MwALT 2023

24th Midwest Association of Language Testers Conference | September 8-9

Language Assessment in Multilingual and Multicultural Contexts

MwALT
Midwest Association of Language Testers

duolingo english test

Innovation in assessment

The Duolingo English Test leverages the latest assessment science and human-in-the-loop AI to empower anyone to test where and when they're at their best.

Digital-first

The Duolingo English Test leverages AI to personalize itself in real-time to every test taker, honing in on their true proficiency more quickly, precisely, and securely than traditional fixed-form tests.

Accessible administration

The Duolingo English Test is designed to radically improve English language proficiency assessment for test takers and score recipients alike, by providing a testing experience that is accessible while remaining accurate and secure.

Secure certification

Unprecedented security technology and a 1:2 candidate-to-proctor ratio ensure that the Duolingo English Test is extremely secure. Every proctor has access to AI tools that monitor dozens of categories of biometrics and behavioral data.

Explore our research

Keep up to date on the latest DET news

englishtest.duolingo.com/research

WIDA Summer Research Internships

Located at the University of Wisconsin–Madison, WIDA offers summer research internships in language assessment to graduate students. Interns will participate in WIDA assessment research projects and collaborate with WIDA researchers on projects that address academic language development in the K–12 context. Research interns have co-presented their work with WIDA researchers at conferences such as LTRC, MwALT, ECOLT, and NCME.

The WIDA Assessment Team pursues a validation research agenda that supports the WIDA suite of language assessments: WIDA Screener, WIDA MODEL, and ACCESS for ELLs. Interns may contribute to various aspects of this research agenda, such as study design, data collection and analyses, manuscript/report authoring and review, and presentation of findings. Quantitative, qualitative, and mixed-methods projects may be assigned to interns, depending on the background of qualified applicants.

Eligibility

- Full-time enrollment in a doctoral program related to language assessment
- Completion of a minimum of two years of coursework toward a doctoral degree, prior to beginning the internship

For more information, visit wida.wisc.edu/about/careers/internship

Apply Now!

Contact widainternships@wcer.wisc.edu

WIDA™

MET

Unlock Success

Certify your English proficiency with the Michigan English Test (MET).

Prove your English

- ✓ Take MET at a test center and at home
- ✓ Available any day of the year
- ✓ Single-section retake: listening, reading, or writing
- ✓ Official results in 5 days

Achieve your goals

- ✓ Apply to colleges or universities
- ✓ Obtain nursing licensure
- ✓ New career opportunities with English language certification
- ✓ Occupational visa to work in healthcare in the United States

Michigan Language Assessment helps people achieve their educational and career goals by providing trusted English language exams that draw on the expertise of two of the world's leading universities.

MICHIGANASSESSMENT.ORG

SPONSORS

School of Literatures, Cultures & Linguistics
Second Language Acquisition and Teacher Education Program
Department of Linguistics
Department of Spanish & Portuguese
Department of East Asian Languages and Cultures
Center for East Asian & Pacific Studies
Center for Advanced Study
European Union Center at Illinois

TABLE OF CONTENTS

Welcome to MwALT 2023 at the University of Illinois at Urbana-Champaign!..... 1

Conference Organization..... 2

Plenary Sessions..... 7

Considering multilingual assessment: Twenty questions—and the beginnings of a response..... 7

Assessing young multilingual learners' language development: Myths, challenges, and promises..... 8

Pre-Conference Workshops..... 9

Best Paper Abstract..... 12

Paper and Demo Abstracts..... 13

Work-in-Progress Abstracts..... 36

Poster Abstracts.....37

Wi-Fi Access: Using IllinoisNet Guest

1. While on campus, connect your device to the Wi-Fi network "IllinoisNet_Guest"
2. A browser window should automatically open; this is our Wi-Fi Captive Portal Page. If this does not happen, open your web browser and navigate to http://illinois.edu.
3. This page will prompt you for your NetID and password to log in. **As a guest or visitor, click the link at the bottom of the page and you can self-register an account.** This account will be valid until 4am the next day, after which you will be prompted to create a new account.
4. After a brief moment the system should then inform you that you have been given Internet access. Visitors should choose **Click Here for Wi-Fi Access**. Keep in mind that any visitor accounts created in this way are only valid until 4am the following day, after which you will need to repeat this process.

Welcome to MwALT 2023
at the University of Illinois at Urbana-Champaign!

Welcome you all to the 24th annual conference of the Midwest Association of Language Testers (MwALT)! The Department of Linguistics and the Department of Spanish and Portuguese at UIUC are delighted to host the MwALT conference. This year's conference theme is 'Language Assessment in Multilingual and Multicultural Contexts'. We hope that this upcoming conference will provide a platform for us to showcase language assessment research that represents a larger variety of contexts and in languages that are underrepresented. Assessment of languages other than English (LOTEs) and language users in bi/multilingual contexts is underrepresented in language testing scholarship, and yet the larger field of applied linguistics has taken a multilingual turn to better represent language use in the real world. At MwALT 2023, we hope to invite more conversations about the similarities and uniqueness of language assessment in multilingual and multicultural contexts.

MwALT is a strong community of graduate students, faculty, and professionals that supports research and practice in language testing and assessment. We place great emphasis on supporting graduate-student research and introducing students to the broader field of language assessment and evaluation. This is achieved through annual awards for student papers and presentations, as well as pre-conference workshops on research methods and assessment skills. In spite of these events and activities, we strive to remain a conference that is affordable to all.

We are grateful for the continuing support of participating universities, organizations such as Educational Testing Service, Michigan Language Assessments, WIDA, and Duolingo; as well as the many departments and areas of research centers at UIUC that sponsor our annual conference and enable us to sustain our mission.

Xun Yan
MwALT President

Conference Organization

Organizing Committee

Xun Yan

Melissa Bowles

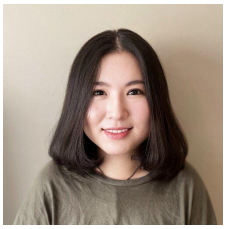
Rurik Tywoniw

Ha Ram Kim

Qiusi Zhang

Ping-Lin Chuang

Huiying Cai



Yulin Pan

Lorena Alarcon

Yu-Wen Lien

Abstract Reviewers

Lorena Alarcon
Shireen Baghestani
Kate Challis
Mark Chapman
Lixia Cheng
Ping-Lin Chuang
Matt Coss
Koen Van Gorp
Haeyun Jin
Haeun Kim
Sebnem Kurt
Erika Latham
Wenyue Ma
Reza Neiriz
Gary Ockey
Aaron Ohlrogge
Agustinus Prasetyo
Haoshan Ren
Sun-Young Shin
Sharareh Taghizadeh Vahed
Paula Winke
Monique Yoder
Qiusi Zhang

Friday, September 8, 2023

Time	Session
9:00 am - 12:00 pm	***Workshop 1: Reviewing for Bias and Sensitivity: Beyond Checklists*** Margaret E. Malone, Camelot Marshall *Location: Room 338, Davenport Hall*
2:00 pm - 5:00 pm	***Workshop 2: Applications of Gaussian Mixture Model in Language Testing and Speech Research*** Yan Tang, Huiying Cai *Location: Room 338, Davenport Hall*

Saturday, September 9, 2023 (Venue: Levis Faculty Center, 919 W Illinois St, Urbana, IL)

Time	Session			
8:00 am - 8:15 am	***Welcome and Opening Remarks*** James Yoon *Location: Room 300*			
8:15 am - 9:15 am	***Plenary 1: Considering Multilingual Assessment: Twenty Questions—and the Beginnings of a Response*** Beverly Baker *Location: Room 300*			
9:15 am - 9:45 am	Coffee Break 1			
9:45 am – 12:00 pm	Concurrent Sessions 1			
	Room 210	Room 304	Room 108	Room 208
9:45 am - 10:15 am	***Test takers' interaction with context videos in a video-based listening test: A conceptual replication and extension of Suvorov (2015)*** *(Best Student Paper)* Jieun Kim	***Promoting multi-cultural and multi-lingual assessment practices among k-12 language arts teachers*** Lucy Belomoina, Iftikhar Haider	***Developing a computer delivered ASL test for diverse audiences: A three-phase feasibility study*** Meg Malone, Camelot Marshall	–

10:20 am - 10:50 am	***Investigating the relationship between automated measurements of nonverbal behavior and language proficiency ratings*** Dylan Burton	***Examining the direct and indirect impacts of verbatim source use on linguistic complexity in integrated argumentative writing assessment*** Huiying Cai	***Using international English proficiency tests in EMI contexts: Exploring the relationships between student performance on TOEFL iBT and EMI course assignments in Korean universities*** Soohye Yeom	***What does elicited imitation measure? Findings from strategy use on an elicited imitation test of Chinese*** Yuyun Lei
10:55 am - 11:25 am	***The influence of Grammarly and ChatGPT on rater behavior and the halo effect*** Jim Ranalli, Shireen Baghestani	***Effects of the type of audio-visual input and listenability on young learners' listening comprehension*** Sun-Young Shin, Senyung Lee	***Uncovering the dimensionality of authorial voice in the integrated writing task: A Confirmatory Factor Analysis study*** Lia Plakans, Yafei Ye, Kwangmin Lee	***Typing vs. hand-writing on L2 writing tasks: Examining test-taker outcomes and perceptions*** Matt Coss, Paula Winke
11:30 am - 12:00 pm	***Translanguaging in assessment: A scoping review*** Chia-Hsin Yin	***Automating silent pause marking and computing fluency measures: A demonstration using ELAN and CLAN*** (Technology Demonstration) Qiusi Zhang	***Linking authentic materials and assessments with Korean language learners' outcomes*** Weejeong Jeong	***Co-occurrence of disfluency features of L2 speech across proficiency levels in controlled and spontaneous tasks*** Yulin Pan
12:00 pm -1:00 pm	Lunch and Business meeting			
1:00 pm - 1:50 pm	Posters and Work-in-Progress Sessions *Location: Room 300*			
	Posters			
	Virtual administration of an oral English proficiency test: Challenges and student perceptions Sharareh Taghizadeh Vahed ***The relationship between response time and performance on a large-scale writing assessment: Examining testing time for English learners*** Junhee Park, Annette Vernon, Jeongmin Ji ***A template for designing self-assessments for LCTL programs*** Paula Winke, Laura Lopez			

