well as an online survey of teachers and parents. In our presentation we discuss the approach to assessment practices adopted in these community schools. We also explore how teachers approach the highly varied language proficiency of children, which was identified as the key challenge reported by the teachers and in past research on heritage language instruction.

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

Author(s): Xiaoxiao Kong

Key words: language assessment for professional purposes, English-as-a-second-language speakers, domain definition

Abstract: Language assessments are increasingly used for professional registration purposes (Knoch & Macqueen, 2020), and this is no exception for teacher registration in Australia. From 2011, both overseas teachers and international graduates of teaching degrees are required to achieve set scores in one of the approved English proficiency tests in order to register and work as a teacher within the Australian early childhood and school contexts (Australian Institute for Teaching and School Leadership, 2011). The current study investigates the linguistic and communicative demands of early childhood and secondary school teachers in Australia, as well as the appropriateness and adequacy of the IELTS Academic, the only English language proficiency test approved by all Australian jurisdictions for teacher registration, as a benchmark measure for assessing English language proficiency for teacher registration. Specifically, multiple sources of data will be collected to inform language demands of early childhood and secondary teaching, including a review of existing literature and position descriptions for early childhood and secondary teachers in a range of contexts in Australia, focus group interviews with teachers, and an analysis of teacher's diaries on communicative tasks undertaken during typical workdays. In addition, individual interviews will be conducted with non-native English-speaking background teachers and their mentors regarding the linguistic and communicative challenges in teaching. Teachers interviewed will also be invited to review the IELTS Academic test tasks and comment on their relevance to the language tasks required in the workplace. Findings from the needs analysis will serve as sources for backing for the domain definition inference in validating the IELTS Academic for teacher registration purposes in the Australian context. Such investigations as proposed in this study provide implications for policy formulation as well as the design and implementation of language assessment for teacher education and employment purposes, which could in turn contribute to student outcomes.

Exploring the relationship between spoken lexical diversity scores and human rater vocabulary scores

Author(s): Thwin Myint Maw

Key words: Lexical diversity, Vocabulary assessment, Objective and subjective assessment

Abstract: There are many different measures of lexical diversity (LD), with different effects from sample length (Treffers-Daller, 2013). While choosing a standard sample length is one way to minimize the differences resulting from different text lengths, when human raters listen to speech samples in a test setting this is not possible. In addition, while previous studies have found connections between sample lexis and human ratings of comprehensibility (Saito et al, 2013; Trofimovich & Isaacs, 2017), there has been little research looking at ratings of LD directly. This experiment seeks to answer which measures of lexical diversity (LD) are most closely correlated with human ratings of lexical resources. This study examines the LD scores of 39 L2 English speech samples (1-2 minutes in length) based on the IELTS speaking test part 2. These samples range in length from 62 to 376 tokens. The lexical resource ratings were done on a 9-point scale based on IELTS. The raters consisted of 10 English L1, and 12 Japanese L1 English teachers living in Japan, as well as 10 Japanese L1 students in Japan. The samples were also self-rated by the people whose speech was recorded. The results show that the lexical resource self-ratings were significantly lower (p&lt;.002) than, and not correlated with, the ratings by others. However, self-ratings were significantly correlated with the following LD measures: MATTR and MTLD (p&lt;.05), D (p&lt;.01), and Guilard's Index, TTR, Types, and Tokens (p&lt;.001). Self-ratings were not significantly correlated with Maas and HD-D. Interestingly, there were no significant correlations between the outside listeners and any of the LD measures except for Maas (p&lt;.05). Future studies will look at larger sample sizes of raters and investigate what rater background factors (teaching experience and L1) may influence correlations between LD measures and human lexical ratings.

Exploring the Sensitivity of the Multifaceted Receptive Vocabulary Assessment Test (MRVAT) in Detecting Changes among Lower-Level English Proficiency Learners

Author(s): Hosam Elmetaher

Key words: Multifaceted Vocabulary Assessment, Receptive Vocabulary Knowledge, Short-term longitudinal study

Abstract: This longitudinal research builds on a pilot study of Author (2022). Author created and validated a new Multifaceted Receptive Vocabulary Assessment Test (MRVAT). The MRVAT contains 60 multiple-choice integrated-skills (listening and reading) questions. The MRVAT questions and responses primarily use the first 5,000 most frequent words in English according to JACET 8000 (Ishikawa et al., 2003) and are divided into three parts (A, B, and C) that cover the ability to recognize three-word aspects (use, form, and usage). A full MRVAT requires 30 minutes to complete and can be processed by the class teacher. The current study aims to explore the MRVAT sensitivity in detecting changes for lower-level English language proficiency learners. The study employs 90 undergraduate, beginner-to-intermediate, English Proficiency level Japanese learners. Participants were 18 years old and in their first year at university. Participants were placed in different compulsory skills-based academic English classes, which met twice a week for 100 minutes per session. Participants took three different equivalent versions of the MRVAT over an academic quarter at the following three test points: 0, 1, and 2 months. The participants' mean percentage scores for the three MRVATs have been statistically processed using T-Test. In general terms, the MRVAT detects significant positive changes in participants' receptive vocabulary knowledge over the collected data points: Time 2-Time 1 (11.13), (t (89) = 8.523, p &lt; .001); Time 3-Time 2 (2.055), (t (89) = 1.921, p &lt; .05); and Time 3-Time 1 (13.185), (t (89) = 12.014, p &lt; .001). The MRVAT might provide a novel, multifaceted, and sensitive receptive vocabulary knowledge testing tool for both language learners and teachers. Future studies could replicate the current study with different L1 populations in order to represent a broader picture of the MRVAT's sensitivity for detecting changes.

