Building Pathways Between Language and Assessment

MwALT 2019

Midwest Association of Language Testers
Bloomington, Indiana
October 4-5

IU SECOND LANGUAGE STUDIES

GRADUATE CERTIFICATE IN TESOL AND APPLIED LINGUISTICS

☐ A practical two-semester program designed to enable students to work as successful teachers of the English language to adult speakers of other languages.

M.A. IN TESOL & APPLIED LINGUISTICS

☐ The most recognized advanced degree in ESL / EFL and a platform for further study.

M.A. IN SECOND LANGUAGE STUDIES

☐ A foundation for advanced research in SLA, encompassing analyses of a variety of first and second languages

PH.D. IN SECOND LANGUAGE STUDIES

☐ A comprehensive, innovative, and pioneering doctoral program, at the cutting edge of contemporary research

---

Application deadlines for the following Fall Semester:

December 1st (international)
January 15 (domestic)

We are always more than happy to speak with prospective students, and further your understanding of SLS at IU. Please do not hesitate to contact any of the SLS faculty with questions concerning our program.

https://dsls.indiana.edu/ dsls@indiana.edu

Over 70 other languages are taught at IU, and learners of English come from many different language backgrounds.

*"Indiana University is a true **candy store** for anyone interested in languages"*

Rex Sprouse

It is a challenging, yet ultimately fulfilling journey.

INDIANA UNIVERSITY
FULFILLING *the* PROMISE

Welcome to MwALT 2019 at Indiana University!

Welcome to the 21st annual conference of the Midwest Association of Language Testers. The Department of Second Language Studies at Indiana University is delighted to host the MwALT conference for the first time in Bloomington, Indiana.

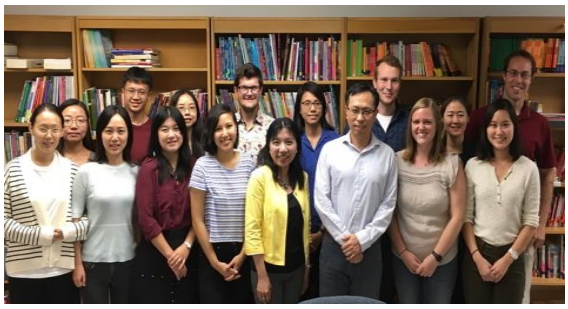
This year's conference theme is *Building Pathways Between Language and Assessment*. We hope that this coming conference will provide a platform for us to think about the term "language assessment" and how language and assessment are connected. While we have seen advancements in the way we assess, it is nevertheless of paramount importance to think about the roots—what we are testing. We hope that this conference provides an opportunity for language testers to address the questions of what, why, how we are testing, and weave the answers together in a holistic manner whereby answers to one question inform answers to other questions. At MwALT 2019, we hope to revisit language in language assessment and strengthen the ties between second language acquisition and how to better assess it.

Sun-Young Shin
Associate Professor, Department of Second Language Studies, *Indiana University*
MwALT 2019 Conference Chair

Conference Organizing Committee

Yena Park & Suchada Sanonguthai (Co-chairs)
Members of the IU Language Assessment Lab

Jaehyun Ahn	Emily Moscaritolo
Bihua Chen	Lucas Murphy
Jaesu Choi	Brian Rocca
Jungyoun Choi	Shaye Smith
Jean Young Chun	Kimberly Wan
Fengming Liu	Lijuan Zhang



Special thanks to Senyung Lee, *Northeastern Illinois University*

Abstract Reviewers

Carol Chapelle, *Iowa State University*
Mark Chapman, *WIDA*
Bihua Chen, *Indiana University*
Jean Young Chun, *Indiana University*
Deborah Crusan, *Wright State University*
April Ginther, *Purdue University*
Ryan Lidster, *Indiana University*
Senyung Lee, *Northeastern Illinois University*
Yuchen Liu, *University of Iowa*
Wenyue Melody Ma, *Michigan State University*
Gary Ockey, *Iowa State University*

Lia Plakans, *University of Iowa*
Dan Reed, *Michigan State University*
Sharareh Taghizadeh Vahed, *Purdue University*
Scott Walters, *UIUC*
Fang Wang, *University of Iowa*
Yangting Wang, *University of Texas at San Antonio*
Paula Winke, *Michigan State University*
Xun Yan, *UIUC*
Xiaowan Zhang, *Michigan State University*

MwALT 2019 Sponsors

Thank you to the organizations and departments sponsoring 21st MwALT at Indiana University:

Indiana University

Department of Second Language Studies
College of Arts and Sciences
Center for Languages of the Central Asian Region
Department of East Asian Language & Culture
Department of Spanish & Portuguese
Department of Germanic Studies

Table of Contents

Questions? Contact mwalt19@iu.edu

Workshops

Workshops will be held in Kirkwood Hall, room 016.

Workshop I:

Introduction to Using Signal Detection Theory (SDT) for Testing Discrimination and Severity for Binary and Rated Response Measures

Workshop Leader: Ryan Lidster, *Indiana University*
Time: 10:00am - 1:00pm, Friday, October 4, 2019

The use of signal detection theory (SDT) measures such as d', c, and A' is already common in fields such as psychology and medicine, but it is becoming increasingly common for a wide variety of L2 domains including language testing for item analysis, examining rater behavior, and especially standard setting, and in L2 acquisition broadly for the analysis of grammaticality judgment tasks and other binary and Likert-scale rating decisions. The primary advantage of using these measures over other measures of item and test performance is that they explicitly separate out sensitivity (i.e. the ability of the person or testing instrument to discriminate between levels of the construct) and bias (i.e. severity or lenience of the raters, items, and so on). The goal of this workshop is to explain how these measures differ from other statistical techniques, delineate the contexts in which using SDT would be preferable to alternatives, and then work through examples with participants in SPSS, Excel, and R to become familiar with calculating and interpreting the results. By the end of the workshop, participants will be able to determine when and how to use SDT measures to enhance the informativeness of test and testee performance data.

Workshop II:

Differential Item Functioning (DIF) Analysis in Language Assessment Research

Workshop Leader: Sun-Young Shin, *Ph.D., MwALT 2019 conference chair*
Time: 3:00pm - 6:00pm, Friday, October 4, 2019

Differential Item Functioning (DIF) occurs when examinees from different groups but at the same ability level respond differently to certain items. DIF analysis allows us to identify items performing differently across different groups and to detect potentially biased items. Detecting DIF items is thus considered an important step in developing a new test, adapting a present test for a new context, or validating inferences and uses of test scores particularly when distinct subgroups are involved in the high-stakes test. This workshop on DIF analysis will introduce the participants to the basic concepts and different DIF detection methods. Participants will also learn how to run difR R package (Magis et. al., 2010) to detect DIF items associated with specific subgroups using the existing testing data.

Plenary

Second Language Speech and Variability in Language Assessment
Okim Kang, *Northern Arizona University*

8:40 – 9:35 am Saturday, October 5
Georgian Room, Indiana Memorial Union

In the field of second language (L2) assessment, there is an increasing need for a comprehensive understanding of linguistic varieties on the part of both test practitioners and test takers. Language variation can advantage or disadvantage a learner beyond day-to-day interactions in the era of globalization. Especially relevant to the new movement toward English as an international language and the growing acceptance of English varieties is the assessment of listening and speaking skills. The presenter will review the effect of test taker individual difference and linguistic variation on their testing performance, discuss the relationship between linguistic properties and rater perception of L2 speech, and explain the variance attributable to rater background and attitudinal variables on assessment of spoken English. She will also address effective practice in listening/speaking assessment that promotes a World Englishes approach and discuss how the variability of L2 speech can be managed in a real-world context.

Kang's research focus lies in the areas of second language (L2) oral assessment and testing, speech production and perception, L2 pronunciation and intelligibility, and language attitudes. Her overall research goal is to investigate the nature of accented speech of non-native speakers of English, which includes several sub-areas of research: (a) how accent is assessed by listeners, (b) how accented speech is characterized linguistically, (c) how the assessment of accented speech is validated through automatic systems, and (d) how speakers with accents can better communicate with others.