	The role of educators in WIDA bias, sensitivity, and content reviews Jason Kemp, Heather Elliott ***An investigation of longitudinal development of integrated writing among Chinese English as a Foreign Language (EFL) learners*** Kwangmin Lee, Yafei Ye, Junhee Park, I-Chun (Vera) Hsiao ***Czech as a foreign language assessments in 2023*** Kate Challis ***Developing an assessment criteria framework for digital multimodal compositions*** Stephanie Lissett Vigil, Chia-Hsin Yin ***Revisiting the issues in L2 proficiency assessment in SLA research*** Jinman Fu, Jungwon Yoo, Sun-Young Shin ***Examining the use of sources in AI-generated writing test responses*** Rurik Tywoniw			
	Work-in-Progress			
	Critical literacy and self-assessment: A study of Spanish students' literacy on sociolinguistic issues Lorena Alarcon			
1:55 pm – 3:00 pm	Concurrent Sessions 2			
	Room 210	Room 304	Room 208	
1:55 pm - 2:25 pm	***The impact of source use features on raters' judgment of argumentation*** Ping-Lin Chuang	***Notetaking as validity evidence: A mixed-methods investigation of question preview in listening assessment*** Rebecca Yeager, GoMee Park, Ray J.-T. Liao	***Using ChatGPT for writing evaluation: Examining the accuracy and reliability of AI-generated scores in comparison to human raters and potential reasons for discrepant ratings*** Shuhui Yin, Haeun Kim, Yasin Karatay, Shireen Baghestani, Jeanne Beck, Leyla Karatay, Sebnem Kurt, Mutleb Alnafisah	—

2:30 pm - 3:00 pm	***The effect of adding an interactive avatar to SDS-mediated oral discussion tasks of an oral communication test: A test taker perception study*** Reza Neiriz	***Using text features to predict rater disagreements*** Shireen Baghestani	–	–
3:00 pm - 3:30 pm	Coffee Break 2			
3:30 pm - 4:00 pm	***Measuring complexity features in young learner spoken language: Human versus contextually applied automated coding*** Haeun Kim, Gordon Blaine West	***The construct of summary writing: how do reading, writing, text features, cognitive operation difficulty impact summarization as an integrated task*** Yu-Chen Tina Lin	***Pre-post elicited imitation: Documenting proficiency development in a two-semester EAP program*** Lixia Cheng, Xiaorui Li	–
4:00 pm - 5:00 pm	***Plenary 2: Assessing Young Multilingual Learners' Language Development: Myths, Challenges, and Promises*** Becky Hsuanhua Huang *Location: Room 300*			
5:00 pm - 6:00 pm	Reception and Award Ceremony *Location: Room 300*			

Plenary Sessions

Considering multilingual assessment: Twenty questions—and the beginnings of a response

Beverly Baker, Director, Language Assessment, Official Languages and Bilingualism Institute, University of Ottawa

Describing the field of multilingual assessment is not at all straightforward; the concept has many meanings, including the assessment of multiple languages (as separate entities or in combination) or the assessment of what can be termed a single language, but in a multilingual context. Whichever perspective you take, it seems necessary to accept that focussing attention on single named language in isolation, as if no other languages exist in a given context, is rarely realistic or appropriate and is often counterproductive. This statement holds true whether we are considering language teaching, assessment, or educational policy.

From this starting point, I began by asking myself two basic questions:

- How is multilingual assessment perceived differently by scholars working with a background of language assessment as opposed to multi/plurilingualism?
- What discoveries have been made by scholars attempting the development of assessment tools that target language use of multiple languages—or across named language boundaries?

While these questions were a useful starting point for my reflections, they resulted in another set of questions! I will share all these questions with you and I will encourage you to add your own. Together they may form the basis of a research program proposal to combine rather than divide the efforts of colleagues within applied linguistics on this topic, which has proven to be surprisingly controversial.

Beverly Baker is Director, Language Assessment, at the Official Languages and Bilingualism Institute at the University of Ottawa, Canada, where she holds the Research Chair in Language Assessment Literacy. She works and publishes widely in the areas of language test validation, language teacher development, multilingual assessment and language assessment competency development. She also consults widely with professional organisations and governments on language testing issues. She is currently Treasurer of the International Language Testing Association and a founding member of its Language Assessment Literacy Special Interest Group. In 2019 she received the International Assessment Award from the British Council.

Plenary Sessions

Assessing young multilingual learners' language development: Myths, challenges, and promises

Becky Hsuanhua Huang, Professor, Department of Teaching and Learning, Ohio State University

Due to globalization and immigration trends, the number of children growing up learning more than one language has been growing rapidly around the world. These young multilingual learners may learn an additional language in a majority/societal language immersion context (e.g., learning English in the United States where English is a societal language) or in a minority/foreign language instructed learning context (e.g., learning English in Taiwan or in Mexico where English is a foreign language). The increase in this population comes with the parallel need for fair and effective assessment of their multilingual development.

In this talk, I will focus on the 4-13 age group (i.e., kindergarten to 8th grade in the US education system) because it's a critical stage of development in language as well as cognitive and social-emotional domains. Assessments of their language development also have strong relevance and implications for education. I will first discuss the myths related to assessing this population. I will then address the challenges of assessing multilinguals in three key areas: 1) the heterogeneity in the young multilingual population, 2) the connections between language and content, and 3) the difficulties in differentiating between typically-developing multilingual development and language learning difficulties (also known as developmental language disorders). I will end the talk with a discussion of the future of multilingual assessment, considering the roles of technology (e.g., Artificial Intelligence) and the application of translanguaging in assessing young multilinguals.

Becky Hsuanhua Huang is a Professor in the Department of Teaching and Learning at Ohio State University, where she is also a faculty associate with the Crane Center for Early Childhood Research and Policy, and affiliate faculty with the Center for Languages, Literatures, and Cultures. Her interests include language assessment and language acquisition, particularly among young multilingual learners. Dr. Huang currently serves on the Editorial Board of Language Testing, Language Assessment Quarterly, and TESOL Quarterly. She has directed two NIH grants focusing on multilingual learners' language/literacy development, and is currently co-directing an IES grant that aims to develop language-focused reading intervention for multilingual learners in upper elementary grades. This year (2023), she won the Mid-Career Award of the AERA Second Language Research Special Interest Group.

Pre-Conference Workshops

Reviewing for bias and sensitivity: Beyond checklists

Margaret E. Malone & Camelot Marshall

Friday September 8: 9 am to 12 pm, Room 338, Davenport Hall

Developing language test items is a huge responsibility and requires adherence to ethical principles and practices. While checklists for item development and review are necessary for the overall process, additional processes are essential to review for bias and sensitivity. This half-day workshop will begin with an overview of the ILTA Code of Ethics and Guidelines for practice. Next, participants will engage with each other to wrestle with realistic scenarios for item writing, review, and production. Based on the discussions from this small group work, the workshop coordinators will share some principles, practices and processes, and practicalities used to help item developers select, review and revise test items, and produce tests. Participants will develop both concrete and abstract approaches to rigorously review items and test forms.

Margaret E. Malone (Ph.D., Georgetown University) is Director of Assessment and Research at ACTFL. She has over three decades of experience in language test development, materials development, delivery of professional development, and teacher preparation through both online and face-to-face methods, data collection and survey research, and program evaluation. Her current research focuses on language assessment literacy, oral proficiency assessment and the influences of the Seal of Biliteracy on language teaching and learning. Previously, she was Associate Vice President of World Languages and International Programs at the Center for Applied Linguistics. Earlier in her career, she served as the first Language Testing Specialist for Peace Corps-Worldwide, where she managed oral language testing for 60 countries and 120+ languages.

Camelot Marshall, PMP, (Ph.D., Bryn Mawr College) is Principal Assessment Manager at ACTFL. She has over 30 years' experience teaching Russian and graduate level second language acquisition courses, as well as designing, developing, and implementing test design and item writing processes. At present, she manages ACTFL's test development program and employs innovative methods and approaches to develop world language proficiency and performance assessments. Prior to joining ACTFL, she was the assessment manager at American Councils for International Education, where she was instrumental in developing and implementing test development processes for ten critical languages.

Pre-Conference Workshops

Applications of Gaussian Mixture Model in Language Testing and Speech Research

Yan Tang & Huiying Cai

Friday September 8: 2 pm to 5 pm, Room 338, Davenport Hall

This is a world filled with Gaussian/Normal distributions. Quantitative data harvested in many practical domains, such as demography, marketing, as well as language testing, can be modelled by Gaussian distribution. Even a real-valued random variable whose actual distribution is unknown can also be assumed to have a Gaussian distribution. It is, therefore, reasonable to suppose that samples sharing a set of 'similarities' and those sharing another set of 'similarities' may constitute distinct groups generated from different Gaussians in either a univariate or multivariate space. With such a fundamental idea, Gaussian Mixture Model (GMM) as a machine learning algorithm models the internal groupings (e.g. clusters) as a mixture of several Gaussian Distributions; each of the clusters is defined by its own mean(s) and (co)variances. As such, GMM can be a useful methodological approach to identifying test-taker and performance profiles in language testing research. This workshop gives a broad introduction to the principle of GMM and how GMM can be used to perform unsupervised clustering and semi-supervised learning in language testing research. The workshop consists of two sessions. The first session will be a lecture-based meeting. During the second session, there will be a tutorial with several hands-on exercises on building GMM-based clustering models using Python for analysing datasets of speech and language testing. Prior Python skills are not essential; the base code for the exercises will be provided during the session.