Indigenous criteria and differential rater behaviour: On the challenges of assessing the complex LSP construct of Teacher Language Competence

Author(s): Olivia Rütti-Joy

Key words: Teacher language competence, rater variability, LSP

Abstract: Teacher Language Competence (TLC) describes the communicative skills that L2 teachers require to teach successfully. While several researchers have attempted to conceptualise TLC and its partial factors, the construct remains highly complex and vague. The profession-related language competence profiles (PRLCP) and their corresponding analytic assessment rubric (PRLC-R) constitute recent language for specific purpose (LSP) tools that propose a more precise description and assessment of TLC. Next to classic assessment criteria for oral speech production such as accuracy, pronunciation or vocabulary, the PRLC-R contain the indigenous criterion “addressee-specificity”, which reflects a teacher’s ability to adapt their linguistic expression to the language proficiency of their addressees. This paper reports on a study that investigated 4 evaluators’ behaviour with reference to their severity or leniency across test performances, criteria and tasks when employing the PRLC-R to assess pre-service teachers’ L2 oral TLC. 48 pre-service language teachers’ test responses in form of 415 audio files elicited through a near-authentic, competence-oriented and online-mediated pre-post LSP performance test constituted the research data. After the rater familiarisation and rater training period, 40% of all audio files were rated double-blind. Krippendorff’s α was computed across all rating criteria for rater pairs and between all raters per criterion, and bias and interaction analyses were conducted by means of a multi-faceted Rasch Analysis and the implementation of a Partial Credit Model. While the bias and interaction analyses revealed significant differential rater functioning overall, the indigenous criterion proved to be the criterion with the highest rater variability and lowest interrater reliability values. The results provide valuable insights into the functioning of the assessment rubric as well as the criterion “addressee-specificity”, indicating that the criterion may constitute its own independent construct rather than a partial aspect of TLC. Furthermore, they highlight the value and challenges of indigenous criteria for describing TLC.

Interaction matters: Automated assessment of interactive features in paired speaking tasks

Author(s): Rena (Wei) Gao

Key words: Automated asssessment, Natural language processing, Interactional competence

Abstract: Automated scoring has been employed and studied in large-scale language assessments for decades (Johnson & McCaffrey, 2023). Currently, most automated scoring tools are used in tasks such as reading aloud, repeating sentences, and constructed responses to test questions. The automated scoring focus on aspects including pronunciation, vocabulary, and fluency in delivery. However, this mainly indicts the performance in pronunciation or use of vocabulary, which lacks the representation of interactional competence in authentic scenarios (Dai, 2023). Thus, this study develops an evaluation model based on key aspects of interactional competence, including topic development, active listener response, and discourse management. These aspects are defined and trained in automated scoring models with downstream fine-tuning and supervised learning (Amorim et al., 2019). This study collected a 60-hour dataset of paired speaking tasks from 120 L2 English speakers whose L1 is Chinese, including two types of tasks: problem-solving tasks and topic discussion. Three recruited raters set the gold standard rating scores after the pilot test of the rating framework. The results indicate that the automated scoring model: (1). can predict and assess the ability of the active listening aspect in the dataset; (2). in assessing the discussion based on a certain topic task, the performance of the automated scoring model is better than the performance in problem-solving tasks in certain contexts; (3). indicates a significant correlation between aspects of interactional competence and L2 proficiency, and further proves the necessity of conducting automated scoring for interactive features in large-scale language assessments.

Investigating test constructs for assessing EMI-readiness of content lecturers in Thai international medical programs

Author(s): Teaka Sowaprux, Jirada Wudthayagorn, Thanakorn Jirasevijinda

Key words: English-Medium Instruction, medical content lecturers, test constructs

Abstract: Although researchers suggest that a minimum of a C1 level on the CEFR would be most suitable for EMI lecturer certification (Dimova, 2021), it remains unclear which language constructs are most important for assessing EMI language proficiency for content instructors, particularly as it relates to medical lecturers in differing local contexts. To investigate this gap, this qualitative study conducts ethnographic observation of 14 pre-clinical lectures on foundational medical science at one academic medical school in Bangkok, Thailand. Employing Atlas.ti, a team of three coders (two university professors and one graduate student) use two-step process coding techniques to encode lecturer transcript data. Unlike previous studies that have involved students with mixed English proficiency, the medical program under examination consists of native bilingual Thai/English college students who do not encounter difficulties in understanding English. The study reveals specific KSAs (knowledge, skills, and abilities) used by medical content lecturers related to ideational organization, speech acts, and classroom management language that are imperative for strategic communication and content comprehension. Based on the findings, this study proposes criteria that could be utilized for EMI teacher certification in Thai international university medical programs.

Language assessment literacy: Teachers' attitudes in ensuring fair classroom assessment practices

Author(s): Karim Rezagah

Key words: Assessment fairness, assessment literacy, Classroom assessment

Abstract: In 2010, Iran's educational system decided to leave teaching English through traditional teaching methods (Grammar-translation method) behind and move toward communicative approaches (Foroozandeh & Forouzani, 2015). In fact, the educational system authorities decided to prioritize communication proficiency in teaching English in schools. According to the new shift in the educational system, the assessment was also supposed to become harmonious with the redirection in teaching by creating a balance between assessment for and of learning (Alavi-moghaddam, et al., 2018). In Iran, as an exam-oriented context in which the assessments play a considerable role in the students' achievements or failures, conducting fair assessments is one of the fundamental expectations that could be anticipated from the teachers specifically (as the protagonists of classroom assessment), and the educational system generally (as a stakeholder). One of the main challenges in Iran is the teachers' specific needs in developing their assessment literacy (Rezagah, 2022) to make their classroom assessment as fair as possible. To make classroom assessment practices of Iranian teachers, sustainable and of enhanced quality, adequate assessment literacy for the involved stakeholders, including but not restricted to teachers, is required. In order to survey the teachers' attitudes towards the characteristics of fair assessment, 15 English language head teachers (English language teachers who are responsible for moderating the group of teachers' activities) were interviewed. The semi-structured interviews were independently coded by the researcher and an expert in assessment, and the content was analyzed. Eight themes appeared as the result of the content analysis, and two of them were related to the purpose of the current presentation, including adjusting testing to teaching and assessment fairness (Rezagah, 2022). This presentation will focus on discussing these themes in more detail and concludes with implications for research on teachers' assessment literacy development.