Conference Schedule

Friday, October 4, 2019 **Pre-Conference Day**			
Workshop 1 10:00am- 1:00pm	**Introduction to Using Signal Detection Theory (SDT) for Testing Discrimination and Severity for Binary and Rated Response Measures** Workshop Leader: Ryan Lidster Venue: Kirkwood Hall 016		
Workshop 2 3:00-6:00pm	**Differential Item Functioning (DIF) Analysis in Language Assessment Research** Workshop Leader: Prof. Sun-Young, Shin, MwALT 2019 Conference Chair Venue: Kirkwood Hall 016		
Saturday, October 5, 2019 **Conference Day**			
7:30-10:30 am	On-site registration (Georgian Room)		
7:30-8:20am	Coffee and light refreshments (Georgian Room) [Sponsored by **WIDA**]		
7:30-8:20am 12:00-12:30pm	Poster setup (Oak Room)		
8:30-8:40am	Opening Remarks (Georgian Room)		
8:40-9:35am	Plenary Speech (Georgian Room): Prof. Okim Kang, *Northern Arizona University* **Second Language Speech and Variability in Language Assessment**		
Paper sessions	Sassafras Room	Redbud Room	Walnut Room
Paper Session 1 9:50-10:20am	**Linking Language and Assessment: Longitudinal Evidence for Oral Proficiency Development in College Foreign Language Learners** Xiaowan Zhang Paula Winke Shaunna Clark	**Concept Mapping for Guiding Rater Training in an ESL Elicited Imitation Assessment** Jie Gao David Crouch Lixia Cheng	**Distinguishing Features in Oral Performance of L2 Chinese Learners** Yuyun Lei
Paper Session 2 10:20-10:50am	**L2 Oral Fluency Development: The Lexico-syntax of Large Fluency Gainers** David Crouch	**Deconstructing Rating and a Rating Scale: How a Binary, Analytic Scale Guides Raters through a Holistic, Profile-based Rating Scale** Hyunji Park Xun Yan	**What are the Linguistic Bases for Construct Definition in Language Assessment?** Carol Chapelle YunDeok Choi
10:50-11:00am	Coffee Break (Georgian Room)		

Paper sessions	Sassafras Room	Redbud Room	Walnut Room
Paper Session 3 11:00-11:30am	**The Relationship between an Interim and a Summative Assessment of English Language Proficiency** Mark Chapman David MacGregor Kyoungwon Bishop	**Assuring Score Quality: A Framework for Making Rater Certification Decisions in Large-Scale Testing** Sharon Pearce Patrick McLain	**Providing Reliability Evidence for Elicited Imitation Using Generalizability Theory** Xiaorui Li
Paper Session 4 11:30am-12:00pm	**Using Summative Assessment for Formative Purposes: The Process of Developing Detailed Learner Feedback for Standardized Testing** Rachel Basse Sally Thelen Luke Slisz Susan Haines	**Optimization through Standardization: Investigating the Efficacy of Standardized Training on Peer Assessment of ESL Writing** Erika Latham Xun Yan	**Assessing the Dependability of a Scenario-based Test of Spoken Argumentation and the Impact of Choice on Performance** Jorge Beltran Zuniga **Winner of 2019 Best Student Paper Award**
12:00-1:00pm	Lunch (Georgian Room) [Sponsored by **Duolingo**]		
12:00-1:15pm	Business Meeting (Walnut Room)		
1:00-1:50pm	Poster Session (Oak Room)		
Paper Session 5 2:00-2:30pm	**Investigating rater (dis)agreement in a local ITA speaking test: A mixed-method study using multilevel modeling and semantic network analysis** Ji-Young Shin	**Applying Kane's Argument-Based Approach to Validity: Score Interpretation for a Test of Listening for Conversational Implicature** Stephen O'Connell	**Assessing L2 Collocation Knowledge: What do Different Test Response Formats Tell Us?** Senyung Lee
Paper Session 6 2:30-3:00pm	**Location, Cost, and Time as Barriers to Test Takers** Geoff LaFlair Burr Settles	**Developing a Scale to Assess L2 Email Pragmatics: A Comparison of a Holistic versus Analytic Scale** Ananda Muhammad Taichi Yamashita	**Three Lexical Progress Measures in the Acquisition of Hebrew by Arabic-speaking High School Students in Southern Israel** Eihab Abu Rabiah

Conference Schedule

Paper sessions	Sassafras Room	Redbud Room	Walnut Room
Paper Session 7 3:00-3:30pm	**Test Review: Pragmatics in the Business Japanese Proficiency Test** Paul Richards	**The Development and Validation of a Contextual Interpretation Ability Test as a Measurement of Language Pragmatic Aptitude: A pilot study** Yoonjee Hong Steven Ross	**Identifying what Learners "Cannot Do": Application of Systemic Functional Linguistics to Development of a Diagnostic Grammar Assessment** Roz Hirch
3:30-3:45pm	Coffee Break (Georgian Room)		
Paper Session 8 3:45-4:15pm	**A Statistical Index of Cheating (and Rule Breaking) in a High-stakes Computer Adaptive Language Assessment** Connor Brem Jenna Lake Geoff LaFlair	**A Rhetorical Model of Directed Self-Placement for Second Language Writers** Zhaozhe Wang	**Developing an Empirically-driven Aural DCT for Pragmatics Assessment** Kathleen Bardovi-Harlig Yunwen Su
Paper Session 9 4:15-4:45pm	**Resolving Mistriangulations between CEFR and the Lexile Scale by Using Both Test Scores and Expert Judgment** Alistair Van Moere Jing Wei Ji-Young Shin	**Validating an English Oral Communication Placement Test** Shireen Baghestani Sonca Vo Gary Ockey	**Developing an L2 Pragmatic Speaking Test Using Conversation Analysis Findings** Shi Chen
5:00-5:20pm	Awards ceremony and closing remarks (Georgian Room)		
5:20-5:40pm	Closing ceremony: **Gayageum Performance** by Eunsun Jung (Georgian Room)		

Paper Session 1 9:50-10:20 am

Sassafras Room

Linking language and assessment: Longitudinal evidence for oral proficiency development in college foreign language learners

Xiaowan Zhang, *Michigan State University*, zhang874@msu.edu
Paula Winke, *Michigan State University*, winke@msu.edu
Shaunna Clark, *Michigan State University*, slclark@msu.edu

There is a paucity of longitudinal research on second language acquisition (SLA), despite the fact that many questions asked by SLA researchers are about developmental processes (Ortega & Iberri-Shea, 2005). This study investigates SLA as a developmental process by examining the oral proficiency development of students studying foreign languages in the postsecondary context. The objective of this study is twofold: (a) to describe the oral proficiency growth of students enrolled in lower-division foreign-language courses; and (b) to understand the effects of four individual differences variables (heritage status, high-school language education, motivation, and L2 contact) on the shape and rate of students' oral proficiency growth.

Three years of oral proficiency data from 1922 students studying Chinese (n = 219), French (n = 483), Russian (n = 120), and Spanish (n = 1100), were collected by means of ACTFL Oral Proficiency Interview by Computer (OPIc), together with survey data on the four aforementioned individual differences variables. A latent growth curve analysis showed that the sampled students on average spoke at a level between Novice-Mid and Novice-High at Level 102 (first year, second semester level) and at a level slightly below Intermediate-Low at Level 202 (second year, second semester level). The total growth from Levels 102 and 202 was slightly more than one sublevel on the ACTFL scale, with slightly more gain occurring between Levels 201 and 202 than between Levels 102 and 201. All individual differences factors were significant predictors of the mean growth curve (as defined by initial proficiency at Level 102 and rate of growth) except for L2 contact. The target language also had significant effects on initial proficiency and change over time, indicating that learners of different languages might progress following different paths in speaking. Findings of this study have important theoretical and practical implications for foreign language education.

Redbud Room

Concept mapping or guiding rater training in an ESL elicited imitation assessment

Jie Gao, *Purdue University*, gao339@purdue.edu
David Crouch, *Purdue University*, crouchd@purdue.edu
Lixia Cheng, *Purdue University*, clixia@purdue.edu

Elicited Imitation (EI) has been integrated into second language (L2) assessments measuring examinees' overall language proficiency (Tracy-Ventura et al., 2014) or examining L2 learners' language development (Ellis et al., 2006). Our study focused on rater behavior when judging L2 learners' EI performances on a local English proficiency test. Implementation of a 5-point holistic rating scale from 0 to 4, with rater training, has rendered high rater agreement

(above .90 for R1/R2 correlation) at the section level. Raters, however, seem to operate with different priorities when making decisions at the lower end of the scale.
We investigated 1/2 rater splits regarding the same item response. Two trained raters rated 56 examinee responses. Of the total 672 EI sentences, 125 1/2 splits were identified. Based on transcriptions and detailed error analyses, a Performance Decision Tree (PDT) was developed with the purpose of fine-tuning the decision-making process at the lower levels of the scale and helping raters align better with each other and with the rating scale. This PDT guides raters to make grammaticality judgements of each item response and then identify semantic deviations at the word level. While the grammaticality judgements cover grammatical accuracy, the semantic comparisons between the examinee's version and the prompt include minor or major meaning deviations resulting from word substitution, addition, omission, or distortion (using a completely different word).

Preliminary results show that 59 of the 125 sentences (47.2%) have grammatical errors. Semantic deviations appear in 98 sentences (78.4%), 50% of which result from word omission. Word addition, substitution, and complete distortion contribute 2%, 15.3%, and 21.4% respectively. The remaining 7% of semantic deviations result from combinations of the aforementioned categories. This study has contributed to our ongoing rater training, with the construction of this PDT to help raters navigate through the lower end of the rating scale.