Yan Tang is an Assistant Professor in Speech Processing at the Department of Linguistics, University of Illinois Urbana-Champaign since 2019. Before that, he was a research fellow and senior research fellow for five years, working on a UK EPSRC project funded by the UK government and the BBC. Yan Tang completed his Ph.D. in Applied Linguistics with Sobresaliente Cum Laude at Universidad del País Vasco, Spain in 2014, focusing on computational speech and hearing. He also received a master's degree in Software Systems and Internet Technology with Distinction at the University of Sheffield, UK in 2008. His research lies in a wide range on the spectrum of speech and hearing science from speech intelligibility enhancement, computational modelling of the auditory system, blind source separation to human speech production and perception in non-ideal listening conditions. His current research interests include computational modelling of informational masking in speech perception, speech-noise interactions in the

spectro-temporal-modulation domain and their impact on speech understanding, and paralinguistic speech processing using machine learning.

Huiying Cai is a fourth-year Linguistics Ph.D. student at the University of Illinois at Urbana-Champaign (UIUC). She is a research assistant for the UIUC English Placement Test, assisting in test development and quality control. Her research interests mainly lie in language testing, with a particular focus on the interplay of language testing and computational linguistics. Her studies used natural language processing techniques on test-relevant corpora (e.g., rater comments and student essays) to provide empirical evidence for test validity arguments. Currently, her research intends to explore an automated scoring system for integrated writing tasks in a local context using machine learning techniques.

Best Paper Abstract

Test takers' interaction with context videos in a video-based listening test: A conceptual replication and extension of Suvorov (2015)

Jieun Kim

Saturday September 9: 9:45 am - 10:15 am *Location: Room 210*

The present study conceptually replicates and extends Suvorov's (2015) investigation of visual information and viewing behavior, examining the relationship between foreign language listening anxiety (Elkhafaifi, 2005) and viewing behavior in context-providing videos. In Experiment 1, 49 participants completed the Foreign Language Listening Anxiety Scale (FLLAS) (from Zhang, 2013), watched three academic lectures from Suvorov (2015), while their eyes were tracked, and answered comprehension questions. In Experiment 2, 64 participants were additionally encouraged to take notes, and the order of the videos was counterbalanced. The relationship between viewing time and comprehension was small and non-significant in both experiments, except when considering note-taking. No listening anxiety effect was found in either experiment. Survey responses and interviews with 31 participants revealed variations in test takers' attitudes toward the videos. The findings indicate that including context videos in listening tests to enhance authenticity does not enhance or compromise comprehension. Additionally, the individual difference factor, namely listening anxiety, did not significantly influence how much individuals look at the videos. The implications for the development of video-based listening tests, at-home testing, and research methodology are discussed.

Paper and Demo Abstracts

Promoting multi-cultural and multi-lingual assessment practices among k-12 language arts teachers

Lucy Belomoina, Iftikhar Haider

Saturday September 9: 9:45 am - 10:15 am *Location: Room 304*

An important part of our mission as teachers of English is supporting and promoting cultural, linguistic, and rhetorical diversity that students and colleagues bring into our courses and programs. Our commitment to such a multicultural and multilingual pedagogical context calls for a new language assessment understanding of both K-12 language arts teachers' pedagogical practices and writing instructors at the college level. Data collected during semester-long empirical research conducted in an introductory linguistics course in spring 2017 at a midwestern university demonstrated that the scientific study of a nonstandardized English language variety, African American English, aided in future elementary and middle school teachers' objectivity vis-à-vis language. This presentation offers effective strategies on how to design assessment tools on the language patterns that diverse students bring with them to school in their acquisition of standardized English and educate K-12 students on the organization and logic in vernacular dialects through the study of their prominent linguistic and paralinguistic features. The presenters will also share implications for sociolinguistic and sociocultural components of language knowledge, particularly focusing on the way in which these components are assessed to convey a range of sociolinguistic meanings in classroom interactions. Suggestions for learning-oriented language assessment tasks, lingua franca communication where norms are situated, and problems related to native speaker benchmarking for classroom assessments will also be discussed. Furthermore, presenters propose a more balanced approach that equally uses and counts both standard-based assessment and alternative assessments as a step toward culturally and linguistically relevant assessment.

Paper and Demo Abstracts

Developing a computer delivered ASL test for diverse audiences: A three-phase feasibility study

Meg Malone, Camelot Marshall

Saturday September 9: 9:45 am - 10:15 am *Location: Room 108*

The Seal of Biliteracy (SOBL) provides opportunities for high school students to certify their proficiency in two languages. However, ASL has been underrepresented in this area; moreover, the diversity of ASL users (L1, children of Deaf adults (CODAs) and world language learners) makes developing and administering such a test not only complicated but a significant equity issue. This work was guided by an external advisory committee comprised of high school and university based ASL instructors and assessment and curriculum experts.

This poster describes a three-phase process to determine the feasibility of a computer-based ASL test for receptive comprehension (RC) and interpersonal interaction (II). The first phase (N=9) of the study focused on the technological affordances of existing technology and whether it could accurately employ matching and multiple-choice responses to RC tasks as well as administer and capture II between a learner and a semi-direct interlocutor. This phase focused on world language learners and their ability to respond to both RC and II tasks, including background surveys, observations and stimulated recalls. The data were then reviewed by ASL subject matter experts to both determine the instrument's effectiveness and recommend improvements. The second phase (N=30) expanded the ASL user pool to L1, CODAs and world language learners to determine the instrument's effectiveness in assessing RC as well as eliciting language for the II phase. In addition to capturing learner responses, we collected background data, conducted real-time test administration observations and administered stimulated recalls (in English, ASL or both, depending on the participants' language background) to explore the learner experience and determine ways to improve overall and for specific groups. The final phase will address ways to develop reliable rating procedures for learner responses. The resulting product will allow ASL users and learners to demonstrate their language proficiency for the SOBL via a reliable, computer mediated instrument at a low cost to the learner.

Paper and Demo Abstracts

Investigating the relationship between automated measurements of nonverbal behavior and language proficiency ratings

Dylan Burton

Saturday September 9: 10:20 am - 10:50 am Location: Room 210

Raters are attuned to nonverbal behavior when measuring second language proficiency (Ducasse & Brown, 2009; May, 2011; Sato & McNamara, 2019). When test takers behave differently during speaking tests, nonverbal behavior can become a salient source of information about communicative language ability (Jenkins & Parra, 2003; Neu, 1993). To date, however, there is little empirical information of the size or consistency of the effect of nonverbal behavior or whether language proficiency may moderate the impact of behavior on proficiency ratings.

In this study, one hundred novice raters watched and scored 30 recordings of test takers taking an international, high stakes proficiency test. The speech samples were each two-minutes long and ranged in proficiency levels. The raters scored each sample on fluency, vocabulary, grammar, and comprehensibility using seven-point semantic differential scales. Nonverbal behavior was extracted using an automated machine learning software called iMotions, which produced three measures: overall facial muscle activation, behavioral valence, and attention. Two sets of ordinal mixed effects regression models were computed for each proficiency outcome, one using mean values of behavior across the samples, and the other using variance in behavior using standard deviations. Interaction terms with proficiency were included in each model.

The results showed that attentional variance predicted fluency, vocabulary, and grammar, but only when accounting for proficiency. Higher standard deviations of attention corresponded with lower scores for the lower proficiency group, but not the higher proficiency group. Comprehensibility was only predicted by mean valence when proficiency was an interaction term. Higher mean valence, or positive emotional behavior, corresponded with higher scores in the lower proficiency group, but not the higher proficiency group. Nonetheless, effect sizes for these predictors were quite small, with small amounts of variance explained. These results have implications for understanding how raters interpret language proficiency in light of visual information.

Paper and Demo Abstracts

Examining the direct and indirect impacts of verbatim source use on linguistic complexity in integrated argumentative writing assessment

Huiying Cai

Saturday September 9: 10:20 am - 10:50 am Location: Room 304

Integrated argumentative writing tasks have been widely used to assess L2 writing proficiency for their authentic nature, where test-takers can use source texts for both content and language support. When used verbatim, source texts may directly enhance the linguistic complexity of writing performance. This enhanced linguistic complexity might present an unequal advantage for test-takers across levels of writing proficiency, engendering validity and fairness concerns. While previous research has mostly examined the relationships between source use characteristics and proficiency levels, the relationship between verbatim source use (VSU) and linguistic complexity remains underexplored. To further unpack these relationships, this study examined (1) whether and the extent to which VSU directly impacts linguistic complexity after accounting for writing proficiency and (2) whether and the extent to which VSU also indirectly impacts linguistic complexity through interaction with writing proficiency. The study included 3,250 essays on an integrated argumentative writing task of a US university's English Placement Test (EPT). Using natural language processing tools and techniques, we examined 34 linguistic complexity features at both syntactic and lexical levels and three VSU features in terms of quantity and length. We performed exploratory factor analysis to identify dimensions of linguistic complexity and applied linear mixed-effect models to examine how VSU features and proficiency level impacted the linguistic complexity dimensions. Post-hoc analyses suggested weak but different direct impacts of VSU features on linguistic complexity but no indirect impact. Specifically, quantity of VSU tended to have positive impacts on linguistic complexity, while length of VSU tended to have negative impacts. These different direct impacts might reflect different essay writing strategies associated with VSU. The findings help unravel the impact of VSU in the integrated argumentative writing assessment and provide empirical evidence for validity arguments for the EPT writing task.