Language proficiency and wages in Korea

Author(s): Junghyun Baik

Key words: Korean proficiency, language policy, language and economics

Abstract: In Korea, disputes over language policies between globalists supporting the strengthening of English education and nationalists working to solidify Korean's position as a native language have intensified; however, discussions have been limited to ideological debates with the lack of empirical studies examining the actual advantages of each language. Concerning this issue, the present study tries to fill the gap by scrutinizing the language-benefit production mechanism, which can potentially guide language policy in Korea. In developing arguments, this study adopts a research approach from "language and economics," an interdisciplinary area bridging sociolinguistics and labour economics. Relying on a sequential mixed-methods approach, this study firstly estimates quantitative language benefits using four national panels - Korean Labour & Income Panel Study (KLIPS), Korean Education & Employment Panel (KEEP), Graduate Occupational Mobility Survey (GOMS), Youth Panel (YP) – that provide information on participants' linguistic, financial, and employment backgrounds. The quantitative phase has applied a series of Mincer-type earnings functions extended to include participants' language proficiency proxied by self-reporting grades in secondary school, and objective test scores, such as standardized English tests and college entrance exam results. Two-stage least squares estimations using instrumental variables, such as interest in language subjects and overseas language training experience, have also been conducted to show the best possible estimates by mitigating statistical biases. The following qualitative thematic analysis uses semi-structured interviews with ten human resource managers, expanding pre-conducting wage analyses to the language-benefit mechanism exploration in the Korean industry. The theoretical roots of both steps come from the "Language augmented theory of production" (Grin et al., 2009) and "Critical realism" (Bhaskar, 1975), which rationalize mixed-methods and interdisciplinary approaches bridging linguistics and economics. Expected results might have policy implications in the multilingual context where demands for additional language skills are widespread and language-ideological debates continue to resonate.

Leveraging Generative AI for Enhanced Content Development in Language Testing: Implications for Item Writer Recruitment, Training, and Engagement

Author(s): Jennifer J. Flasko, Michelle Y. Chen

Key words: Content development, Generative AI, Item writer training

Abstract: This study investigates the integration of generative AI tools into content development for language testing. It explores the impacts of AI-assisted tools on traditional item writer recruitment and training and the willingness of item writers to adopt these tools. This study contributes to advancing content development practices by addressing the challenges of integrating AI tools into item writer training and content development. More specifically, the purposes of this study are threefold: (1) to explore the interest and willingness of item writers to adopt new AI tools into their work; (2) to identify additional recruitment and training considerations for item writers to use generative AI tools efficiently; and (3) to propose strategies and interventions to improve item writers' engagement and success in adopting new AI technologies. To achieve these research objectives, a mixed-methods approach is employed. The research begins with a review of the literature on traditional item writer training, AI-assisted tools in content development, and the factors influencing user adoption of new technology. Surveys and interviews are conducted to gather data on item writers' perceptions, training needs, and willingness to adopt generative AI tools. Additionally, potential updates to the current recruitment and training program for a large-scale language assessment will be proposed to incorporate generative AI tools. In the future, the effectiveness of these updates could be further evaluated. This research is expected to offer insight into the impact of generative AI tools on traditional item writing and content development. Strategies and interventions to improve item writers' engagement and success in adopting AI technologies will be proposed to promote the smooth integration of AI-assisted tools into content development. Overall, this study contributes to the advancement of content development practices and the responsible implementation of AI technologies into language testing.

Linking translation tasks to the Common European Framework of Reference (CEFR): The case of the General English Proficiency Test (GEPT)

Author(s): Jason Fan, Ute Knoch, Ivy Chen, Jessica Wu

Key words: linking, CEFR, translation test

Abstract: Despite the numerous studies that link language tests to the Common European Framework of Reference (CEFR), little research has focused on translation tests or tasks. Spurred on by the recent publication of the illustrative descriptors of mediation in the companion volume to the CEFR (Council of Europe, 2018), this study aimed to link a Chinese-English (C-E) translation task (i.e., Part 1 of the General English Proficiency Test writing test) to the CEFR. A 'twin-panel' approach was adopted to compare the judgements of those who were familiar with the GEPT but with little or no experience in C-E translation assessment and practice (Group A, n = 8) with the judgements of those with substantial experience in translation assessment and practice (Group B, n = 4). In addition, this study explored the mental processes of the two groups of panellists (n = 4, with 2 from each group) in linking GEPT translation samples to the CEFR levels, through a verbal protocol think-aloud study. Two examinee-centred standard-setting methods (i.e., the 'Contrasting Groups' and 'Borderline Group' methods) were used in combination in the alignment process. Many-facets Rasch analysis was implemented to explore whether significant differences existed in the severity of the judgements made by the two groups of panellists. The qualitative data was thematically coded and analysed. Findings indicate that (a) the panellists' judgements were consistent, but their severity levels differed, and (b) the two groups of panellists engaged in different mental processes and prioritised different criteria in the alignment process. While providing the GEPT developer, users and relevant policy makers with credible linking results, the findings also shed light on several crucial issues in linking a translation test or task to the CEFR, such as the selection and use of scales and descriptors and the recruitment and training of panellists.

Measuring L2 reading assessment processes via user experience

Author(s): Sarah Goodwin

Key words: reading processes, user experience, L2 reading assessment

Abstract: To accurately, consistently, and fairly evaluate reading skills, reading proficiency tests must demonstrate construct and content validity, as well as cognitive validity (Weir, 2005). Assessing advanced reading abilities involves examining higher-order thinking skills, global and local text analysis, and understanding aspects beyond the text's surface level (Alderson, 2000; Grabe & Stoller, 2013; Koda, 1988; 1990). Apart from comprehension questions to determine if test takers understood the material, alternative research methods like stimulated recall (SR) interviews can reveal cognitive processes involved in reading. SRs involve using a stimulus to help individuals recall or reflect on something they read, heard, or experienced (Ericsson & Simon, 1980; 1993; 1996; Gass & Mackey, 2000; 2016). This demonstration shows how assessment researchers can collect cognitive validation evidence using web-based user experience research software, which captures computer screen and webcam recordings of participants as they complete internet-based test tasks. This can reveal whether test takers employ the same cognitive processes during reading for a simulated high-stakes assessment context as the processes they would activate while reading in non-testing situations. The demo will present case studies of examinees that summarize their advanced reading processes. Data were collected from 10 adults who took reading assessments, including a highlighting-text task assessing understanding of main ideas and important details, and comprehension questions requiring global and local reading skills. Candidates click-and-drag their cursor over text on screen within the passage to answer comprehension questions. Asynchronous think-aloud stimulated recalls took place right after the reading tests, with no researcher present. This methodological decision was made intentionally to reduce the impact of the observer's paradox. The demo will reveal how examinees skimmed-and-scanned to locate specific information, and how they tackled comprehension questions. Results include the challenges examinees faced, the strategies they employed, and how the test design effectively assessed their higher-order reading skills.