Walnut Room

Distinguishing features in oral performance of L2 Chinese learners

Yuyun Lei, *University of Illinois at Urbana-Champaign*, lei13@illinois.edu

Distinguishing features are crucial for characterizing second or foreign language development for both teaching and assessment purposes. Yet little research has been devoted to distinguishing features used in scoring oral performance of second language (L2) learners of Chinese, with only a handful of studies dealing with advanced L2 Chinese learners (Jin & Mak, 2013; Chen, 2015). The present study includes all levels of L2 learners typically found in a university-level Chinese language program and explores what quantifiable features can distinguish different levels of oral performance holistically judged by experienced language instructors. Twenty L2 Chinese learners and eight native speakers at a Midwest US university were asked to complete an oral narrative task. Their oral responses were scored and analyzed for an array of individual features, including features of fluency, lexical and grammatical accuracy, lexical variation and sophistication, and grammatical complexity. The results showed that features from each category could help distinguish assessed levels of L2 Chinese oral performance. Among them, rate features of speech, length of pausing, number of lexical and syntactic errors, word type, vocabulary profile, and mean length of clauses could reasonably be selected as descriptors of L2 Chinese oral performance across levels. The study provides implications for the development of rating scales of speaking assessment in L2 Chinese contexts.

Paper Session 2 10:20-10:50 am

Sassafras Room

L2 oral fluency development: The lexico-syntax of large fluency gainers

David Crouch, *Purdue University*, crouchd@purdue.edu

The theory underlying L1 and L2 oral fluency has focused on cognitive processes, particularly proceduralization (Anderson, 1983; Raupach, 1987) and linguistic constructs, especially lexico-syntactic formulation (Levelt, 1989). Towell et al. (1996) argued that development of formulaic language enables automatic speech production. However, no research has studied the longitudinal development of L2 oral fluency concurrently with any of the following lexical variables: lexical frequency profile, formulaic sequence use, and D measure.

In the present study, I calculated oral fluency using the transcribed oral test responses of 100 L1-Chinese EAP learners at the beginning and end of a required two-course (two semester) EAP sequence at a US university. The task completed was a computer-administered, two-minute "express your opinion" task. This study included eight oral fluency measures: speech rate, mean length of speech run, articulation rate, phonation-time ratio, average length of silent pause, average length of filled pause, silent pauses per minute, and filled pauses per minute. For the ten participants who made the largest percentage-wise oral fluency gains (in terms of the oral fluency variable associated with the largest Cohen's d effect size), I analyzed the oral transcripts to compute descriptive statistics for the three lexical variables mentioned above and three syntactic variables: coordinate clause ratio, dependent clause ratio, and words per T-unit.

Eight pre-post paired sample t-tests (with Bonferroni-adjusted alpha-level of .00625) indicated significant change ($p<.00625$) in all of the oral fluency measures, except average length of silent pause and average length of filled pause. The gain in mean length of speech run exhibited the largest effect size (Cohen's $d=0.41$) of the six significant variables. Furthermore, of the lexical and syntactic variables, the only large longitudinal change was associated with coordinate clause ratio, which increased by 34.33%. Implications will be discussed in the presentation.

Redbud Room

Deconstructing rating and a rating scale: How a binary, analytic scale guides raters through a holistic, profile-based rating scale

Hyunji Park, *University of Illinois at Urbana-Champaign*, hpark129@illinois.edu
Xun Yan, *University of Illinois at Urbana-Champaign*, xunyan@illinois.edu

This two-phased, sequential mixed-methods study investigates how rater behavior is influenced by different rating scales on a college-level English as a second language (ESL) writing placement test. In Phase I, nine certified raters rated 152 essays using a holistic, profile-based scale; in Phase II, they rated 200 essays using a binary, analytic scale developed based on the holistic scale and 100 essays using both rating scales. Rater performance was examined both quantitatively through Rasch modeling and qualitatively via think-aloud

protocols and semistructured interviews. Findings from Phase I revealed that, despite satisfactory internal consistency, the raters demonstrated relatively low rater agreement and individual differences in their use of the holistic scale. Findings from Phase II showed that the binary, analytic scale led to much improvement in rater consensus and rater consistency. The results suggest that the binary, analytic scale helped the raters deconstruct the holistic scale, reducing their cognitive burden. This study represents a creative use of a binary, analytic scale to guide raters through a holistic rating scale. Implications regarding how a rating scale affects rating process are discussed.

Walnut Room

What are the linguistic bases for construct definition in language assessment?

Carol Chapelle, *Iowa State University*, carolc@iastate.edu
YunDeok Choi, *Sungkyunkwan University*, yundeokchoi@gmail.com

Representation of language over the course of development is central to theory and research in both language assessment and second language acquisition (Ellis, 2017). In language assessment, the definition of the construct assessed by a test is required for test creation and validation (Chapelle, 1989). Assumptions about a construct include, for example, its components, scope, and differences at the lower vs. the higher levels. In view of the shared concerns between language assessment and second language acquisition, one might expect these two areas of applied linguistics to engage with the same empirical and theoretical concepts about language. Yet, in published research on language assessment, references to such a shared domain seem to be limited even though no studies to date have explicitly investigated such connections.

This paper examines if and how language testers draw upon conceptions of language consistent with those in second language acquisition. Our corpus for analysis consists of over 100 articles published in *Language Testing* from 2007 through 2018, including all articles that refer to a language construct as part of the study. The papers were divided into sub-groups based on the language construct(s) defined in the paper (e.g., reading, writing, listening, speaking, vocabulary, grammar, pragmatics, and academic language). We examined the construct definitions provided by authors to identify any connection between their construct definitions and the theory or research perspectives in second language acquisition. Judgments of commonality were made based on standard references on second language acquisition and were agreed upon by the two authors, who sought additional opinions as needed. Results are reported chronologically and by sub-group representing the various constructs investigated, and conclusions are drawn about the shared domain of understanding about language based on this analysis.

Paper Session 3 11:00-11:30 am

Sassafras Room

The relationship between an interim and a summative assessment of English language proficiency

Mark Chapman, *WIDA*, mark.chapman@wisc.edu
David MacGregor, *WIDA*, macgregor3@wisc.edu
Kyoungwon Bishop, *WIDA*, kyoungwon.bishop@wisc.edu

Annual summative assessments of English language proficiency (ELP) can provide valuable information on the developing ELP of English learners (ELs). However, they provide only a snapshot of students' ELP, and often the results are delivered long after the test administration. Therefore, schools may look to interim assessments as measures of developing ELP that can be administered during the school year and provide immediate results. To be maximally useful, there should be a strong, positive predictive relationship between the interim and the summative assessments.

In this study we examined the predictive relationship between WIDA's MODEL, an interim assessment of ELP, and the summative assessment ACCESS for ELLs. Both assessments are operationalized from the WIDA English Language Development (ELD) Standards and written to similar test specifications, but feature different content and test designs. Data from 1,162 grade 1-12 students from two districts in member states of the WIDA Consortium was used. All students took MODEL some time prior to taking ACCESS. A multiple regression was run, using the grade level of the student, the number of days between the administration of the two tests (Days), and the student's scale score on MODEL (Scale score) as covariates. In addition, the interaction Days*Scale Score was investigated. Regression models were run for the domains of Listening, Reading, Writing, and Speaking, along with three composite scores.

Results on the Overall Composite score showed a significant positive relationship between scores on the two tests (R= 0.88), with a narrow dispersion of scores in the plot. Days and MODEL Overall composite scale score had significant positive parameter estimates, while the Days*MODEL Overall composite scale score has a small but significant negative parameter estimate. These results support the use of MODEL as an appropriate interim assessment to complement the use of ACCESS.

Redbud Room

Assuring score quality: A framework for making rater certification decisions in large-scale testing

Sharon Pearce, *Michigan Language Assessment*, pearce.s@michiganassessment.org
Patrick McLain, *Michigan Language Assessment*, mclain.p@michiganassessment.org

Before training raters to score constructed response performances, test developers must define an acceptable level of scoring accuracy. Due to rater variation (McNamara, 1996), expecting raters to consistently achieve exact matches with true scores is not realistic. Therefore, a variety of factors must be taken into consideration in determining what a

comfortable distance from the true scores looks like. Such considerations include the stakes and purpose of the test, the rating instrument and the level at which tolerances can be set (analytic criteria level, overall score level, etc.), the proposed tolerance's effect on the overall rater pass rate and the impact this has on capacity for test delivery, as well as the number of ratings a given performance will receive to arrive at a final score. The current study investigates several tolerance scenarios for certifying raters on a high-stakes, multi-level speaking assessment (CEFR levels A2-C1) and provides an example of how test developers can approach creating a framework for making certification decisions.

Rater training data from over 300 prospective speaking examiners was utilized to investigate the effects of a variety of different certification criteria. The effects of these different criteria were examined using many-facet Rasch measurement and generalizability theory. Many-facet Rasch measurement analysis was conducted to examine the effects of the different certification criterion on the overall leniency/severity of the examiner pool relative to the target ratings, and generalizability theory was used to examine and compare the reliability of the different scenarios. The results of these analyses will be discussed as well as the determination of which scenario yielded the best quality outcome - one that balances the need for robust quality assurance with the practical considerations of training speaking examiners in a large-scale testing environment.