Paper and Demo Abstracts

Using international English proficiency tests in EMI contexts: Exploring the relationships between student performance on TOEFL iBT and EMI course assignments in Korean universities

Soohye Yeom

Saturday September 9: 10:20 am - 10:50 am Location: Room 108

With the increased popularity of English-medium instruction (EMI) in higher education, many East Asian universities are using international English proficiency tests to make admissions and placement decisions. Since these tests were not originally designed for the EMI contexts, validity evidence is needed to support the use of these tests in this new context (Bachman & Palmer, 2010). This study focused on the extrapolation inference in an interpretive validity argument (Chapelle et al., 2008), by examining whether students' scores on a test can be interpreted as indicators of students' performance in university EMI courses. This study investigated (1) the relationships between students' TOEFL writing scores and their writing performance in EMI courses, and (2) the comparability of the quality of the students' writing on the TOEFL writing tasks and in course assignments. Participants included 111 undergraduate and graduate students at Korean universities enrolled in EMI content courses in various disciplines who took the TOEFL writing test and submitted their course assignments. Data analyses included (1) regression analyses on the relationships between the writing test scores and course indicators of writing performance across disciplines, and (2) repeated-measure two-way ANOVA examining whether there are differences in the quality of student writing between task types (the TOEFL Integrated/Independent tasks vs. EMI course assignments) across disciplines. The findings showed that only the Integrated task score was a moderate and significant predictor of EMI writing performance, and the quality of the writing in the Integrated task was more comparable to that of course assignments than the Independent task, providing insights into language test development. Also, varying relationships were found between the students' performance on the TOEFL tasks and in course assignments across disciplines, suggesting that the students' performance on the test could be extrapolated to disciplinary performance to different extents depending on their academic area.

Paper and Demo Abstracts

What does elicited imitation measure? Findings from strategy use on an elicited imitation test of Chinese

Yuyun Lei

Saturday September 9: 10:20 am - 10:50 am Location: Room 208

Elicited imitation (EI) has gained popularity in recent years with the field's quest for efficient and flexible measures of second language (L2) proficiency. Despite the surge of interests in EI, questions remain as to what specific linguistic knowledge, skills, and strategies EI measures. Test-takers' strategy use form an important source of validity evidence for understanding the construct of a task (AERA, APA, and NCME, 2014). Therefore, as an effort to elucidate the constructs of EI, this study explored the nature of test-takers' strategy use on EI and its effect on EI performance.

Twenty-four L2 learners and eight native speakers of Chinese were recruited to complete an EI test of Chinese and a strategy use questionnaire after the test. The EI test consisted of 72 sentences featuring varying lexico-grammatical complexity levels. The strategy use questionnaire gathered information about participants' processes and strategies employed to complete the EI test. Participants' questionnaire responses were qualitatively analyzed to identify strategies employed for the EI test. Descriptive statistics were calculated to examine possible trends of strategy use across proficiency levels. Linear mixed-effects regression analyses were conducted to investigate the effects of the use of different strategies on EI performances.

Results revealed that participants mainly employed five types of strategies, including approach strategies, cognitive strategies, metacognitive strategies, communication strategies, and test-wiseness strategies. Except for test-wiseness strategies, other strategies are common to the strategies test-takers would use for listening and speaking tasks. Moreover, the linear mixed-effects regression analyses showed that participant proficiency level, test item length, and item complexity level explained the largest test score variance, while the use of different strategies only accounted for a small proportion, among which test-wiseness strategy did not contribute positively to EI performances. These findings provide validity evidence for EI and suggest that EI taps into speech processing and production.

Paper and Demo Abstracts

The influence of Grammarly and ChatGPT on rater behavior and the halo effect

Jim Ranalli, Shireen Baghestani

Saturday September 9: 10:55 am - 11:25 am *Location: Room 210*

Although the halo effect, a tendency of raters to assign similar ratings across multiple criteria, has been studied extensively, little research has focused on its potential alteration with the use of language enhancement tools or its influence on individual analytic criteria when certain elements of the writing are systematically controlled. This study aimed to address these gaps by examining the impact of Grammarly and ChatGPT, two such language enhancement tools, on trained raters' behavior. The study design presented raters with three versions of essays: an original version, a version improved with Grammarly, and a version enhanced with ChatGPT. Grammarly improvements operationalized controlled enhancements in 'Grammatical Range and Accuracy' and 'Lexical Resource', whereas ChatGPT also improved 'Coherence and Cohesion' according to the IELTS rating scale. Six trained raters rated these essays using the IELTS analytic scale, which includes 'Task Response' as an additional criterion. To avoid potential order effects, the sequence of essay versions was randomized and with one-week intervals maintained between rating sessions for the different versions of the same essay. The raters were blind to the fact that they were rating different versions of the same essays, and a many-facet Rasch model was used to account for rater severity. The study aimed to uncover potential correlations between the ratings across different categories as evidence of the halo effect, especially in terms of Coherence and Cohesion, which was only manipulated by ChatGPT, and Task Response, which was not manipulated by either language enhancement tool. With data collection still in progress at the time of abstract submission, the results were pending. This investigation thus sought to expand understanding of the halo effect, especially in the context of the introduction of language enhancement tools and their potential influence on rater behavior.

Paper and Demo Abstracts

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

Sun-Young Shin, Senyung Lee

Saturday September 9: 10:55 am - 11:25 am *Location: Room 304*

Numerous studies have compared the effect of input mode (video-based vs. audio-only) on adult second language (L2) learners' listening comprehension. However, there has been no investigation into how different types of audio-visual input in a listening test affect young learners' listening comprehension. To address this gap, this study examined the effects of audio-visual input type and text listenability on the listening comprehension performance of 128 Grade 3-5 students in the U.S., including 50 English language learners (ELLs) and 78 non-ELLs. Participants took a computer-delivered English listening test comprising three test forms with six passages each, including three video passages and three audio-only passages. The videos were distributed counter-balancing two variables: the listenability of the text, as measured by Lexile Audio Measure, and the type of visuals, including speaker-only, visual-only, and speaker-and-visuals. The study found that students performed equally well on the video-based and audio-only passages, and the type of visuals did not significantly affect students' listening comprehension scores, regardless of their ELL status. However, the listenability of the text significantly affected the students' listening scores. Furthermore, a post-test survey revealed that students had a positive perception of all three types of audio-visual input.

Uncovering the dimensionality of authorial voice in the integrated writing task: A Confirmatory Factor Analysis study

Lia Plakans, Yafei Ye, Kwangmin Lee

Saturday September 9: 10:55 am - 11:25 am *Location: Room 108*

Research has long noted the importance of establishing an authorial voice as a hallmark of academic writing. However, despite purported connections between academic writing and integrated writing tasks (Gebril & Plakans, 2014; Huang & Hung, 2013), little attention

has been paid to examine the role and effect of authorial voice on writing quality in a test task that involves use of source texts. The primary intent of this study is to shed light on the internal structure of authorial voice in an integrated reading-to-write task by way of Confirmatory Factor Analysis (CFA). The participants of this study included 315 Chinese college-level EFL students who wrote for a task addressing the topic of "eating for professional situations".

In coding the data collected, we adopted Hyland's (2005) model of authorial voice whereby the construct is categorized into stance and engagement. Stance, referring to "writer-oriented features of interaction" (p. 178), is undergirded by hedges, boosters, attitude markers, and self-mentions. On the other hand, engagement consists of reader pronouns, personal asides, appeals to shared knowledge, directives, and questions. To analyze the data collected, we first rated the collected essays based on each of the stance and engagement traits on a 1-5 scale. We took a model comparison approach to CFA. The results suggested that the current data are best characterized as a unidimensional construct; it represents one single factor. Furthermore, the Structural Equation Modeling (SEM) analysis indicated verbatim source use is inversely associated with quality of authorial voice, corroborating our hypothesis that improper source copying negatively influences the rater's perception of a test-taker's authorial voice. Practical implications regarding rubric use and instruction are discussed.

Typing vs. hand-writing on L2 writing tasks: Examining test-taker outcomes and perceptions

Matt Coss, Paula Winke

Saturday September 9: 10:55 am - 11:25 am *Location: Room 208*

The extent to which writing modality (i.e., hand-writing vs. keyboarding) impacts second-language (L2) writing assessment scores remains unclear. Research has demonstrated that for alphabetic languages like English, writing test scores are equivalent when test takers type or hand-write (e.g., Barkaoui & Knouzi, 2018). However, for non-alphabetic languages like Chinese, recent research has shown that scores differ significantly, with typed responses consistently out-scoring hand-written responses in

beginner (e.g., Zhu et al., 2016), pre-intermediate, (e.g., Zhang & Min, 2019), and advanced proficiency groups (e.g., Bourgerie et al., 2023).

Our goal was to examine whether the differential scoring on L2 Chinese writing tests is detectable at the intermediate level. As such, we collected two writing assessment samples from 64 intermediate-level university learners of L2 Chinese (N=128), with prompt and modality both counter-balanced. We also collected test-takers' reports of their writing process and their perceptions of the writing tasks. We invited two certified raters of Chinese-language writing proficiency tests to rate all writing samples using a three-criterion analytic rubric (complexity, accuracy, and lexicogrammatical diversity) adapted from a widely-used world language proficiency test in the U.S. (AVANT, n.d.).

A linear mixed-effects model revealed that regardless of modality or prompt order, the 128 test takers wrote significantly more and received significantly higher scores on typed responses, even when language proficiency, multilingual experiences, and Chinese writing and typing habits were taken into account in the model. Together with the prior documentation of substantial differences between writing processes and perceptions when hand-writing compared to typing, these findings present evidence that hand-written and typed L2 Chinese test tasks may tap into different constructs, and therefore have implications for L2 writing assessment-score use (e.g., classroom-based assessment, placement testing, proficiency tests) as well as assessment research.

Translanguaging in assessment: A scoping review

Chia-Hsin Yin

Saturday September 9: 11:30 am - 12:00 pm Location: Room 210

This scoping review focuses on assessing bi/multilingual learners with pedagogical translanguaging (TL) framework (García , 2007; García & Li Wei, 2014). In a TL classroom, students are allowed to flexibly deploy their entire linguistic repertoires (García, 2016). Particularly, TL facilitates literacy development (e.g., García & Kleifgen, 2019) and ideological shifts, impacting school language policies (Menken & Sánchez, 2019). However, little focus is on what and how pedagogical TL is utilized as an assessment. As such, this

scoping review provides a systematic mapping (Brant & Booth, 2009; Peters et al., 2015) of the status quo of how pedagogical TL is deployed in K-16 classroom assessment to reimagine the traditional assessment by attenuating the translingual-monolingual binary (Lee, 2016) and to illustrate the multilingual realties in language use (Baker & Hope, 2019). The research questions are, in K-16 contexts, (1) what are the characteristics of pedagogical TL assessment and (2) what are the washback effects of pedagogical TL assessment. Three rounds of rigorous search in research databases were conducted using the keywords: (translang') AND (assess' OR evaluat' OR measur' OR test' OR screen'). 15 peer-reviewed journal articles were selected and analyzed.