Nurturing Assessment for Learning through Assessment of teachers: Praxis-pedagogy for teacher/student development

Author(s): Harsha Dulari Wijesekera

Key words: assessments for learning, content teacher professional development, praxis-pedagogy

Abstract: Bilingual education (BE) in Sri Lanka serves two primary purposes: content acquisition and second language (English) acquisition through content. However, using conventional testing methods to fulfill both purposes while adhering to testing principles such as construct validity is questionable. Conventional testing fails to minimize unintentional construct-irrelevant sources, particularly the language in which students are not proficient. Additionally, the lack of knowledge and skills among BE teachers due to the absence of professional development opportunities and an excessively examination-oriented culture in Sri Lanka makes it challenging to change teachers' mindset and encourage innovative approaches to assessment. This paper illustrates how teachers engaged in alternative forms of assessment as part of their postgraduate diploma program. These assessments placed equal emphasis on the process, product, and collaboration, critical thinking aspects of learning. Two of the assessments were final evaluations of two courses, requiring teachers to create e-portfolios in collaboration with their students. The third assessment was a Continuous Assessment, where teachers engaged in individual lesson planning, sharing, reviewing, and eventually writing group lesson plans in preparation for their teaching practicum. Teachers' reflections on these experiences were gathered through written responses and interviews. The analysis of reflections reveals significant changes in teachers' agency, attitudes, knowledge, and skills. Teachers emphasize that without this practical exposure, they would never have learned these innovative assessment methods, which they consider crucial for fostering an assessment culture beneficial not only to their students but also to teacher development programs. They highlight that these alternative assessments served as learning tools for both students and themselves, particularly in enhancing language proficiency and digital skills. They believe that if designed reliably, such alternative methods can be utilized for school-based assessments and could predict students' achievements in externally designed tests. Overall, this study underscores the significance of a praxis-oriented approach to teacher development.

Opening the black box of experience in language assessment literacy (LAL) research: A sociocultural perspective

Author(s): Xuan Minh Ngo

Key words: language assessment literacy, sociocultural theory, perezhivanie

Abstract: There has been a growing number of studies on language assessment literacy (LAL), particularly among teachers. This body of literature indicates that experience, together with context, is a major mediator of teacher LAL development. Nevertheless, few studies have attempted to define what experience constitutes and show how experience interacts with context to mediate teacher LAL development. It is this gap that the current presentation aims to bridge. Specifically, this presentation argues that Vygotsky's perezhivanie is a useful analytical tool to operationalise the construct of experience in LAL research. Since perezhivanie is a unit of analysis where both "personal characteristics and situational characteristics" (Vygotsky, 1994b, p.342) are represented, it can help shed light on how the interactions between context and experience mediate teacher LAL development. Moreover, as a unity of cognition and emotion, perezhivanie offers a powerful theoretical concept to examine not only the intellectual but also affective dimension of teacher LAL development. To illustrate the benefits of using perezhivanie to operationalise the construct of experience, the presentation incorporates some key findings from a recently completed narrative inquiry into teacher LAL development in Vietnam.

Oral Recall as Assessment of Reading Comprehension for Intermediate-level Chinese as Second Language Learners

Author(s): Shuyi Yang

Key words: oral recall, text-based model, situation model

Abstract: Reading in a second language (L2) is a process of constructing text-based model and situation model. The common reading assessment, multiple-choice, focuses on individual textual pieces instead of coherent models of the entire text. Recall after reading, on the other hand, probes both text-based and situation model and provides richer information than a numeric score. Although recall has been used among L2 learners, most studies have focused on written recall and alphabetic language learners. Research on the effectiveness of oral recall as a classroom-based reading assessment for Chinese L2 readers is lacking. It also calls for further investigation into readers' recalls, especially the recall structure, in order to design targeted instruction. This study examined the effectiveness of oral recall to measure reading comprehension among 62 English-speaking, intermediate-level Chinese L2 learners. Participants silently read an instructional-level expository text and orally recalled it. They also completed a reading proficiency test. Recalls were analyzed by the recalled textual propositions, a holistic scale including textual content, relevance, and organization, and Pathfinder Network which converted recalls into networks and compared them to the text and expert recalls. The recalled propositions, the textual content and relevance in holistic scale, and the network similarity to the text measured text-based model, whereas the organization in holistic scale and the network similarity to expert recalls measured situational model. Results showed significant, positive correlations between oral recall performance and proficiency test, indicating its effectiveness in measuring comprehension. Both text-based and situation model effectively predicted comprehension, with the latter emerging as a stronger indicator. Learner recalls approached a hierarchical structure and better recalls interwove textual information with readers' comments. The findings offered support for oral recall as an effective assessment in daily classroom, extended the application of Pathfinder Network to L2 settings, and suggested pedagogical practices of text-structure-analysis and elaborative inference.

Plain or precise: How writers do "readability" in health information for multiple audiences

Author(s): Jeanie Henchman

Key words: Health information, Readability, Writing processes

Abstract: 'Readability' and 'plain language' are widely-used constructs for estimating the extent to which written information will be understood by a range of readers and for guiding the production of public information. Although it is important that authoritative sources of information about health are available and accessible to the literacy range of the general public, readers have diverse needs, including to inform policy decisions and service delivery. Readability is often measured through calculations of word and sentence length, whereas plain language is prescribed by global and federal authorities through style guides and accessibility standards. To what extent do these tools guide writers who are tasked with producing health information that is a suitable basis for major policy decisions? This study investigates how readability and plain language are operationalised by health information writers from the Australian Institute of Health and Welfare. Using data gathered via think aloud writing sessions and semi-structured interviews, a reflexive thematic analysis (Braun & Clarke, 2006) was carried out to develop a model of writers' processes and perceptions of responsible health information writing. Findings about the writers' decision making regarding the linguistic features of their texts as well as their concepts of the texts' communicative purposes provide insight into their processes of constructing health information. Writers managed competing concepts of clarity between traditional views of readability and technical accuracy, demonstrating carefully considered linguistic decisions that directly conflict with guidelines. A distributed authoring process and a sense of responsibility to intended audiences contributed to the writers' content and language decisions. The writers' insights show that current guidelines are insufficient in capturing the complexity of producing effective and useful health information texts for multiple audiences. The findings reveal the complex role of the writer in the governmental and health domains, demonstrating what could be lost in over-emphasising the use of widespread tools for effectively reaching intended audiences.