Walnut Room

Providing reliability evidence for elicited imitation using Generalizability Theory

Xiaorui Li, *Purdue University*, li1828@purdue.edu

Elicited imitation (EI) is a widely used approach to assess second language (L2) proficiency. In an EI task, examinees are provided with a series of sentence stimuli with target language structures embedded, and examinees are asked to repeat the sentences as accurately as possible (Larsen-Freeman, 1991). In the late 1970s, EI received a series of critiques regarding its validity (e.g. Hood & Lightbown, 1978; Hood & Schieffelin, 1978; McDade, Simpson & Lamb, 1982). The major criticism is that examinees may complete EI tasks using mere rote repetition instead of L2 knowledge. The aim of the present study is to provide reliability evidence for EI via using Generalizability Theory (GT) and offer suggestions for the future improvement of the test administration. The EI task used in the present study was developed locally at Purdue University as a part of the Assessment of College English-International (ACE-In) to screen the English language proficiency of undergraduate students whose first language (L1) is not English. The test scores of one hundred fifty-nine freshmen were analyzed in this study. The result from the Generalizability study (G study) shows that examinee effect is accounted for 71.41% of the EI score variability. As the majority of the variance is contributed by the examinee effect, the study result suggests that the current EI test is a reliable measure of L2 proficiency. Meanwhile, more rater training sessions is also desirable as the rater effect (12.52%) claims the second main source of score variability. Although there are four forms of the current EI task, the form effect as well as the form-by-rater interaction effect are very small. A Decision study (D study) with 3 raters and 2 forms is also performed to explore further options of test administration. The generalizability coefficient of the given D study is 0.96.

Paper Session 4 11:30-12:00 pm

Sassafras Room

Using summative assessment for formative purposes: The process of developing detailed learner feedback for standardized testing

Rachel Basse, *Michigan Language Assessment*, basse.r@michiganassessment.org
Sally Thelen, *Michigan Language Assessment*, thelen.s@michiganassessment.org
Susan Haines, *Michigan Language Assessment*, slisz.l@michiganassessment.org
Luke Slisz, *Michigan Language Assessment*, haines.s@michiganassessment.org

The increase of high-stakes testing in evaluating the proficiency of students of English worldwide has raised questions about the purpose and use of these tests (Shohamy, 2017). While such testing gives learners useful information on proficiency level and is used to shape educational policy, it is also important to consider guiding student learning on a classroom level by incorporating elements of formative assessment (assessment for learning) (Stiggins, 2006). Providing meaningful feedback to learners is a key element of formative assessment. Such feedback should deliver valuable information to students about their performance, provide information that can help shape teaching, and close gaps between current and desired level (Nicol & Mcfarlane, 2006).

This paper describes the development of detailed feedback descriptors for a multi-level writing test (CEFR levels A1-B1) for English language learners aged 11 to 15. We describe the process of analyzing the rating tool (a comprehensive 17-point checklist-type scale) to determine meaningful categories for the feedback to be provided to test takers. We then discuss the process of drafting feedback descriptors which provide information on strengths and areas to improve in their writing. The feedback was written at a level appropriate for test takers with a positive and motivational tone. Finally, we describe the ways in which we were able to determine a flowchart to represent every possible outcome in test-taker performance. We also include some of the challenges which arose during this process, such as tying feedback to learning activities and automating the feedback creation. Finally, we address the measures taken to ensure the reliability and validity of the feedback given to test takers. This presentation aims to present an innovative way of combining formative assessment elements into high-stakes testing.

Redbud Room

Optimization through standardization: Investigating the efficacy of standardized training on peer assessment of ESL writing

Erika Latham, *Center for Applied Linguistics*, e.latham.academic@gmail.com
Xun Yan, *University of Illinois at Urbana-Champaign*, xunyan@illinois.edu

The importance of corrective feedback in L2 writing instruction is generally acknowledged. While corrective feedback is an important part of L2 writing instruction, peer review remains somewhat controversial as a valid assessment method. Many of the validity concerns surrounding peer assessment can be mitigated by training students in feedback practices, much like rater norming for standardized language tests. Unlike rater norming, however, the

way peer review training is delivered in practice can vary greatly even within the same writing program. This lack of standardization limits the effectiveness of peer review training and thereby the quality of students' comments.

This quasi-experimental, mixed-methods study compared the effects of a standardized online peer review training program with those of unstandardized classroom training. Data was collected from 37 university-level ESL writing students using a pre-review background questionnaire, peer review worksheet (eliciting comments on a sample essay), and a post-review questionnaire (eliciting participants' perceptions of peer review). Participants' worksheet comments were qualitatively coded based on their focus, specificity, and politeness, then converted to frequencies and percentages for quantitative group comparison. Responses to the post-review questionnaire were examined qualitatively to identify trends in student perceptions of peer assessment and training.

Quantitative results revealed that standardized training increased participants' specificity, politeness, and focus on content-related topics. Conversely, unstandardized classroom training had little effect on these variables and instead kept participants focused primarily on surface-level features like lexico-grammar and rhetorical structure, suggesting a need for ESL writing programs to reinforce, in both instruction and assessment, the equal importance of content and structure in academic writing genres. Qualitative results reveal a variety of perceived learning outcomes for all participants, demonstrating the value of peer review for Assessment as Learning rather than solely a source of corrective feedback.

Walnut Room

Assessing the dependability of a scenario-based test of spoken argumentation and the impact of choice on performance

Jorge Beltran Zuniga, *Teachers College, Columbia University*, jlb2262@tc.columbia.edu
**Winner of 2019 Best Student Paper Award*

When it comes to the study of L2 skills, one communicative function that has been readily explored in the context of writing but has not been examined in depth for speaking ability is the elaboration and defense of an argument. However, certain real-life tasks require that language learners display their ability to build and defend an argument orally (e.g., class discussions, debates). Therefore, a speaking test that resembles an argumentation cycle should be developed and examined.

The current study aimed to determine whether a scenario-based speaking test successfully elicited argumentative language. Moreover, it sought to examine the possible effects of choice on performance by implementing different testing conditions with a group of 71 EFL learners. An experimental group was able to choose the position to defend throughout the test (+choice), while two control groups (-choice) were assigned a position (either the position in favor or against the policy described in the scenario).

Results from multivariate G theory analysis suggest the test had a relatively high degree of dependability, corroborating the findings of classical test theory analyses. The analytic scales used to score the test, however, seemed to contribute differently to the composite test score of

the test. An examination of the -choice and +choice conditions suggests that rather than the variable of choice itself, differences in dependability across test forms could be attributed to the content of the prompts of the tasks.

Paper Session 5 2:00-2:30 pm

Sassafras Room

Investigating rater (dis)agreement in a local ITA speaking test: A mixed-method study using multilevel modeling and semantic network analysis

Ji-Young Shin, *Purdue University*, shin209@purdue.edu

Literature on speaking tests has examined sources of rater (dis)agreement in relation to reliability, construct relevance, and fairness issues, primarily focusing on rater-related variables (McNamara, 1994). Incorporating task-, rater-, and examinee-related sources into investigating rater disagreement may provide comprehensive insights into understanding complex characteristics of rater agreement. The present study extensively investigated sources of rater disagreement in a local English speaking test that screens international teaching assistants (ITAs) (N=1,276) from a mixed-method approach. First, sources of rater disagreement in three categories were quantitatively examined: task-related (e.g., item types, prompt types, integration), examinee-related (e.g., sex, academic background, first language (L1), country, Indian speakers, the current test scores), and rater-related factors (e.g., rater L1, experience, confidence, shared L1 among raters and examinees). Considering the nested data structure, multilevel modeling with an ordinal outcome (two- and three-level random-intercept models) was used (Snijders & Bosker, 2012). Next, a semantic network analysis of rater justification, including microscopic examination on selected themes, was conducted. Multilevel modeling results revealed marginal, non-significant variance from examinee L1 (Level-3) (ICC=0.05%, p=0.393) but large, significant variance from examinee (Level-2) (ICC=42.25%, $\chi^2(1,259)=7,324.629$, $p<0.001$), which indicated rater agreement was not systematically associated with examinee L1 but with specific examinee characteristics. Follow-up two-level, one-predictor random-intercept models identified three significant sources: score types (item versus overall scores), L1 sharing, and the ITA speaking test scores. Less proficient examinees' item scores, when two raters and an examinee shared L1, tended to have higher levels of agreement, although practical significance was not large. Semantic network analysis and thematic analysis of rater justification supported the quantitative findings, by highlighting more variability among adjacently-scored responses than exactly-agreed responses. The study findings generally supported construct relevance of the ITA test score interpretations, while making implications for rater training to increase exact agreement rate and decrease a L1 sharing impact on advanced-level responses.

Redbud Room

Applying Kane's argument-based approach to validity: Score interpretation for a test of listening for conversational implicature

Stephen O'Connell, *University of Maryland, College Park*, soconn@umd.edu

Kane's argument-based approach to validity has grown in prominence in the language testing field in recent years (Chapelle, 2018). However, as Chapelle (2018) pointed out, there is a need for more researchers who cite Kane to explicitly demonstrate the application of his argument-based approach to the score interpretations of their language assessments. A clear exemplification of how the approach is applied serves two goals: the primary one of showing the validity of a parwaticular score interpretation, and a secondary one of improving understanding and generating discussion around the complex task of demonstrating score validity.