Findings suggest that TL assessment has been employed in both ESL and EFL multilingual and multicultural contexts among elementary children, college students (Mbirimi-Hungwe, 2021), in-service bi/multilingual teachers (Schissel et al., 2019, 2021), bilingual professors (Baker & Hope, 2019), and EFL school administration system (Gandara & Randall, 2019) with qualitative research paradigm as the main methodology. Predominant research indicates positive washback effects of TL assessment, such as positive perception (Ascenzi-Moreno, 2018), identity affirmation (Shi & Rolstad, 2022), and broader understanding of socio-cultural themes (Wang, 2022). Yet, challenges lie in the tension of integrating TL assessment in practice and validating students' dynamic ways of languaging. Pedagogical implications and future research directions will be discussed.

Automating silent pause marking and computing fluency measures: A demonstration using ELAN and CLAN (Technology demonstration)

Qiusi Zhang

Saturday September 9: 11:30 am - 12:00 pm Location: Room 304

In this presentation, I will demonstrate how to use two publicly available, free software programs, ELAN (Version 6.5) (MPI, 2023) and CLAN (MacWhinney, 2000), for automatically marking silent pauses in second language speech and extracting temporal fine-grained fluency measures that involve silent pauses. These measures include speech fluency measures (mean length of run and articulation rate) and breakdown fluency

measures (silent pause rate and silent pause duration) (Skehan, 2003). Fine-grained measures have been argued to better capture the multi-dimensional nature of second language proficiency (Larsen-Freeman,
2006; Skehan, 1998; Wolfe-Quintero et al., 1998). However, the traditional approach of manually marking silent pauses has been labor insensitive, and challenges associated with human coding and software knowledge have contributed to the lack of research in L2 speech compared to L2 writing (Yu and Lowie, 2020).

Through this demonstration, I aim to demystify the use of speech analysis software for language testing purposes. ELAN serves as an annotation tool for audio and video recordings. CLAN, short for Computerized Language Analysis, is a powerful tool enabling audio and video transcription, as well as the automatic computation of a wide range of fluency, complexity, and accuracy indices. Specifically, I will showcase the following steps 1) using the "silence recognizer MPI-PL" in ELAN to automatically mark silent pauses; 2) manually checking and revising the identified pauses; 3) employing CLAN to extract the number and duration of silent pauses; 4) calculating the four fluency measures. The implementation of this technology is time-effective and can encourage more researchers to conduct studies related to L2 speech and further enrich our understanding of second language speech learning and teaching.

Linking authentic materials and assessments with Korean language learners' outcomes

Weejeong Jeong

Saturday September 9: 11:30 am - 12:00 pm Location: Room 108

This study investigates how authentic teaching materials are linked to authentic speaking assessment and learners' outcomes in second-year Korean language courses at a mid-western university. Authentic materials have been found to be beneficial for language learning (Pinner, 2013; Zyzik and Polio, 2017). Similarly, authentic assessment, which mostly refers to real-world tasks or problems (Wooster, 1993), has been found to have a positive impact on learning and teaching. This study aims to answer the following research questions: (1) What types of authentic materials have been used in the classroom? (2) How

is the use of authentic materials linked to authentic assessment? (3) To what extent is authentic assessment related to the outcomes of Korean language learners? The data consists of authentic materials, such as video clips, photos, cultural artifacts, handouts, online materials, and student grades during the 2021-2022 academic year and the 2022 fall semester. This study employed content analysis, descriptive statistics, and a series of bivariate correlation analyses. The results showed that using a thematic format for lesson plans made it possible to incorporate authentic materials into the classroom and connect them to authentic oral assessments. The statistical analyses revealed that students' performance on the authentic oral assessment was significantly and positively correlated with their performance on the written tests of receptive language skills (r = .75). It was also found to be moderately and positively related to other grading indicators, including workbook homework (r = .48), vocabulary quizzes (r = .56), attendance and participation (r = .61), and the video project (r = .40). The results demonstrate that incorporating authentic materials into language instruction and assessment can lead to learners' positive outcomes. This suggests that authentic assessment can be an effective way to measure learners' Korean language proficiency and it can complement traditional forms of assessment.

Co-occurrence of disfluency features of L2 speech across proficiency levels in controlled and spontaneous tasks

Yulin Pan

Saturday September 9: 11:30 am - 12:00 pm *Location: Room 208*

As a core construct in measuring L2 speaking proficiency, fluency is reflected by temporal features of speech in speed and smoothness (Lennon, 1990). Depicting the smoothness of speech, disfluency features, including breakdown and repair features, often occur consecutively at one location or across multiple locations in one utterance (Riggenbach, 1991; Shriberg, 1994). While previous literature tends to investigate disfluency features in isolation using regression models (e.g., De Jong et al., 2013; Révész et al., 2016), the current study explores the co-occurrence of fine-grained disfluency features. A speech corpus of responses from 58 L1 Chinese speakers of English and 13 native English speakers on the

Elicited Imitation Task (EIT) and oral Listen-to-Summarize Task (LST) sections of a post-admissions English placement test is used. Each participant was analyzed based on 21 disfluency features using hybrid hierarchical-based k-means clustering to extract disfluency profiles. The association between different categories of disfluencies and proficiency levels was then examined using contingency tables. Results showed that on both tasks, speech involving more lengthy repairs shows a negative association with proficiency, and speech involving fewer hesitations and repair shows a positive association. However, task-specific disfluency profiles were also observed. In the EIT responses, different profiles were observed featuring co-occurrence of both repair and breakdown, indicating a stronger cognitive demand of the task and a recall-like behavior among all proficiency levels. On the LST task, disfluency profiles tend to feature repair and breakdown features separately. Implications for both L2 speaking and language testing research will be discussed.

The impact of source use features on raters' judgment of argumentation

Ping-Lin Chuang

Saturday September 9: 1:55 pm - 2:25 pm *Location: Room 210*

The scoring of L2 integrated writing tasks is a complex process that involves decisions in content, linguistic, and source use features (e.g., Cumming et al., 2002; Weigle & Montee, 2012). Although researchers (Gebril & Plakans, 2014) have found that raters consider source use characteristics including citation mechanics and source use quality, it remains unclear if and how raters perceive these features when determining the quality of argumentation in a source-based argumentative writing task. This study aims to explore how source use features impact raters' judgment of argumentation performance.

106 participants were recruited to complete a rating task. The participants were categorized into novice and experienced raters according to their experience in teaching and assessing L2 writing. The rating task simulated the scoring assignment of a local English Placement Test (EPT). Written responses in the rating task were paragraphs adapted from essays written by EPT test-takers. Source use features—source use quantity and quality—were manipulated in a 2x2 factorial design (high vs low). Source use quantity was operationalized as the number of source text ideas used while quality considered the

relevance and effectiveness of source use. Rater scores were analyzed using the many-facet Rasch model where argumentation scores were compared across four source use conditions and two rater groups.

Results revealed that raters assigned the highest argumentation scores to paragraphs with high source use quantity and quality, followed by high-quantity-low-quality, low-quantity-high-quality, and low-quantity-low-quality paragraphs. While raters showed similar levels of severity on most types of responses, they differed substantially in the rating of high-quantity-high-quality paragraphs. Novice raters were significantly more severe on paragraphs with high source use quantity and quality than experienced raters. They also assigned relatively similar ratings to high-quantity-low-quality paragraphs and low-quantity-high-quality paragraphs compared to their experienced counterparts. Implications for rater training and assessment research will be discussed.

Notetaking as validity evidence: A mixed-methods investigation of question preview in listening assessment

Rebecca Yeager, GoMee Park, Ray J.-T. Liao

Saturday September 9: 1:55 pm - 2:25 pm *Location: Room 304*

Recent scholarship has questioned the cognitive validity of listening tests with preview, in which test-takers can see test questions before listening (Koyama et al., 2016; O'Grady, 2021). Students report using more testwise strategies, such as guessing and keyword matching, in tests with preview (Field, 2011; 2012; Yeom, 2016), while evidence from eye-tracking and Functional Near-Infrared Spectroscopy (fNIRS) indicates patterns consistent with shallow listening in while-listening performance tests with preview (Aryadoust et al., 2022). Notetaking is less subjective than self-report data and more interpretable than eye-tracking and fNIRS, making it a useful window into test-taker listening processes. Further, it is an essential skill for success in academic contexts (Asaly-Zetowi & Lipka, 2019), underscoring the need for positive washback of listening assessment on notetaking behaviors. This study applies a convergent mixed-methods design to explore student notes for evidence of cognitive processes in multiple-choice listening tests with and without preview. Qualitative analysis of notes (n = 94) indicated

that test-takers in the preview condition were more likely to systematically omit information, highlight keywords, and engage in shallower representation of propositional structure, and were more likely to attempt to apply those strategies when listening without preview later. Conversely, Kruskal-Wallis tests revealed that test-takers who listened without preview took more notes, especially of main ideas and details, and had better coverage of the lecture. However, correlation and hierarchical linear regression analyses found these notetaking achievements did not predict higher scores in the no-preview condition, while in the preview condition, only note quantity and focus on details predicted scores. Both strands of data suggest that students' cognitive processes were shaped by the format of the exam they experienced. These findings may bear on validity arguments for listening assessment as well as inform the way that language instructors prepare their students for academic listening.