Policy intentions and realities: Assessment of non-common foreign languages in China's secondary education

Author(s): Chenyang Zhang

Key words: Non-common foreign language assessment, Language policy, Agency

Abstract: Following the promulgation of the Belt & Road (B&R) initiative in 2013, increasing attention has been devoted to supporting non-common foreign language (NCFL, i.e., foreign languages other than English) education in China. Under policy support, senior secondary students can now access NCFL schooling and select a NCFL subject, instead of English, in the National Entrance Examination (a.k.a. Gaokao) in China. Given the limited scholarly attention to this significant policy shift in China, this study aims to explore the intentions of national and local government policies for NCFL education and testing; in addition, it also aims to investigate the agency of local stakeholders by exploring their perceptions and actions, including school administrators and program coordinators when they appropriate the top-down policies and enact school-level policies in China's senior secondary schools, as well as teachers and students when they respond to the school-level NCFL assessment policies and prepare for the NCFL subject in the Gaokao. Data will be gathered in three ways: in-depth interviews with 4 school administrators, 4 program coordinators, 6 teachers, and 20 students in two senior secondary schools in Shanghai, classroom observation, and documents (national policies, curriculum standards, teaching materials, and students' journals) collection. Positioned at the intersection of language testing and language policy, this study will contribute to language testing by shedding light on the significant role of agency that policy actors exercise in test impact and consequences, thus enriching the current understanding and theorisation of the impact of language tests embedded in policy contexts.

Redefining Language Assessment in the wake of AI invasion and Technological Innovations in Language Testing

Author(s): Mohammad Haseen Ahmed

Key words: language testing tools, educational technology, fair language assessment system

Abstract: The field of language assessment has undergone a significant transformation in recent years, primarily driven by the invasion of artificial intelligence (AI) and rapid technological advancements. If we explore the benefits and challenges associated with AI-based language testing, including automated scoring systems, chatbots, and virtual reality simulations, we do come across some of the ethical considerations and potential biases that arise with AI integration in language assessment. Considering these developments, a new framework could be suggested that combines human expertise with AI capabilities to create a comprehensive and fair language assessment system. The proposed framework emphasizes the importance of a balanced approach, leveraging the strengths of AI while maintaining the essential role of human assessment in ensuring validity, reliability, and fairness. This paper serves as a call to action for educators, policymakers, and assessment developers to collaboratively redefine language assessment practices to meet the evolving needs of learners in the 21st century.

Rethinking assessment for inclusivity. Considerations on academic writing in multilingual contexts

Author(s): Ana Maria Benton Z

Key words: academic writing, inclusive assessment, decolonial views

Abstract: Most universities acknowledge nowadays their diverse cultural and multilingual student body, and often proudly aim for 'equity and inclusion'. Critical perspectives of international education highlight some of the challenges international and multilingual students can face (Pennycook, 2021; Canagarajah, 2023). While universities see their international and diverse student body as enriching for all, they sometimes fall short at providing key support to some struggling students. PG and PhD students build and develop their specialised knowledge and are assessed in a significant way through their academic writing. Still, the expectations regarding academic writing are often elusive for them. Furthermore, English academic writing can become a language barrier. I speak here of my experience completing my doctoral studies, as a Mexican PhD student in a New Zealand university, and how my struggle with academic writing allowed me later on to develop an 'Academic Writing Learning Framework' (Benton, 2021) which considers students' sociocultural and linguistic backgrounds and that has been helpful for some students in my role of language learning adviser. This learning framework suggests a way for the scaffolding of students' academic writing and the promotion of their autonomy. Some students' experiences would be shared in here as well as some considerations to expand the short writing programme going forward.

Standardised tests of English proficiency of international students: Evaluating the state of play

Author(s): John Read

Key words: standardised tests, test validity, international students

Abstract: With the return of international students to universities following the upheavals of the Covid-19 pandemic, it is timely to review developments in the design and implementation of the standardised tests that these students are required to take to provide evidence of their proficiency in academic English. For a long period the market for such tests was dominated by TOEFL and IELTS, with the Pearson test as a newer, more fully computerised competitor. In the 2010s the Duolingo English Test emerged as an avowed disruptor of the standardised testing industry, with a low-cost, high-tech model of remote delivery that gave the test a competitive advantage when the pandemic took hold. Now that national borders have reopened, we find a rather complex situation where numerous providers offer an array of tests that vary significantly in their key features. It is important for language testers to keep abreast of these developments to be able where necessary to counter the claims of those who market the tests. This paper will first give an overview of the range of tests currently available. It can be argued that there is a continuum represented at one end by the paper-based IELTS and at the other by the Duolingo test. Other tests and their versions can be ranged along the continuum in terms of criteria such as the nature of the test tasks; the mode of test delivery and administration time; the degree to which scoring is automated; and the extent to which speaking and writing skills are assessed. This will lead to discussion about the nature of the construct and the validity of each test as a measure of academic proficiency. If the tests are seen primarily as screening measures, it highlights the need to complement the test scores with post-admission assessments as appropriate.