This is the two-fold goal of this paper. A study was conducted to examine the relationship, in terms of subskills, between comprehension of conversational implicature and non-implicature listening comprehension. As part of the study, an English listening comprehension test was created to assess comprehension of conversational implicature and was administered to 251 Spanish L1 learners of English. The score interpretation claim requiring support was that successful performance on the test of conversational implicature was an indicator of an ability to successfully comprehend and thus engage in everyday conversation in English. Starting with the vast target domain of being able to interact competently in everyday conversation, and proceeding through the chain of inferences in Kane's approach (as laid out in Chapelle, 2011), the evidence collected during the study from a number of analyses (standard Rasch analyses, partial credit Rasch analyses, principal components analyses of residuals, and logistic regressions on participants' CEFR levels) will be examined in relation to the inferences in the chain that need support in order for the score interpretation to be defensible. The appropriateness and sufficiency of the evidence will be discussed, as will the appropriateness of the manner in which Kane's approach was applied.

Walnut Room

Assessing L2 collocation knowledge: What do different test response formats tell us?

Senyung Lee, *Northeastern Illinois University*, S-Lee65@neiu.edu

This study compares four collocation test response formats to determine how informative they are about adult L2 learners' collocation knowledge, knowledge of how to combine words in the target language. Although various test response formats for assessing L2 collocation knowledge have been proposed (e.g., Bonk, 2000; Gyllstad, 2007), psychometric properties of those test formats remain under-explored. Two research questions are addressed: To what extent do different collocation tasks distinguish among levels of L2 collocation knowledge? To what extent do different collocation tasks show different levels of difficulty?

Four elicitation tasks were developed to capture different aspects of academic English collocation knowledge: a sentence writing task, fill-in-the-blank task, eight-option multiple-choice task, and Yes/No acceptability judgment task. Four groups of ESL learners from

precollege to matriculated university students (*n* = 205) and two groups of NSs of English, undergraduates and ESL teachers (*n* = 85), completed the four tasks. Each task targeted the same sixty-four collocations selected from the literature (e.g., Ackerman & Chen, 2013), including verb-noun (e.g., *commit a crime*), adjective-noun (e.g., *wide variety*), adverb-adjective (e.g., *readily available*), and adverb-verb (e.g., *clearly indicate*) collocations. Corpus frequency, mutual information, and NS ESL teachers' judgments were used in the collocation identification process.

Results from the item response theory analysis showed that the eight-option multiple-choice task distinguished best among learners of different levels of L2 collocation knowledge, followed by the fill-in-the-blank task, sentence writing task, and Yes/No acceptability judgment task. Among the four tasks, the sentence writing task was most difficult, and the Yes/No acceptability judgment task was least difficult. The results suggest that if differentiating learners based on L2 collocation knowledge is important, using the eight-option multiple choice task would be effective. Results also suggest that using the fill-in-the-blank task alone would suffice when assessing productive knowledge of L2 collocations.

Paper Session 6 2:30-3:00 pm

Sassafras Room

Location, cost, and time as barriers to test takers

Geoff LaFlair, *Duolingo*, geoff@duolingo.com
Burr Settles, *Duolingo*, burr@duolingo.com

The warrant authorizing the utilization inference states that test scores should be useful for stakeholders (Chapelle et al., 2008). This warrant presupposes that one set of stakeholders, the test takers, have access to the test. Many high-stakes tests are administered in a test center. This model requires test takers to register and travel to a location for the test, which favors people who live in large cities, have the resources to travel, and/or have advance notice of a need to provide an indicator of language proficiency. This same model disadvantages people who live in rural areas, people who may not be able to travel to a test center location (e.g., refugees), and people who may need to show a certificate of language proficiency without much advance notice. High-stakes tests that are administered online have the potential to address these accessibility issues.

This paper explores test accessibility by comparing high-stakes language tests that follow the test center model with a high-stakes language test that is administered online on three properties: 1) access to test administrations (number of test centers per million people) and internet access per million people, 2) differences in test price at purchasing power parity (PPP) and 3) time difference between registration/administration and time between test completion and score reporting. Results of preliminary analyses show that on average there is one test center per 1.9 million people in the world while there is an internet connection for 1 in 2.2 people on average worldwide. Additionally, analyses of differences in test price at PPP indicate that the cost of the prevailing model (test center) accounts for up to 25% of countries' gross

domestic product (adjusted for PPP) whereas the online model accounts for up to 7%. Implications for the utilization inference will be discussed.

Redbud Room

Developing a scale to assess L2 email pragmatics: A comparison of a holistic versus analytic scale

Ananda Muhammad, *Iowa State University*, nanda@iastate.edu
Taichi Yamashita, *Iowa State University*, taichiy@iastate.edu

L2 pragmatics assessment is a recently investigated area of language testing, and current research mainly focuses on spoken pragmatic competence. Meanwhile, very few studies have addressed written pragmatic competence, for example, in an email communication context. Even when addressed, the studies seem to only assess the appropriateness of the speech act (e.g., apology, request, complaints) produced in the emails (Chen & Liu, 2016; Ishihara, 2010). However, email communication does not only consist of the speech act, but also other features that could contribute to the overall communicative adequacy of the message conveyed. Moreover, there appears to be no agreement yet on whether a holistic or analytic scale would be more appropriate to rate emails. Therefore, this study attempts to partially address the gap in L2 pragmatics assessment research, first by developing a holistic and analytic scale to assess L2 email pragmatics in the U.S. academic context, and then applying these scales to ratings of emails produced by ESL students at a U.S. university setting. The scales were developed by analyzing requesting emails produced by 26 ESL students. Features such as frequent pragmalinguistic and sociopragmatic errors, as well as frequently missing essential email elements (e.g., subject line, greetings, closings) were identified by the researchers to create the scale descriptors. Next, two raters were trained to use the rating scales before they rated a different set of requesting emails also produced by ESL students. A rater questionnaire was administered to capture raters' attitudes towards the scales. Reliability of the rating scales, comparison between raters' attitudes towards the holistic versus analytic scale, and implications for the assessment of email communication tasks will be provided.

Walnut Room

Three lexical progress measures in the acquisition of Hebrew by Arabic-speaking high school students in southern Israel

Eihab Abu Rabiah, *Ben-Gurion University of the Negev*, aehab@post.bgu.ac.il

This study evaluates three lexical measures of progress in the acquisition of Hebrew by fifty ninth- and eleventh-grade Arabic-speaking high school students in Israel's Negev: level of abstractness (i.e. the proportion of abstract nouns relative to all the nouns in the text tokens), lexical diversity by counting the types (distinct words) in relation to the overall number of tokens of words and lexical density (i.e. the ratio between content words and function words). The corpus consisted of fifty expository essays written in Hebrew. Each essay was tested using the three measures. The quantitative analysis reveals several findings. First, the abstractness of Hebrew increases with subjects' age – the abstractness measure that clearly showed this emerged as an effective measure at both ages. On the other hand, the lexical density test did not show the expected increase with age, and I suspect the reason for this is

that relatively high density in younger pupils indicates a lack of cohesion and not a high level of linguistic competence, as expected. In the eleventh grade, the Hebrew language curriculum of Israel's Arab high schools focuses on cohesion, and then the density decreases. Third, the lexical diversity It evaluates the effectiveness of these measures in testing lexical competence of Hebrew in the groups examined. T-Test also did not yield significant findings, apparently because this test is influenced by the length of the text, which in the ninth grade is still not sufficiently long for a reliable test. Conclusions: Lexical density and lexical diversity tests, at least in their current form, should be postponed until the eleventh grade, whereas the abstractness test is already reliable in the ninth grade and shows the expected increase with age and exposure to Hebrew as a second language among Arabic speakers in Israel.

Paper Session 7 3:00-3:30 pm

Sassafras Room

Test review: Pragmatics in the Business Japanese Proficiency Test

Paul Richards, *Indiana University*, pauricha@iu.edu

This test review describes and analyzes the components of the Business Japanese Proficiency Test (BJT), a standardized test used to measure “the ability [of non-native speakers of Japanese] to process and respond appropriately to provided information” in business contexts (Japanese Business Proficiency Test, n.d.). The BJT is a rare example of a commercial test that emphasizes pragmatic knowledge. For instance, items on the BJT require examinees to identify sentences with appropriate address terms and/or politeness markers for a variety of business situations with given power dynamics.

As the assessment of pragmatics is relatively understudied, this review highlights the pragmatic features assessed by the BJT and the types of tasks used for assessing those pragmatic features. The validity of the methods of pragmatics assessment are then evaluated in terms of Bachman & Palmer’s (1996) model of test usefulness and the published literature on measuring L2 pragmatics ability (Bardovi-Harlig, 1999, Bardovi-Harlig & Shin, 2014). Although the BJT makes use of several novel conventions that may be useful for designing pragmatic assessments for other languages (e.g., some prompts instruct the examinee to select the desired way an employee should respond in a given scenario), the validity of the spoken and written components of the test are suspect, as these constructs are assessed via multiple-choice items.

Redbud Room

The development and validation of a contextual interpretation ability test as a measurement of language pragmatic aptitude: A pilot study

Yoonjee Hong, *University of Maryland, College Park*, yhong125@umd.edu
Steven Ross, *University of Maryland, College Park*, sross@umd.edu

Even though context and language are both at the core of pragmatics, research in L2 pragmatics has focused more on testing the skills of language, and less on the role of context.