Using ChatGPT for writing evaluation: Examining the accuracy and reliability of AI-generated scores in comparison to human raters and potential reasons for discrepant ratings

Shuhui Yin, Haeun Kim, Yasin Karatay, Shireen Baghestani, Jeanne Beck, Leyla Karatay, Sebnem Kurt, Mutleb Alnafisah

Saturday September 9: 1:55 pm - 2:25 pm *Location: Room 208*

In language assessment, ChatGPT (Open AI, 2023) has been found to be useful in evaluating student writing and providing formative feedback. However, questions regarding the usefulness of ChatGPT in generating "scores" in assessing second language writing are yet to be answered. In addition, while AI prompt engineering has emerged as a new important field of research (Teubner et al., 2023), there is a limited amount of research showing the extent to which ChatGPT-generated scores can vary depending on the prompt that the user provides. Therefore, the current study aims to investigate the accuracy and reliability of the scores that ChatGPT produces in comparison to human raters; to understand the factors that result in discrepant ratings between humans raters and ChatGPT; and to find optimal ways of prompting ChatGPT to generate scores based on a rubric for an integrated writing task.

First, a random sample of 90 argumentative essays was taken from Iowa State University's English Placement Test (EPT) Corpus of Learner Writing (2017). Each essay was independently rated by three raters, and the 70 essays which received the most consistent ratings based on a many-facet Rasch model were selected for subsequent use. Ten were provided to ChatGPT as benchmark responses (i.e., two essays per each of the five levels on the scale) along with the EPT writing scale descriptors. ChatGPT was asked to rate the remaining 60 essays, and the accuracy and reliability of its ratings were compared with the human ratings in various prompting conditions. Finally, for the responses that led to discrepant ratings, we asked ChatGPT to provide justifications for the scores and analyzed its responses to see if there were any recurrent patterns/themes. In this presentation, we will present our comparison results and discuss the impact of prompts on the rating performance of ChatGPT.

The effect of adding an interactive avatar to SDS-mediated oral discussion tasks of an oral communication test: A test taker perception study

Reza Neiriz

Saturday September 9: 2:30 pm - 3:00 pm *Location: Room 210*

Spoken dialogue systems (SDS) have been studied in the context of language learning and testing due to their promising practicality and consistency in delivery. These systems can create the impression of a dialogic oral interaction by presenting plausible responses to the user input. Because of this quality, they can be used as a partner in paired oral discussions in tests of oral communication (Ockey & Chukharev-Hudilainen, 2020) or role-play tasks (Gokturk, 2020; Karatay, 2021). However, SDSs used in testing contexts mainly use a still image related to the task (Ockey & Chukharev-Hudilainen, 2020; Gokturk, 2020; Karatay, 2021; Timpe-Laughlin et al., 2017). This leaves test takers with only verbal and intonational clues about turn management and misses other important metalinguistic information that can help manage oral interaction. This makes SDSs less comparable with human-to-human oral interaction and unable to tap fully into different aspects of oral communication, such as body language. To study the effect of adding a dynamic avatar to a web-based SDS in the context of language testing, an SDS was designed with ten discussion tasks related to student life topics, such as whether it is

better to live in a dorm or an apartment. A 3D avatar was also designed in Blender 3.4.1 (Community, 2018) and was exported as a GLB file to be handled by WebGL through ThreeJS. The mouth movements reflecting speech were created in real-time using the viseme information from the text-to-speech module of the SDS. Additional body language features such as hand gestures, head movement, and eye movement were programmed into the system. Half of the ten conversations were randomly selected to be delivered with the avatar and the other with a still image for each participant. The order of the conversations was also randomized for each participant.

Eighty-four L2 speakers of English (Female = 45, Male = 39, Female age mean = 27.82, SD = 7.42, Male age mean = 28.28, SD = 6.74, undergraduate = 19, Master = 18, Ph.D. = 47) at a large midwestern university participated in the study. After each conversation, the participants received a short survey, one question of which specifically asked how helpful the avatar or the still image was. The participants answered this question on a five-point Likert scale, with five being very helpful and one being the least helpful. The means of responses to this question from the two with and without avatar conditions were calculated for each participant. Then a paired-sample t-test was conducted to compare these means. The results indicated that students found the avatar condition generally more helpful (t=2.817,df=83,p=0.006, D=0.307). Test takers were also interviewed to learn more about the reasoning. The main reasons for preferring avatars were visual backchannelling, human-like interaction, and a non-judgmental impression of the avatar. Those who did not prefer the avatar indicated that they found the avatar weird due to its unnatural body and facial expressions. Some others found it strange because they reported that they did not expect a high level of responsiveness from an avatar. For those who preferred still images, the main reason was the contextual clues that the still images provided. Overall, most of the participants perceived the dynamic avatar positively, but further improvements are required to make the avatar more natural. Moreover, the lack of important information, such as backchannelling through body language in SDSs, was significant for the participants, and the results suggest that a dynamic avatar can address this issue.

Paper and Demo Abstracts

Using text features to predict rater disagreements

Shireen Baghestani

Saturday September 9: 2:30 pm - 3:00 pm *Location: Room 304*

In performance-based assessment, inter-rater reliability is critical to the valid interpretation of test scores (Knoch & Chapelle, 2018). While there is a large body of research showing that attributes of the raters may cause score disagreements (Barkaoui, 2010), less research focuses on disagreements caused by responses that are "difficult to score" (Wolfe et al., 2016). Therefore, this study aims to understand what makes a written response difficult to score in terms of its textual characteristics (content, organization, grammar/vocabulary, and source use). The context for the study is an English placement test of writing (EPT-W) at a U.S. university. The study will focus on the first task of the EPT-W: a summary of two reading passages.

In this study, 55 EPT-W summary tasks will be coded for coherence, summarization, accuracy, citations, grammatical complexity, lexical richness, length, spelling, and textual borrowing. The first four variables will be manually coded using rubrics and the last five variables will be coded using computer-based tools. Each summary will be assigned a holistic rating by 6 raters who have previous experience rating for the EPT-W and have undergone a 1-hour norming session. The agreement of their scores will be calculated using mean absolute deviation (MAD; Heumann et al., 2023). Separate simple linear regression models will be run for each text feature, with MAD as the dependent variable. Qualitative data will be collected in the form of a stimulated recall task and an interview.

I will present the outcomes of the analyses in terms of 1) which text features, if any, predicted MAD (difficulty of agreement), and 2) a preliminary summary of raters' qualitative responses indicating why it was difficult to agree on a score for particular texts. The findings will have implications for rater training by shedding light on features raters may have difficulty agreeing on.

Paper and Demo Abstracts

Exploring the effect of passage difficulty level on ChatGPT's test item generation

Jean Young Chun, Natalia Barley

Saturday September 9: 2:30 pm - 3:00 pm *Location: Room 208*

The recent development of ChatGPT shows great promise in its ability to create test items, expediting the test development process. However, it remains unclear whether ChatGPT can maintain item quality across different levels of passages, particularly those at higher levels that involve personal views or abstract ideas. This study explores the variation in quality of ChatGPT-generated items across different levels of listening passages and investigates the factors contributing to such differences. Two sets of Korean listening passages were utilized, categorized by difficulty levels according to the Interagency Language Roundtable (ILR) scale. The first set consisted of 10 L2 instructive-mode passages that feature factual information, while the second set contained 10 L3 evaluative-mode passages that are abstract, analytical, and inferential in nature. Researchers guided ChatGPT to generate level-appropriate multiple-choice questions. A total of 60 items (three per passage) were evaluated against five criteria: (a) level appropriateness, (b) question clarity, (c) presence of a single correct answer, (d) distinctiveness of distractors, and (e) plausibility of distractors (Haladyna, 2004). Four raters reviewed and rated each item on a five-point Likert scale, providing justifications for their ratings. Inter-rater reliability was examined, and The Mann-Whitney U test was performed to analyze differences in ratings between L2 and L3 items for each criterion. Furthermore, a qualitative analysis of rater comments, listening passage characteristics, and item characteristics was conducted to explain the changes in item quality between the two levels. The study's findings highlight the potential of ChatGPT for enhancing practicality in test item development and provide insights into the precautions necessary when utilizing ChatGPT for item writing.

Measuring complexity features in young learner spoken language: Human versus contextually applied automated coding

Haeun Kim, Gordon Blaine West

Saturday September 9: 3:30 pm - 4:00 pm *Location: Room 210*

Analyzing complexity features is useful in understanding how test takers' language varies in response to different assessment tasks. Meaningful differences in features can help identify issues related to task effects and evaluate the validity of score interpretations (Cumming et al., 2006; Plakans et al., 2019). Advances in natural language processing technology have provided researchers with several tools to automatically code complexity features; however, existing automated tools are often used as universal tools and are not adaptable to specific contexts. Since language is shaped by the context in which it is produced, contextualization is critical to the accuracy of coding in assessment research. This paper focuses on the development of a contextualized automated coding tool to analyze complexity features in young learner speech. Due to the intricacies of young learner language produced in response to computer-based speaking assessment, we decided to develop an automated coding tool that was adapted to the specificities of our dataset. The data consists of test responses from the speaking component of ACCESS for ELLs Online, a large-scale computer-mediated English language proficiency assessment used in US K-12 schools. Multilingual learners (n = 360) from five grade levels responded to six different speaking tasks. The transcribed data was coded for verb tenses, modal verbs, verb phrases, and cohesive devices by trained human coders and an automated coding program written with Python. Using criterion values, the accuracy of both codings were compared. Automated coding demonstrated higher levels of accuracy and consistency compared to human coders across all four measures and grade levels. The results show how automated coding can become a powerful tool when adjusted for the context of the data being analyzed. Developing contextually adjusted automated tools is a necessary next step in measuring the many varieties and types of language produced in our increasingly interconnected and multilingual societies.

Paper and Demo Abstracts

The construct of summary writing: how do reading, writing, text features, cognitive operation difficulty impact summarization as an integrated task

Yu-Chen Tina Lin

Saturday September 9: 3:30 pm - 4:00 pm *Location: Room 304*

This study examined a hypothesized structural equation model (SEM) to understand the performance of 358 EFL Taiwanese university students in two summary writing tasks (A & B) with different text features. The aim was to investigate the influence of L2 general reading, writing, familiarity with text features, and text reading comprehension on the perception of difficulty with cognitive operations and summary writing performance.