Student ethical considerations on the use of language assistance tools for assessed academic writing

Author(s): Elpida Petraki, Averil Grieve, Amir Rouhshad, Alan Bechaz, David Wei Dai

Key words: academic integrity, translation software, English as an Additional Language

Abstract: Lack of clear university policies on the use of Language Assistance Tools (LATs), such as AI-based machine translation, in assessed academic writing and related academic integrity concerns (Alley, 2005) fuel debates about the ethicality of using such technology (e.g. Correa, 2014; Harris, 2010). Although there has been research on the use of machine translation, such as Google Translate, and other LATs like online dictionaries, and word processing software in translation studies and additional language acquisition (e.g. Garcia & Pena, 2011; Jin & Deifell, 2013; Liu et al., 2022) not much is known regarding university students' ethical perceptions of using such tools for assessed academic writing in non-language focused disciplines (e.g. e.g. Ata & Debreli, 2021; Jolley & Maimone, 2015). Acquiring a clear perception of English as an Additional Language (EAL) student views on the use of technology in this context is vital to inform university policies and address academic integrity concerns (Mundt & Groves, 2016). This project examines the ethical positions of 23 EAL nursing and midwifery students on using LATs for assessed academic writing in undergraduate and postgraduate nursing education in a large Australian university. Thematic analysis of student responses collected through semi-structured interviews reveal three main considerations when making ethical decisions concerning the use of LATs for assessed academic writing: 1) ownership of language and ideas; 2) fairness in relation to other students, society and institutions; and 3) personal growth. At the same time, students discuss a variety of arguments, conclusions and approaches to the ethics of using LATs, which go beyond academic integrity. The findings indicate that students' ethical decisions are scalar, strategic and dynamic, rather than dichotomous. The examination of their complex and nuanced ethical decision-making stances can inform policies concerning the use of LATs in assessed academic writing and academic integrity at a tertiary level.

Studies informing test equivalency tables – how well are they serving test users?

Author(s): Ute Knoch, Jason Fan

Key words: concordancing, test equivalence tables, policy

Abstract: Test providers are required to publish test equivalency tables to help test users and other stakeholders to compare and interpret test scores in light of other, more familiar or established tests. To arrive at these tables, testing agencies either conduct concordancing studies and/or link their tests to an established proficiency framework, such as the CEFR. Such studies are not without problems, however, and are rarely scrutinized for quality. The aim of this study was to examine the publically available studies/documentation available for four major international language tests underpinning test equivalency tables. To evaluate the concordancing studies, we followed two steps. First, we compiled a list of best-practice criteria from a review of both the literature (e.g., Pommerich & Dorans, 2004) and various standards documents in the field of educational measurement and language assessment (ILTA Guidelines for Practice, 2020; AERA/APA/NCME Standards, 2014). Next, we carefully read and re-read the publically available studies by the four tests and compared the studies to the criteria created in the first step. The results showed that none of the concordancing studies fulfilled all the criteria we identified, but that some were more robust than others. We also reviewed the CEFR linking stuides in light of the strengths of their 'linking argument'. These studies varied widely in the methodologies employed and the arguments for linkage put forward by the test providers. Based on the findings of this study, we critically evaluate the usefulness of the test equivalence tables put forward by test providers to test users and policy-makers.

Test takers' attitudes towards at-home testing during and post pandemic

Author(s): Jieun Kim

Key words: at-home testing, remote proctoring, test takers' perspectives

Abstract: The recent pandemic disrupted the administration of TOEFL iBT at test centers, leading to the introduction of TOEFL iBT At Home Testing (formerly TOEFL iBT Home Edition) (Papageorgiou & Manna, 2021). While stakeholders' preferences regarding at-home testing vary, online testing has become increasingly prevalent in language assessments, with IELTS and ACTFL also adopting this format. However, concerns about the security and validity of at-home language tests have been raised (Isbell & Kremmel, 2020). Despite growing interest, limited research exists on test-takers' perceptions and experiences of at-home testing. This netnographic study (Kozinets, 1997) collected 1685 posts from online forums used by Korean test-takers discussing the TOEFL iBT At Home Testing (Kim, 2017). Korean test-takers were chosen due to their significant presence as TOEFL examinees (Malone & Montee, 2014) and international students in English-dominant countries (Kim, 2017). Initial analysis of 102 posts revealed common topics such as equipment and environment requirements, score report dates, scheduling and rescheduling, and general reviews of test-taking experiences. Interestingly, test-takers also discussed issues related to malpractice prevention, sharing experiences of test cancellations due to noise during breaks. These candid insights provide valuable information on test-takers' concerns and shed light on the challenges associated with at-home testing. By examining test-takers' perceptions and experiences, this study contributes to a better understanding of the complexities involved in at-home language testing. The findings offer insights into test-takers' perspectives and can inform improvements in test administration, ensuring the validity and reliability of at-home language assessments.

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

Author(s): Kyle Smith

Key words: Bourdieu, Critical Language Testing, Power

Abstract: 30 years ago, Elana Shohamy (1993) observed that language tests appear to have an unrivalled power to alter test-takers' behaviour and dictate decisions made by educators and policymakers. Shohamy (2001) explored the nature and source of this power in ground-breaking scholarship at the intersection of language testing, the philosophy of Michel Foucault and the sociology of Pierre Bourdieu. This paper takes up the latter thread in Shohamy's work to further investigate the 'power of tests,' emphasising the crucial Bourdieusian concept of reflexivity. In language testing research, Bourdieusian reflexivity means that researchers 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 analysis of assessment practices in Direct Entry Programs based on documents and interviews with staff from two university English centres in Australia. The paper argues that the social relations within the language testing field in 1979, at the time of the first Language Testing Research Colloquium, are instituted in Direct Entry Programs in Australia today. In other words, the contemporary 'power of tests' even at a local level can be understood as originating in the historical relations among language testers and the large-scale standardised assessments they have constructed and legitimised.

The relationship between writing tasks and second language writers' written stance

Author(s): Giang Tran

Key words: Written stance, writing tasks, English as a foreign language (EFL)

Abstract: Stance has been found to be a key linguistic feature that is mastered by expert writers of academic English but at the same time is a major source of challenge for novice writers. The current study is set in the context of Vietnam, a country in serious need of qualified EFL teaching staff. In one of Vietnam's leading foreign language teacher training institutions, written stance receives insufficient treatment in instruction but is commonly integrated in the assessment criteria for writing in high-stakes standardized tests. Moreover, little has been known regarding the interaction between the writing tasks featured in these tests and the learners' use of stance markers. Therefore, the current mixed-method study aims to examine how Vietnamese EFL learners use linguistic resources to express stance and the factors that influence such language choices, with a specific focus on features of the writing tasks. The data includes a corpus of 550 English essays written by 110 EFL learners from one university. Each participant wrote five essays in response to five writing tasks differing in terms of topics, genres, and audience specifications. The same participants also responded to a short task-perception questionnaire and a longer writing-perception questionnaire. The corpus was annotated based on Hyland's (2011) framework of interactional metadiscourse, and the corpus statistics were then analysed together with quantitative questionnaire responses in linear regression models. Responses to open-ended questionnaire items will be analysed qualitatively to complement the quantitative findings. Initial quantitative findings have indicated that the essay writing prompts played a major role in predicting the participants' use of stance markers, while their university study majors also exerted a moderate degree of influence. The findings will hopefully have pedagogical implications for EFL writing teachers, writing test designers, and teacher educators in Vietnam and in similar EFL or related contexts.