Using Kane's argument-based approach to validity framework (2006), this study proposes interpreting situational context as a type of individual difference that might affect L2 pragmatic ability. In order to validate this, a non-verbal contextual interpretation ability test was developed using muted video materials. A total of twenty-four extended discourse completion items were designed to represent a wide variety of situational context components. Each item represented one turn of the interlocutor, and the participant had to guess what the interlocutor said by interpreting non-verbal situational cues in the video. The scoring rubric, whose score ranged from one to three, reflected two core issues in pragmatics testing: illocutionary acts and preference structure. Using Rasch analysis, the results of the study demonstrated that the dispersions of the test takers and the items were reliable, and that almost all of the test items fit the model well. The design of the test items and scoring rubric, the reliability of the test, and the fit statistics of the items served as backings for the warrant, the Rasch analysis. This in turn validated the argument that adults without language disability showed variation in interpreting situational context.

Walnut Room

Identifying what learners "cannot do": Application of systemic functional linguistics to development of a diagnostic grammar assessment

Roz Hirch, *Iowa State University*, rhirch@iastate.edu

Diagnostic assessment has gained attention in recent years as a pedagogical tool to enhance language learning. One complication from a language testing perspective is that diagnostic assessments focus on learners' errors, or what learners cannot (yet) do, as this gives the most diagnostic information for stakeholders' decision-making processes. Researchers suggest that this negative focus breaks with current practices of "can do" statements (Alderson, 2007; Lee, 2015). Additionally, although several large-scale diagnostic tests have been developed, such as DIALANG and DELTA, in the absence of a generally accepted theory of language acquisition in applied linguistics, diagnostic tests are difficult to make broadly applicable (Alderson, 2005). A preferable method is that tests be developed for classroom use, to assist learners, teachers, and other stakeholders with making decisions about where to focus instruction at the individual, classroom, and program levels. Designing an assessment for the classroom means having a theory that is useful for the classroom; that theory then directs development of the assessment.

This presentation will discuss the process of developing a diagnostic grammar assessment, focusing particularly on the application of theory and how that affected the construct and design of the test. This test utilized systemic functional linguistics (SFL) as the theory of grammar because of the way it integrates grammatical and topical knowledge. In SFL, grammar is a series of language choices, both conscious and unconscious, that speakers make in order to convey ideas; researchers can then analyze those choices to understand how speakers constructed meaning (Ma & Slater, 2015; Thompson, 2014). The presentation will discuss how SFL was used in the development of multi-tier tasks to analyze grammar errors – what students cannot do – as well as a hierarchical model for error analysis of the tasks, to develop feedback to help learners and instructors change cannot do to can do.

Paper Session 8 3:45-4:15 pm

Sassafras Room

A statistical index of cheating (and rule breaking) in a high-stakes computer adaptive language assessment

Connor Brem, *Duolingo*, connor@duolingo.com
Jenna Lake, *Duolingo*, lake@duolingo.com
Geoff LaFlair, *Duolingo*, geoff@duolingo.com

The strength of the evidence for inferences in an interpretation/use argument is undermined when test takers cheat. The interpretations and uses of test scores are threatened when test--takers cheat, which can result in incorrect decisions.

Cheating on traditional fixed--form tests can be mitigated by controlling the test environment, using alternate forms, and through post--administration analyses of response patterns (e.g., Angoff, 1974; van der Linden & Sotaridona, 2004) and differences in subscores (Garenpayeh, 2014). Computer adaptive tests (CATs) that are administered remotely in at--home settings require a different model for detecting cheating because the environment cannot be manipulated, CATs ensure a unique test each administration, and there are not other test takers to copy.

CATs generate non-traditional data (e.g., biometric data and external device detection) and traditional data (e.g., subscore differences) that can be leveraged in algorithms to detect cheating. This paper reports on the development and use of such an algorithm—the risk index (RI)—which indicates the probability that a test taker has cheated. The RI is a logistic regression, and we present the results from a training on 28,000 tests, which includes 23 predictor variables derived from human proctor flags, biometric data (not demographic data), and response data (e.g., length and scores on subskills). The RI is calculated on tests between an initial round of combined human/automated proctoring (conducted after the administration of the test) and a second round of human proctoring. It provides the second--round human proctors with a probability of cheating index that can be used to prioritize risky administrations and ensure that all tests are reviewed and released in a timely manner (48 hours). We describe the selection procedures we used to arrive at the final model and how the RI is used in practice.

Redbud Room

A rhetorical model of directed self-placement for second language writers

Zhaozhe Wang, *Purdue University*, wang2839@purdue.edu

In this presentation, I propose a rhetorical model of directed self-placement (DSP) to be implemented in a second language writing program, which aims to fully recognize student agent's position, deliberation, negotiation, and appropriation in relation to the placement decision, and to engage the student in a "rhetorical rehearsal" before signing the placement contract. The proposed model emerged in response to the need to address concerns with validity and social justice issues related to current common practices of DSP across the U.S. I

begin with a review of the development and assumptions of DSP in current scholarship while situating the discussion in theories of rhetorical agency. Then, I introduce the rhetorical model of DSP, which engages students in a rhetorical act of "rhetorical rehearsal" — a trial performance of rhetorical positioning, deliberation, negotiation, and appropriation before making placement decisions. In practice, during a DSP rhetorical rehearsal session, staff from the writing program will distribute to each student a packet that contains the instructions — an introduction to the program and curriculum, a DSP questionnaire collecting students' basic information about students' literacy history, a prompt that guides students to write a literacy history essay that complements the questionnaire, and a prompt that directs students to justify their placement decisions. The deliverables of a rhetorical rehearsal session include the completed literacy history questionnaire, a brief literacy history essay, and a justification essay. Students will present the deliverables in the form of a DSP profile to academic advisors, who will then review their profiles, make recommendations, or re-place certain students into what they deem more suitable courses. Lastly, I illustrate the rhetorical model of DSP with modified DSP procedures at my institution to concretize and contextualize it with attention to administrative and material affordances and constraints.

Walnut Room

Developing an empirically-driven aural DCT for pragmatics assessment

Kathleen Bardovi-Harlig, *Indiana University*, bardovi@indiana.edu
Yunwen Su, *University of Utah*, yunwen.su@utah.edu

Second language pragmatics assessment is still dominated by six measures from Hudson, Detmer, and Brown (1992, 1995; Bardovi-Harlig & Shin, 2014) and few tasks have been designed for EFL learners (cf. Liu, 2007). This study explores the use of two tasks, an oral DCT (discourse completion task) that requires oral responses, but is more practical to score than a role-play (Roever, 2011), and an aural DCT (modelled after Liu's, 2007 written multiple-choice DCT for EFL learners and Teng and Fei's, 2013, aural DCT for Chinese instruction) to test EFL learners' knowledge of pragmatic routines.

The oral DCT presents 20 situations by computer; twelve items require responses to spoken turns, and eight require initiating turns. The aural multiple-choice DCT tests the same items, each presenting 4 options (played twice) produced by learners on earlier tests of the oral DCT (Bardovi-Harlig, 2009). 200 first-second year college students from three universities in Xi'an, China were given the two tests, counter-balanced one week apart.

Preliminary analyses of one-third of the multiple-choice DCT tests (N=67) show that first-year students scored an average of 7.31/20, in line with Roever's (2006) finding that test-takers without exposure to English-speaking environments scored low on a test of pragmatic routines. An item analysis shows that 7/20 items had an item difficulty below .30, suggesting that they were extremely difficult. This set of items, however, represents different scenarios from the difficult oral DCT items identified previously (Bardovi-Harlig, 2009). A possible explanation is that while the multiple-choice DCT lessens production demands, it increases listening demands; excluding distractors also adds to task difficulty. Factor analysis will be conducted and compared to the oral DCT data to further investigate the reliability of the assessment task.

This study provides insight into some issues in developing an informative aural multiple-choice task for pragmatic assessment.

Paper Session 9 4:15-4:45 pm

Sassafras Room

Resolving mistriangulations between CEFR and the Lexile Scale by using both test scores and expert judgment

Alistair Van Moere, *MetaMetrics*, avanmoere@lexile.com
Jing Wei, *MetaMetrics*, jwei@lexile.com
Ji-Young Shin, *Purdue University*, shin209@purdue.edu

When one scale is mapped to many tests, it is often the case that the triangulations of every mapping do not align. For example, Tests A, B and C might map to the Common European Framework of Reference (CEFR), but when they are all compared, the mappings between tests do not line up (e.g. deJong, 2009). The Lexile Framework is a universal reading scale (from approx. 0L to 1600L) which has been mapped to five international English proficiency tests that are also aligned to the CEFR (Smith & Turner, 2016). However, when Lexile and CEFR score ranges are all tabulated, it can be seen that adjacent levels overlap, making it difficult to classify borderline students whose Lexile measures fall within the overlapping CEFR ranges.