716 summaries were graded by two raters using a four-criterion analytic rating rubric (score: 0-4, reliability: 0.8-0.87). After writing each summary, students completed two tasks: (1) a Likert scale questionnaire on familiarity with text features (reliability: .82-.88) and perception of difficulty with cognitive operations (reliability: .87-.97), and (2) a multiple-choice reading comprehension test of the text being summarized. A few weeks before writing summaries, students completed TOEFL-iBT reading and TOEFL independent writing. 358 TOEFL essays were graded by the e-rater Criterion and two raters (reliability = 0.84).

In both text A and text B SEM models, text reading comprehension, L2 general writing proficiency, and perceived cognitive operation difficulty directly and significantly affected summary writing quality. Although familiarity with text features and L2 general reading proficiency did not directly impact summary writing, general reading proficiency indirectly contributed to summary writing through text reading comprehension. Familiarity with text features only indirectly influenced summary writing through cognitive operation difficulty and had an impact on reading comprehension only for the more difficult text B.

This theoretical framework has the potential to systematically research integrated writing assessment by explaining a significant amount of variance in summary writing (A: 77%; B: 59%). It also provides insights into the contradictory findings of previous studies (Yu, 2005; Li, 2014) on the relationship between reading, writing, and summary writing. The implications of these findings are discussed in relation to assessment theory, test design, and pedagogy.

Paper and Demo Abstracts

Pre-post elicited imitation: Documenting proficiency development in a two-semester EAP program

Lixia Cheng, Xiaorui Li

Saturday September 9: 3:30 pm - 4:00 pm *Location: Room 208*

Elicited Imitation (EI), a psycholinguistic task eliciting oral production of language chunks, is believed to measure L2 linguistic competence, through tapping into language learners' implicit grammar and language processing capabilities (e.g., Davis & Norris, 2021; Van Moere, 2012). Our study examined a locally developed and rated, internet-based EI test used in an English for Academic Purposes (EAP) program at a large public U.S. university to document international freshmen's English proficiency development and provide empirical evidence for program evaluation. We conducted total population sampling on the FA 2017–SP 2018 cohort of students enrolled in the two-semester EAP course sequence. A total of 261 students' EI item scores and section-total score were analyzed in a three-stage quantitative study:

1. Classical Test Theory (CTT) analyses of measurement properties (e.g., interrater reliability, test form comparability, item difficulty, item discrimination, and test reliability)
2. Pre-post comparisons of section-total scores and a subsequent calculation of the effect size for score changes
3. Computation of an Instructional Sensitivity index (Hedge's g) per item

This EI test demonstrates excellent measurement properties; in addition, the statistical testing in Stage 2 indicates that the 261 first-year students made significant gains, with a large effect size, in their EI post-test as compared to the pretest seven months earlier.

While the 12 EI items in each test form were sensitive to student improvement as a function of the ESL support received, only 17 (35.4%) of the total 48 EI items were found to be instructionally sensitive. We then discussed why the other EI items had a less than desirable Hedge's g and which linguistic features contribute to high instructional sensitivity. This detailed item-level information (e.g., difficulty, discrimination, and instructional sensitivity) has practical value in informing test revision, especially for better capturing learner development and reflecting the integration of assessment and instruction.

Work-in-Progress Abstracts

Critical literacy and self-assessment: A study of Spanish students' literacy on sociolinguistic issues

Lorena Alarcon

Saturday September 9: 1:00 pm - 2:00 pm *Location: Room 300*

Previous studies revealed that Spanish students increased their critical language awareness (CLA) after receiving instruction (Beaudrie et al., 2019; Beaudrie et al., 2020; Quan, 2020; Gasca Jiménez & Adrada-Rafael, 2021). These studies mainly explored heritage speakers' CLA, which is an approach to reflect on practices related to language in society (Clark & Ivanič, 1997). However, little has been said about second language and heritage speakers' assessment of sociolinguistic issues and critical literacy defined as the development of knowledge of language issues in society (Del Valle, 2014). Besides their awareness, it is necessary to study second language learners' and heritage speakers' literacy because students interact in society and need to know about these topics (Holguín Mendoza & Taylor, 2021).

This study explores Spanish students' critical literacy and assessment of sociolinguistic issues. Participants included 19 students in the experimental group and 12 students in the control group. Those in the experimental group received instruction on sociolinguistic topics. All students completed a survey to self-assess their perceived sociolinguistic knowledge. Participants also completed an oral recorded interview, which allowed the researcher to understand students' perceptions of their self-assessed literacy. Preliminary findings suggest that those in the experimental group self-assessed their knowledge higher than the control group participants in some sociolinguistic topics, such as Spanglish, code-switching, language maintenance, etc. These findings showed CLA instruction helped these Spanish students to increase their self-assessed literacy on sociolinguistic issues. Preliminary results also revealed that heritage speakers in both groups displayed some knowledge of these issues because of their interaction with others in society.

Poster Abstracts

Virtual Administration of an Oral English Proficiency Test: Challenges and Student Perceptions

Sharareh Taghizadeh Vahed

Saturday September 9: 1:00 pm - 2:00 pm *Location: Room 300*

In American universities, assignment of international graduate students to teaching assistant positions has been a practice since the increase in the number of international students with the Open Doors policy after the 2000s. Many international students in the United States are dependent on graduate teaching and research appointments to fund their studies. When it comes to International Teaching Assistant (ITA) certification, speaking skills are usually the focus of assessment as the concern with ITAs' direct communication with undergraduates mostly involves the subskill of speaking. The assessment of ITAs' oral English skills is almost always done in person, whether the assessment method is direct (i.e., in the form of an oral proficiency interview), or semi-direct (i.e., delivered by a computer in a test center). In the context of school and test center shutdowns during the Covid-19 pandemic, language testers and university officials encountered difficulties in assessing ITAs' oral English proficiency for ITA certification purposes. Local language testing programs, with limited staff and limited budget, needed to quickly come up with a solution to meet their institutions' need to assess their ITAs in order to either certify them or place them in post-entry language support programs. This poster will demonstrate how a large public university in the Midwest was able to transfer their computer-delivered oral English proficiency test to an online test invigilation platform, what challenges were encountered, and what test-takers' perceptions were regarding online administration of a speaking test. The presenter will discuss the logistics of partnering with online invigilation providers, troubleshooting and fake testing procedures, and the local testing program's role in administering an oral English proficiency test virtually. The presenter will also discuss why their local testing program decided to continue administering their oral English proficiency test virtually post-COVID.

Poster Abstracts

The relationship between response time and performance on a large-scale writing assessment: Examining testing time for English learners

Junhee Park, Annette Vernon, Jeongmin Ji

Saturday September 9: 1:00 pm - 2:00 pm *Location: Room 300*

This study explores the association between response time and writing performance on an evidenced-based writing assessment as part of the state ELA assessment in Iowa. Grade 6 to 11 English learners and non-English leaners were compared based on how much time they have spent on completing the writing tasks and their overall writing score as well as scores on four writing traits, including Prompt Task, Development, Organization, and Language Use. For this writing task, students were asked to understand text- or data-based evidence and use the source into their writing. This state accountability test does not have any imposed time limits, however local testing administration may vary sight to sight. For the analysis, an ordinal logistic regression was used to model score probability across response time. The results showed that there was significant time differences in order for EL and non-EL students to maximize the probability of attaining scores across traits and grades. For example, EL students in grade 6 have a maximum probability of receiving a 2 (not-yet-proficient) out of 5 for prompt task at 28 minutes and a score of 3 (proficient) at 51 minutes compared to the 14 and 33 minutes for their non-EL pairs. Results from this study indicate that English Learners across grades do better with more time spent on writing. It is important for research such as this to be communicated to different stakeholders in order for these students to have the best possible chance at success.

A template for designing self-assessments for LCTL programs

Paula Winke, Laura Lopez

Saturday September 9: 1:00 pm - 2:00 pm *Location: Room 300*

Assessment in LCTLs can be notoriously difficult. In addition to having fewer standardized testing options than commonly taught languages, LCTL programs can run into difficulty

with standardized assessments not being aligned with their curriculum. The LoLA project seeks to address the need for locally-relevant assessments within LCTL programs by developing test-template kits for LCTL placement and diagnostic tests. The project provides LCTL instructors with the training and open access resources needed to develop and implement local assessments within their LCTL programs. Specifically, LoLA is developing a self-assessment template kit and a group orals template kit. Although specifically designed for LCTL teachers, LoLA's resources are broadly targeted at any language program in grades 9 through 16 that would need diagnostic or placement information on students entering and participating in their programs. This poster focuses on the presentation of the self-assessment template kit.

The self-assessment template kit guides language teachers to create ACTFL-proficiency-based self-assessments for their programs using locally customized Can-Do Statements. Attendees will learn how they can use the 10-item self-assessment to monitor learner progress, to obtain feedback on student and curricular success, and to facilitate student placement within their program. Attendees are provided with sample template materials and asked to give feedback to LoLA on how easy they are to understand and follow. Future assessment development plans will also be shared, and participants will be invited to engage in piloting those assessments and in providing feedback.

The role of educators in WIDA bias, sensitivity, and content reviews

Jason Kemp, Heather Elliott

Saturday September 9: 1:00 pm - 2:00 pm *Location: Room 300*

This poster will provide conference attendees with insight into the yearly Bias, Sensitivity, and Content Review of ACCESS test content that WIDA undertakes with educators. WIDA ACCESS for ELLs is an English language proficiency assessment taken by more than 2 million students annually in 41 U.S. states/territories. Test takers are K-12 students identified as English learners. Students' performance on ACCESS contributes to reclassification decisions.

Poster Abstracts

The WIDA Assessment Team and our partners at the Center for Applied Linguistics strive to create test content that is: reflective of classrooms in topics and difficulty; accessible to all students; representative of the multilingual student population. None of the work we do would be possible without the invaluable input from educators during the various stages of test development. Educators are involved in the development of ACCESS content from the item writing to the approval of operational test forms. This poster focuses on an important step in the test development process – the Bias, Sensitivity, and Content Review. WIDA relies on educators' experiences and perspectives to ensure that ACCESS test content is grade-level appropriate, free of bias, and free of sensitive topics.
During this poster presentation, we will share details about who participates in our reviews, including our recruitment process. We will cover how we conduct the review sessions and the checklists for review we provide to reviewers. Finally, we will address the impact of the review and how we update test content in response to educator feedback.