The State of Classroom-based Assessment in Japan

Author(s): Adam Murray, Taiko Tsuchihira

Key words: teachers' beliefs, teachers' self-efficacy, assessment practices

Abstract: Although the field of teachers' self-efficacy is well-established in the field of general education, relatively little research has been published about language teachers' self-efficacy (see Klassen et al., 2011, Wyatt, 2018). During the 2019 academic year, the researchers started development on an instrument named the Classroom-based Assessment Self-Efficacy Scale (CBA-SES) which consisted of three sections: teachers' beliefs, teachers' self-efficacy, and teaching practices. This instrument was first piloted with 30 teachers. Based on the participants' feedback, the researchers created additional items and added another section. This resulted in a revised questionnaire with four sections: teaching context (5 items), teacher beliefs (11 items), self-efficacy (10 items), and teaching practices (12). We administered this instrument to a convenience sample of 29 Japanese Teachers of English (JTEs) to get a better understanding of JTEs working in various teaching contexts (primary, secondary, post-secondary) with the intention of revising the instrument and conducting a large-scale study during the 2023 academic year. The respondents at the university level had the strongest beliefs. Also, there is a connection between experience and beliefs. As JTEs gain more classroom experience, their beliefs become stronger. Finally, teachers believe there is a need for assessment tasks that resemble real-life language use. In the first semester of the 2023 academic year, we will administer the latest iteration of instrument (61 items) on a nationwide scale to both JTEs and native speakers of English. Our presentation will report on the further development of our instrument to better measure teachers' attitudes towards CBA and their actual implementation of CBA. We will also share preliminary results about the data that has been collected. We will conclude with suggestions for teacher education programs for pre- and in-service language educators.

To what extent does an instructional rubric affect Japanese junior high school students' Eiken writing performance and assessment literacy?

Author(s): Chiho Young-Johnson, Noh Kawase, Jerami L. Vanderholm

Key words: rubrics, writing performance, assessment literacy

Abstract: Rubrics for evaluating writing can provide information to teachers about students' performance and proficiency (East & Cushing, 2016) as well as instruction to students about effective writing (Andrade, 2001). Previous studies have explored the influence of rubrics on students' writing performance and assessment literacy (Andrade, 2001; Iwamoto, 2020; Sundeen, 2014; Turgut & Kayaoğlu, 2015), yet few have discussed the specific benefits and methods of rubric presentation in classroom practice. To explore how instructional rubrics influence Japanese secondary school students' Eiken test writing performance and their assessment literacy, this study examined the performance of students on three Eiken-based writing tasks, as well as students' answers on questionnaires. This research was conducted in third-grade classes in a public secondary school in Japan. First, both students in the experimental (N=23) and control groups (N=23) wrote the first essay on the same topic without receiving a rubric or instruction. Only students in the experimental groups received instructional rubrics and instruction for the second and third essays. The study found that while students in the experimental groups showed a slight improvement in overall scores, there were no significant differences between performances before and after the presentation of rubrics, nor were there notable differences in performance between the control and experimental groups. On the other hand, questionnaires revealed marked differences in terms of their assessment literacy. The experimental groups became more aware of not only grammatical and lexical accuracy but also lexical diversity, while the control groups tended to focus on accuracy. These findings have important implications for writing prompts and rubrics as well as for writing pedagogy in Japan and abroad. It is proposed that presenting and explaining rubrics can have a positive influence on students, especially if the language of the rubric is adapted to their age and level of familiarity with assessment criteria.

Unpacking the Drivers of Citation Counts in Language Assessment Research: A Bibliometric Study

Author(s): Zhang Sai, Vahid Aryadoust

Key words: Bibliometrics, Citation counts, Language assessment

Abstract: With the growing significance of evaluative bibliometrics in academia, citation counts have been extensively acknowledged as a good proxy to quantify the impact of research output. Factors affecting citation counts have also become a topic of interest among scholars in various disciplines, but remain poorly understood in the language assessment literature. To fill this gap, in this study we examined the potential drivers of citation counts besides scientific quality, using a dataset of 479 articles published in four Quartile-1 language assessment journals during a five-year window (2017-2021). The detailed bibliometric data of each article, covering journal-related, author-related and paper-related features, were retrieved from Scopus (as of May 2023). For statistical analysis, we first illustrated the citation landscape of language assessment research over time. A negative binomial regression model was adopted, which allowed us to assess the association between multiple predictors and the over-dispersed count outcomes. The results revealed that among the total ten predictor variables, seven exerted significant effects on citation counts: publication venue, h-index (first author), international co-authorship, region of author's affiliation, article age, characteristics of references, and subfield of language assessment. However, no statistical significance was found in terms of number of authors, mode of access, and funding. This suggests that citation practices in the language assessment community are associated with a number of key attributes of a publication besides scientific merit. Subsequent discussions in this area of research are encouraged to help language assessment researchers improve their article citations and academic performance.