To address these issues, the current study employs a mixed-method design by supplementing a test-score approach with a content approach in linking the Lexile Scale to the CEFR. Data from previous linking studies were aggregated into a single data set and was visually represented in a scatterplot. The data was examined to understand patterns of change in students' Lexile measures as their CEFR levels progressed, and boundaries were validated through a modified item-descriptor-matching method (Ferrara, Peril, & Johnson, 2008). Forty reading passages representing the full range of Lexile measures were selected and presented to a panel of judges in random order. The judges analyzed the reading demand of each passage and formed an initial judgment on CEFR level. Next, the aggregated judgments were shown to the judges, who could change their judgments if needed. After the CEFR level of each reading passage had been finalized, the association between passages' CEFR levels and Lexile measures was examined to evaluate the appropriateness of the boundaries. Findings from the study provide an example for how to resolve mis-triangulations among multiple test mappings.

Redbud Room

Validating an English oral communication placement test

Shireen Baghestani, *Iowa State University*, shireenb@iastate.edu
Sonca Vo, *Iowa State University*, soncavo@iastate.edu
Gary Ockey, *Iowa State University*, gockey@iastate.edu

Post-entry English language assessment is a growing practice at universities worldwide and much of this testing is carried out through locally-developed English language placement tests. These tests require continuous evaluation to ensure they are working effectively and achieving their purpose. Validity arguments are a useful tool for carrying out such evaluation as they help test creators articulate the attributes required for a test to be defensible in its interpretation and use in a particular context.

The current study uses an argument-based validity framework to examine a locally-developed oral communication placement test at a large Midwestern university. The study investigates whether there is adequate evidence to support the Explanation and Extrapolation inferences by asking teachers to rate the oral communication ability of students who did not pass the placement test and were required to take an oral communication course. Seven instructors of this course were asked to rate their students (n = 85) on the extent to which they needed and were benefiting from the class. Instructors used a 7-point scale in which 1 indicated that the student needed the class but was benefiting little (the class was too difficult) and 7 indicated that the student did not need the class and was not benefiting in any way. The results showed that 80 students (94%) were judged as needing and/or benefiting from the class, whereas only 5 were judged as not needing or benefiting from the class at all. Closer inspection of these 5 students' placement test scores showed that they were all very close to the cut score needed to pass the test. These results provide preliminary evidence that the placement test is doing a good job of identifying students who need and/or can benefit from the oral communication class.

Walnut Room

Developing an L2 pragmatic speaking test using Conversation Analysis findings

Shi Chen, *Northern Arizona University*, sc2592@nau.edu

Due to cultural differences, it is difficult for second or foreign language learners to orally produce proper pragmatic skills in an academic context in an English-speaking country. Pragmatic competence research in assessment has shown that pragmatic competence can be measured using different tasks, and performance can be differentiated across distinct levels (Roever & Kasper, 2018; Youn, 2015). The practicality of administering L2 pragmatic competence assessment remains problematic since such tests involve high labor cost. Therefore, the researcher developed a pragmatic speaking test using Conversation Analysis (CA) findings based on the discursive approach to L2 pragmatics in PPT Slideshow. The turns delivered by the interlocutors in this pragmatic speaking test were pre-designed and pre-recorded. The discursive approach to L2 pragmatics serves as the predominant theoretical framework of the study, and some pragmatic theories (e.g., politeness theory, speech act theory) are employed. Moreover, target language use (TLU) domain guides the study to

ensure the test is representative of real-life EAP situations. The test includes 20 test items (e.g., Asking about the due date, Missed class, Discuss the topic for the final presentation) were developed based on on-campus scenarios. 33 test takers participated in the study, which takes 30 minutes for each test taker to finish. Multi-face Rasch Measurement (MFRM) was employed to examine whether the items and rating scale function as intended. The results reveal that test takers were able to take the test smoothly. However, the test items are relatively easy for the students. The CA findings provide a practical framework for test developers to create a L2 pragmatic speaking test.

MICHIGAN LANGUAGE ASSESSMENT

MwALT 2019

Attend Our Presentations

Assuring Score Quality:

A Framework for Making Rater Certification Decisions in Large-Scale Testing

Saturday, October 5, 2019 11:00 a.m.–11:30 a.m.

Sharon Pearce & Patrick McLain

Using Summative Assessment for Formative Purposes:

The Process of Developing Detailed Learner Feedback for Standardized Testing

Saturday, October 5, 2019 11:30 a.m.–12:00 p.m.

Rachel Basse, Sally Thelen, Susan Haines, & Luke Slisz

Research and Internship Opportunities

Michigan Language Assessment conducts and supports research to inform test development and revision, to provide evidence of quality and validity for its tests, and to contribute to knowledge in the language testing field. Visit our website to learn more about summer internships in language assessment and about funding opportunities under the Spaan Research Grant Program.
Michiganassessment.org/about-us/research

"I feel I not only learned a lot about language assessment but also met great people. The friendly environment where all staff members are eager to help each other and learn new things was one of the best parts of the internship."

– Senyung Lee, Indiana University, 2018 intern

"I am certain that this experience will have a great impact on my future career."

– Mayu Miyamoto, Purdue University, 2017 intern

"Throughout the internship, I genuinely felt that the work I was doing mattered, and I greatly appreciate all of the learning opportunities I had."

– Dan Isbell, Michigan State University, 2016 intern

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

Cambridge Assessment English

UNIVERSITY OF MICHIGAN

MichiganAssessment.org

Get your degree with a focus on Language Testing at MSU

Second Language Studies Ph.D. / MA TESOL

Our interdisciplinary 4-year **Ph.D. in Second Language Studies** and 2-year **MA TESOL** programs provide students with a firm foundation in a range of research techniques and approaches. Fundamental to our program is **student-faculty interaction**, whether in the classroom as students or outside the classroom in **joint research projects**. Students graduate with an **outstanding skill set** that allows them to contribute to Language Testing research, practice, and innovation.

www.sls.msu.edu / www.linglang.msu.edu

There are ample opportunities for students to receive full funding during their studies through a combination of assistantships, grants, and fellowships through the SLS program and MSU's **Center for Language Teaching Advancement** (CeLTA), **English Language Center** for international students, and a well-developed **Less Commonly Taught Languages** teaching program. Graduate assistants in the English Language Center's Testing Office receive hands-on experience in classroom-based and large-scale test development and research.

Learn more about our program, students, and graduates through our social media channels:

College of Arts & Letters
MICHIGAN STATE UNIVERSITY

https://www.facebook.com/msuslsphd/

https://twitter.com/SLSPhdProgram

http://msuslsspotlight.blogspot.com

https://vimeo.com/user18640935

Poster Session Georgian Room 1:00-1:50 pm

Abstracts appear alphabetically by first author's last name.

Peer assessment: A complementary tool to promote students' autonomy

Jose Fabian Elizondo Gonzalez, *Universidad de Costa Rica*, josefabian.elizondo@ucr.ac.cr

This research study explores the implementation of peer assessment in an impromptu group discussion set in an English as a Foreign Language (EFL) class of 20 students. Peer assessment is incorporated in this class as a way to enrich the students' own learning process and promote autonomy, based on the premise that, in conjunction with traditional testing, peer assessment helps students to develop a better understanding of the subject matter, their strengths and weaknesses, and their learning process in general (Crisp, Sambell, Mc Dowell & Sambell as cited in Thomas, Martin & Pleasants, 2011). The results of the study showed that the students took a proactive role in completing the checklists conscientiously and writing comments on their peers' performances, focusing on delivery, pronunciation, and grammar and vocabulary, respectively. Careful preparation of the activity and appropriate guidance to students were key to obtaining the results desired –students' autonomy, self-confidence, cooperation and motivation (Brown, 2004). Overall, the results obtained not only show that peer assessment can promote autonomy and cooperation among students, but it also has practical implications for instructors, as we learned that we can improve the instruments to generate more insightful learning experiences through peer assessment in future activities.

Rating duration as a factor in rating accuracy in second language oral assessment: A proposal study

Sondoss Elnegahy, *Iowa State University*, sondoss@iastate.edu

This proposed study aims to explore if the rating time is a factor in rating accuracy in English Placement Test for Non-Native Speakers of English: Oral Communication (EPT OC), in a Large Midwestern university, since rating takes place simultaneously while test takers are taking the test. The study also examines if there is a difference between time as a factor in the two different types of tasks in this test: an 'individual' task and a 'pair' task. Following an explanatory mixed-method approach, rater participants would be asked to rate an EPT OC of a previously video recorded and then would be asked to follow a stimulated recall protocol that would take place immediately after participants complete the. This qualitative data would be analyzed inductively and then thematically coded to help explain the quantitative results. The quantitative analysis would be conducted via two Multiple Regression analyses would determine to what extent is time a factor in rating accuracy between the two task types.

Gaining insights into university administrators' language assessment literacy

Roz Hirch, *Iowa State University*, rhirch@iastate.edu
Gary Ockey, *Iowa State University*, gockey@iastate.edu

Assessment literacy originally referred to the knowledge that educators required to improve education standards, but has since been expanded to include a wide variety of stakeholders,

including administrators; the same is true for language assessment literacy (LAL) (Ginther & Elder, 2014). One of the recent advances in LAL is the recognition that, although there are many stakeholders who require knowledge of language assessment in order to make relevant decisions, not all of them need the same amount of knowledge that language assessment researchers do (Taylor, 2013). In the case of administrators, they may not require knowledge of theory or technical skills, but they should at least have knowledge related to scores and sociocultural values. However, how much knowledge do administrators actually have? What do the people who use language assessments for admissions decisions know about the tests, and how do they use this knowledge?