An investigation of longitudinal development of integrated writing among Chinese English as a Foreign Language (EFL) learners

Kwangmin Lee, Yafei Ye, Junhee Park, I-Chun (Vera) Hsiao

Saturday September 9: 1:00 pm - 2:00 pm *Location: Room 300*

An abundance of research evidence has been accumulated about the validity and reliability of integrated writing tasks in second language (L2) testing and assessment. However, most, if not all, of the evidence was collected cross-sectionally, leaving us with a limited understanding about how L2 learners develop longitudinally for a test task that requires integration between multiple language modalities.
The primary intent of this research study is to examine the longitudinal development of integrated writing among 20 Chinese college-level English as a Foreign Language (EFL) learners. Integrated assessment was administered three times within a semester, with a three-week interval between measurements. All the collected 60 essays were rated on a 1-5 scale based on the following seven traits: complexity, accuracy, source importance, source integration style, organization, development of ideas, and authorial voice.
We carried out multilevel modeling (MLM) to describe the patterns of change in each of the scoring traits we considered as a function of time. The way we conducted MLM was progressively complicated, proceeding from fitting intercepts and slopes based on simple

regression equations to those using more complex ones. We first examined an empty means, random intercept as a baseline. The result showed that complexity was the only measure where the addition of a fixed linear time slope was significant ($p < .001$), such that after a predicted value of 2.917 ($p < .001$) at Session 1, complexity scores increased by 0.250 ($p < .05$) per session on average. However, the addition of a variance of the random linear time slope was not statistically significant, if approximating significance at $p = .0563$, indicating non-significant individual differences in the linear rate of change. Implications for the L2 writing classroom and assessment are discussed.

Czech as a foreign language assessments in 2023

Kate Challis

Saturday September 9: 1:00 pm - 2:00 pm *Location: Room 300*

Czech has a situational context approximately the inverse of English; whereas around 92% of Czech speakers globally are native first language (L1) speakers, for English, approximately 0.2% are L1 speakers (Eberhard, 2020). Understandably, this means that most of the resources developed for both Czech as a Second Language (CSL) and Czech as a Foreign Language (CFL) are written by L1 Czech speakers who may not have extensive background in understanding how L1 English speakers specifically learn Czech. CFL/CSL teachers generally either adapt ESL or GSL (English and German as a Second Language, respectively) materials and techniques, often through a grammar-translation method (Škodová et al., 2013). A small number of corpus-based tools for Czech learning (Challis, 2022; Challis 2023), but it is not yet known the extent to which they are useful for language learning. In spite of it all, Czech is learned, and assessments exist to measure Czech proficiency. This study presents the current state of CFL assessments available to L1 English speakers located in the United States, which can generally be categorized into four groups: 1) tests administered by American universities, 2) tests administered by Czech universities, 3) online for-profit tests, and 4) tests administered by large language companies. This study compares all available tests from each category in content, length, cost, format, stakes, and grading scale, providing a case study of how to examine an assessment context that is currently understudied.

Poster Abstracts

Developing an assessment criteria framework for digital multimodal compositions

Stephanie Lissett Vigil, Chia-Hsin Yin

Saturday September 9: 1:00 pm - 2:00 pm *Location: Room 300*

This conceptual paper supports the use of digital multimodal composition (DMC) projects in K-12 EFL/ESL settings and proposes a product-based assessment rubric for digital stories (DSTs). Despite the intrinsic multimodality of communication (Kress, 2003), the assessment of English language learners' (ELLs) language proficiency skills has been predominantly decontextualized and monomodal (Liang & Lim, 2021). Support for using DSTs highlights the benefits of narratives as communicative vehicles for ELLs to share their experiences and relate with others (Kovach, 2018). Although previous empirical studies have attempted to apply different frameworks in assessing both the production processes and final products of the DMC practices of ELLs (e.g., Hung et al., 2022), the resulting assessment rubrics have focused on researcher purposes rather than in-practice teacher use. As such, this paper proposes a triangulated approach to evaluating 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 interpretation of the story). Lastly, this study employs the sentential coordination index, which can measure language complexity. 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 paper seeks to take steps toward DMC assessment validation. Research and pedagogical implications will be discussed.

Poster Abstracts

Revisiting the issues in L2 proficiency assessment in SLA research

Jinman Fu, Jungwon Yoo, Sun-Young Shin

Saturday September 9: 1:00 pm - 2:00 pm *Location: Room 300*

Assessing second language (L2) proficiency, whether on a global or local scale, is crucial for research on second language acquisition (SLA). Proficiency itself is often considered a dependent variable, impacting other independent variables, or an important independent variable influencing the developmental sequences of specific aspects of target L2 knowledge in cross-sectional research designs. Recognizing the significance of precise assessment and reporting of L2 proficiency levels in SLA research, Thomas (1994) conducted a comprehensive survey of SLA literature published in major journals. She discovered that L2 proficiency was sometimes inappropriately evaluated, jeopardizing the validity and generalizability of research findings. However, with the advancements made since Thomas published her article nearly 30 years ago, the techniques for assessing L2 proficiency have significantly improved. The introduction of computer adaptive tests, integrative tests, web-based tests, and automated scoring systems now enables SLA researchers to measure and utilize L2 proficiency in a more reliable and valid manner than ever before. It is thus worthwhile to investigate how L2 proficiency is currently assessed and reported in recent SLA studies published in major journals, aiming to identify any advancements in L2 proficiency assessment within SLA research. Additionally, there has been limited exploration into how L2 proficiency influences various aspects of SLA outcomes across different studies in a meta-analytic manner. To address these gaps, we conducted a study examining data-based articles published in major SLA journals during the six-year period from 2017 to 2022. Our investigation focused on identifying common methods of measuring and determining L2 learners' proficiency levels. This paper presents the results of a meta-analysis of L2 proficiency as an independent variable in diverse SLA contexts, as well as improvements in practices related to L2 proficiency assessment within the SLA research context.

Poster Abstracts

Examining the use of sources in AI-generated writing test responses

Rurik Tywoniw

Saturday September 9: 1:00 pm - 2:00 pm *Location: Room 300*

Artificial intelligence tools offer new opportunities for efficient natural language generation, but also present problems for the measurement of individual writing ability. As AI tools become more available to the public, the way we compose texts is changing. It remains the responsibility of language testers to understand what aspects of written production are within the domain of assessing a writer's capability and what aspects are reliably handled by writing aids such as ChatGPT.

One aspect of natural language generation in AI chat tools which has come under scrutiny is how the tools utilize or fail to utilize sources of factual information. Two research questions are investigated in this study. First, in what proportion of essays generated by AI is inaccurate or non-factual information present in the essay, and second, how is source use different depending on whether the AI is given specific information or left to utilize information in its training data. 50 AI-generated responses to writing prompts in a university standardized English placement test are examined in two different conditions: giving the AI tool the writing prompt only, or giving the AI tool the writing prompt and integrated reading sources. These essays were rated for overall quality, the use of factual information, and the level of reliance on sources of information. Linear modeling is used to compare the presence of false information in texts in the two conditions, as well as the level of reference to sources between the two types of texts. Results indicate that providing the AI with sources reduces factual errors, but increases direct borrowing in the output. Implications for test-development and scoring are discussed.

More than four decades of research supports the validity of the **TOEFL® Family of Assessments**

Choi, J. S. & Loewen, S. (2022). Exploring young learners' strategic behaviors in a speaking test. *TESOL Quarterly*, 56(4), 1384–1396. **http://doi.org/10.1002/tesq.3136**

Frost, K., Clothier, J., Huisman, A., & Wigglesworth, G. (2020). Responding to a TOEFL iBT integrated speaking task: Mapping task demands and test takers' use of stimulus content. *Language Testing*, 37(1), 133–155. **https://doi.org/10.1177/0265532219860750**

Hsieh, C.-N., & Wang, Y. (2019). Speaking proficiency of young language students: A discourse-analytic study, *Language Testing*, 36(1) 27–50. **https://doi.org/10.1177/0265532217734240**

Hui, B., Wong, S. S. Y., & Au, R. K. C. (2022). Reading aloud listening test items to young learners: Attention, item understanding, and test performance. *System*, 108, 102831. **https://doi.org/10.1016/j.system.2022.102831**

Kim, M., Nam, Y., & Crossley, S. (2022). Roles of working memory, syllogistic inferencing ability, and linguistic knowledge on second language listening comprehension for passages of different lengths. *Language Testing*. **https://doi.org/10.1177/02655322211060076**

Kyle, K., Eguchi, M., Choe, A. T., & LaFlair, G. T. (2022). Register variation in spoken and written language use across technology-mediated and non-technology-mediated learning environments. *Language Testing*, 39(4), 618–648. **https://doi.org/10.1177%2F02655322211057868**

Papageorgiou, S., Davis, L., Ohta, R., & Garcia Gomez, P. (2022). Mapping TOEFL® Essentials™ test scores to the Canadian Language Benchmarks (TOEFL Research Report No. 100). ETS. **https://doi.org/10.1002/ets2.12357**

Papageorgiou, S., & Manna, V. F. (Eds.) (in press). *Meaningful language test scores: Research to enhance score interpretation*. John Benjamins.

Révész, A., Michel, M., Lu, X., Kourtali, N., Lee, M., & Borges, L. (2022). The relationship of proficiency to speed fluency, pausing, and eye-gaze behaviours in L2 writing. *Journal of Second Language Writing*. Advance online publication. **https://doi.org/10.1016/j.jslw.2022.100927**

Roever, C. & Ikeda, N. (2022). What scores from monologic speaking tests can(not) tell us about interactional competence. *Language Testing*, 39(1), 7–29. **https://doi.org/10.1177/02655322211003332**

https://www.ets.org/toefl/research/

ETS TOEFL®