Validation of language self-assessment descriptors for use with the Defence Force

Author(s): Ksenia Zhao

Key words: Self-assessment, validation, military and LSP assessment

Abstract: The use of language self-assessment (SA) as a learning tool has been relatively well-established, however its efficacy for use in particular workplace contts a replacement for more conventional language tests is yet to be determined. This study explores the accuracy and usefulness of a new SA tool developed for use in the Australian Defence Force (ADF), a context where language instruction and assessment involve multiple languages and proficiency levels. In the ADF, language assessment in Languages other than English (LOTE) is guided by the Australian Defence Language Proficiency Rating Scale (ADLPRS). Recently, the ADF, in collaboration with language testing experts, developed a self-assessment instrument to accompany the ADLPRS for use as a language screening tool when no conventional LOTE test is available, or when test administration would be impractical. In this session I will describe the methodology for developing the scale, and the methods used to explore its accuracy (by, for example, comparing the self-assessed level of 98 trial participants on the new SA tool with the recent scores of these participants on a validated ADF LOTE proficiency test). Additionally, I will outline the approach adopted to gauge the usefulness and practicality of the SA tool through analysis of participant feedback (both quantitative and qualitative) on the experience of completing the SA. Finally, I will report on some preliminary findings of this research and reflect on its practical implications for the testing of military personnel who use LOTE in their work and theoretical implications for SA scale design for workplace contexts.

What are the Barriers to Developing English-as-a-Second/Foreign-Language (L2) Learners' Metaphor Awareness? Evidence from the Development and Validation of Metaphor Awareness Instruments

Author(s): Ting Ma, Lawrence Jun Zhang, Judy Parr

Key words: metaphor awareness, instruments development and validation, L2 learners

Abstract: Studies have shown that raising L2 learners' metaphor awareness contributes to the acquisition of figurative language and development of language skills (see for instance, Littlemore et al., 2014; MacArthur, 2021; Wang et al., 2020). However, the instruments measuring metaphor awareness, in the majority of relevant research, seemed not to have undergone robust methodological procedures of checking validity and reliability, thus compromising the authenticity of the measurement and problematising the interpretation of the results from the measurements adopted (O'Reilly & Marsden, 2021). In addition, both theoretical and empirical research tends to frame metaphoric competence within the territory of linguistic and conceptual metaphors, neglecting the communicative functions of metaphor in discourse (O'Reilly & Marsden, 2021). As an attempt to fill these research gaps, this study developed three instruments—two tests and a questionnaire—for measuring L2 students' metaphor awareness in linguistic, conceptual, and communicative dimensions. Our Rasch analysis of data from 293 Chinese undergraduate English majors show that the instruments demonstrate good validity and reliability. The participants were well aware of the communicative functions of metaphor but found it challenging to identify metaphorical prepositions, adverbs, and adjectives. Establishing the correspondences between the source domain and target domain of conventional conceptual metaphors was equally daunting. Based on these findings, we conclude this presentation with a discussion of the implications for L2 metaphor measurement and classroom instruction.

What linguistic features do raters rely on when rating borderline cases?: Empirical examination of Complexity, Accuracy, Fluency of OPI tests

Author(s): Myoyoung Kim, Jee Eun Gaetz, Sunkwang Bae

Key words: speaking proficiency test, Complexity, Accuracy, Fluency (CAF) , borderline performance

Abstract: Complexity, Accuracy, Fluency (CAF) dimensions have been widely employed in studies to examine distinguishing features of performance across different levels in speaking tests. (Kang and Yan, 2018; Seedhouse et al., 2014; Yan et al, 2021). However, research has mainly been conducted on English tests. The current study investigated what CAF features differentiate borderline performance in Korean (KOR) Oral Proficiency Interview (OPI) administered at Defense Language Institute Foreign Language Center (DLIFLC). The OPI at DLIFLC is a high-stakes standardized assessment instrument measuring speaking abilities in foreign languages, using the Interagency Language Roundtable (ILR) as a rating scale. The dual-rating system of the OPI yields 3 possible rating types: 1) agreed-by-raters; 2) initially split between two adjacent levels and finalized as the lower level; 3) initially split and finalized as the higher level. Varied characteristics of borderline performance are a potential source of split rating, which affects the degree of inter-rater reliability. The purpose of the study is to examine what CAF features of borderline cases induce different rating types. A total of 862-minutes Korean (KOR) OPI samples were selected at 5 ILR levels ranging from 1 to 3. Each level includes samples of 3 rating types. After being transcribed using KOR markup guidelines, the KOR samples were coded according to a KOR CAF codebook comprised of 39 lexical, grammatical, discourse, fluency features and error types. The results of one-way ANOVA analysis showed that CAF features distinguishing 3 rating types vary per level. There was a significant effect of rating type on fluency features across levels except level 1, and on the number of errors at level 2. The findings will provide evidence for the evaluation inference of the OPI validity argument. Implications of the findings for rater training, limitations of the study and a way forward will be addressed as well.

Who will teach the teacher educator? Findings and implications of promoting Brazilian teacher educators' Language Assessment Literacy

Author(s): Isadora Teixeira Moraes

Key words: language assessment literacy, teacher educators, continued education

Abstract: In Brazilian educational policies, the goal of assessment is to promote learning (Brasil, 2018), but, in practice, its summative function prevails, and it is often used as a tool for control and punishment (Scaramucci, 2006). The lack of assessment training in language teacher education programs in Brazil (Quevedo-Camargo, 2020) and elsewhere (Stiggins, 1991) is well-reported in the literature. Therefore, this research's main goal was to identify and promote Brazilian English teacher educators' language assessment literacy (LAL) (Giraldo, 2018; Inbar-Lourie, 2008), aiming at a positive impact in these educators' praxis (Freire, 2014) and contexts. To do so, a critical qualitative methodological framework characterized as case study was adopted (Cohen, Manion & Morrison, 2018). Participants were 15 teacher educators from seven public universities in the state of Paraná, Brazil, who took part in an online language assessment workshop. Data generation instruments were a) transcribed workshop sessions; b) learning artifacts; and c) questionnaires. Findings show that the three LAL components – knowledge, skills and principles – are present in the policies that guide the educators' programs, but, in their praxis, the knowledge and skills are more prominent. Tasks in the workshop that promoted educators' LAL the most were those that fostered collaboration among participants, leading to the development of a community of practice (Wenger, 1998), and those that promoted their knowledge in LAL, which differs from research with pre-service teachers (Giraldo & Murcia, 2018). The most recurrent implications were the broadening of the knowledge and principles in LAL. Some educators also mentioned actions in terms of promoting other stakeholders' LAL, such as students and colleagues. We therefore advocate in favor of more initiatives for collaborative and contextually situated LAL development for teacher educators, as this development might, besides improving these stakeholders' concepts and practices, have a cascading effect on other stakeholder groups.