This poster presents a work in progress regarding LAL of university administrators in the United States. The first stage was collecting publicly available data regarding which universities have language requirements for international admissions, what tests are generally accepted, and score requirements for those tests, focusing particularly on R1 universities. The second part of data collection will take a different methodological approach from most studies in this area, which tend to use case studies; instead, this study will develop a survey for university administrators who make decisions about and based on language tests. The poster will present a summary of the data that has been collected so far as well as share the subject areas of the survey that will be sent to administrators.

Development and evaluation of a web-based rating system to measure reading-to-write ability

Haeyun Jin, *Iowa State University*, haeyunj@iastate.edu

Given that discourse synthesis is a common practice in college-level academic writing, reading-to-write (RTW) tasks have been adopted widely in many L2 writing assessment contexts (Knoch & Sitajalabhorn, 2013). Researchers have mostly agreed on the augmented authenticity of these tasks (e.g., Shin & Ewert, 2015) in that they replicate the actual practices in university classrooms. Despite such a benefit, however, raters are faced with challenges due to the complex nature of RTW tasks because they have to focus on both reading- and writing-related dimensions (Gebril & Plakans, 2014; Weigle & Montee, 2012). Existing studies mostly suggested providing more robust rater training to address the complexities of rating RTW tasks. The proposed study seeks to explore how the affordances of technologies can be employed to address these challenges through the use of a web-based system that provides enhanced rating input for raters. The current work-in-progress presentation describes the process in which Patton's (2008) Theory of Action framework guided the development and evaluation of a web-based system for measuring RTW ability in the context of an English placement test at a Midwestern US university. The rating system is designed to present enhanced rating input with features intended a) to assist in rating the construct of source-text use and b) to reduce raters' cognitive loads such as highlighted verbatim words from the readings, reading texts embedded as a dialog box and test taker essay presented in integration with the rating rubric. This presentation will also discuss how initial validity evidence will be gathered for the use of the system focusing on rater's rating process, their perceptions, and reliability of the rating output. The current study will demonstrate an innovative application of technology to minimize potential rater variability in rating RTW tasks.

Technology enhanced items and young English learners: What construct are we measuring?

Ahyoung Alicia Kim, *WIDA, University of Wisconsin-Madison*, alicia.kim@wisc.edu
Rurik Lol Tywoniw, *Georgia State University*, rtywoniw1@gsu.edu
Mark Chapman, *WIDA, University of Wisconsin-Madison*, mark.chapman@wisc.edu

With the advancement of technology, online language assessments have incorporated potentially more authentic and engaging item types in the form of technology enhanced item (TEI). TEIs are computer-delivered items that include specialized interactions for collecting response data, such as

hot spot and *drag and drop* items. These items differ from traditional multiple-choice items in terms of how students engage with the test interface and select a response. For example, hot spot items allow test-takers to click highlighted areas within a larger picture. The development of TEIs reflect a desire to create items that reflect the overall construct measured on the assessments (Sireci & Zenisky, 2006). Although there has been research on TEIs in math assessments (Crabtree, 2016), few studies exist in the field of language assessments for young learners, suggesting the need for more research (Bryant, 2017).

This study aims to compare the performance of Grades 1-12 English learners (ELs) on TEIs vs. traditional multiple-choice items in the listening and reading domains of ACCESS for ELLs. ACCESS is a large-scale standardized English language proficiency assessment that is annually administered to over 2 million K-12 ELs in the U.S. The study explores which aspects of the listening and reading constructs TEIs can address. Also, it examines characteristics of TEIs such as item difficulty, time required for completion, and whether students demonstrate a higher use of universal tools (e.g., magnifier, highlighter) that are embedded in the items. Item specifications and students' ACCESS listening and reading test scores are analyzed, along with telemetry data that records students' clicks and keystrokes during test administration. Data are analyzed considering both the students' grade and proficiency levels. Findings provide implications for enhancing the quality of TEIs to better assess ELs' listening and reading proficiency.

The effect of dialogues on the task complexity of narrative writing tasks

Haeun Kim, *Iowa State University*, haeunkim@iastate.edu

Narrative writing tasks are widely used to assess the writing competence of beginning level students or younger learners. Compared to expository writing tasks that require students to make claims and arguments, narrative writing tasks are considered to be a better choice for this test-taker population as they place less cognitive load on the writers (Beauvais et al., 2011). However, little research has been conducted on the inherent complexity of narrative writing tasks and how various differences in complexity of these tasks can affect scores on these tasks. Informed by Robinson's (2001, 2003, 2005, 2011) Cognition Hypothesis, this study investigates the degree to which [± spoken] features in narrative writing prompts may influence task complexity, and how this complexity impacts scores on these tasks. The [± spoken] aspect of the prompt is operationalized by two types of prompts: one that asks

students to continue a story which contains dialogues, and one that requires students to do the same task but gives a story without dialogues in it. The narrative writing tasks will be administered to twenty graduate students who are highly proficient in English and are enrolled in the same course. Participants will complete ten different narrative writing tasks (i.e., five of each prompt type) throughout the semester. Repeated measures MANOVA will be conducted to determine how the writing performances of students differ in terms of lexical competence, syntactic complexity and accuracy, cohesion, and the use of direct speech/dialogues in writing. Initial results from the analysis of the student writings will be presented, suggesting how this spoken feature can influence task complexity. Expressing spoken language in written forms can be a cognitively demanding variable, which can inadvertently increase the task complexity of narrative writing tasks if gone unnoticed by test task developers and instructional materials designers.

Exploring EFL teachers' use of technology for English language assessment

Ananda Muhammad, *Iowa State University*, nanda@iastate.edu
Haeun Kim, *Iowa State University*, haeunkim@iastate.edu
Timothy Kochem, *Iowa State University*, tkochem@iastate.edu

Technology-enhanced language learning continues to be a growing trend within language instruction for most of the world. However, there is still a dearth of opportunities for EFL teachers to learn how to integrate technology into the language classroom effectively (Godwin-Jones, 2015; Kessler & Hubbard, 2017), especially when used as a medium for formative and/or summative assessment. Without appropriate training, language teachers may use technology superfluously; that is, they might use technology for the sake of using technology. Moreover, we are concerned that language teachers view technology for language assessments merely as a means to make the assessment more efficient to implement, neglecting to make adequate justifications for the test usefulness (Chapelle & Douglas, 2006).

The current study is a small-scale exploratory study of a larger ethnographic study conducted on a Global Online Course (GOC), *Using Educational Technology in the English Language Classroom*, offered by the American English e-Teacher program. Participants integrated technologies into instructions for vocabulary, grammar, reading, writing, speaking, and listening. Some participants integrated the technologies into their lesson plans, while others integrated technology into their assessments. For the purposes of this study, we looked only at those participants who integrated technology for assessment purposes. To do so, we employed thematic coding following an ethnography approach using grounded theory (Glaser & Strauss, 1967) to engage in data coding of all assignments, quizzes, and discussion forums. From this coding, several themes emerged, and consequent targeted coding followed based on these emergent themes. Prominent themes that arose from our data were the use of open educational resources (OERs), constraints with language assessment both with and without using technology, and meaningful versus superfluous use of technology for assessment purposes. With this information, teacher training programs could identify problematic areas in the integration of technology for language instruction and assessment.

Test takers' and raters' perceptions of and reactions to pronunciation in a paired-discussion task

Liberato Silva dos Santos, *Iowa State University*, liberato@iastate.edu

Pronunciation assessment has recently gained traction (Isaacs & Trofimovich, 2017), reflecting an understanding that it is a key element of second language (L2) speaking ability. Recent studies have shown that pronunciation is important in determining test takers' speaking proficiency (Ma et al., 2018), but further research is needed to better understand the pronunciation construct and develop rubrics with clear descriptors for pronunciation. The purpose of this work-in-progress study is to investigate test takers' and raters' perceptions of and reactions to test taker pronunciation in a paired-discussion task. A qualitative analysis will be conducted with 20 L2 English test takers and 8 raters who participated in the paired discussion task of a university's English placement test. During individual verbal protocol sessions, test takers will watch video recordings of their own performance during the paired discussion task and will be asked to share their thoughts on their discussion partner's pronunciation and how their perceptions influenced their performance. Raters will also be asked to share their thoughts on test taker pronunciation and how it influenced their rating processes.

To help the participants focus on the construct of pronunciation, they will be asked to use a pronunciation rubric that was developed specifically for this study, following suggestions from Ockey & French (2016) and Ma et al. (2018). The rubric will help participants focus their thought-sharing on seven pronunciation categories: vowels, consonants, word stress, sentence stress, intonation, rhythm, and connected speech. This presentation will share the current version of the rubric and discuss its development process, including the feedback provided by test takers and raters. The findings and reflections obtained from this study may have implications for the development of oral communication assessment instruments that account for pronunciation. They may also have implications for speaking and pronunciation instruction and rater training.

Development of an online version of the MLAT

Charles Stansfield, *Language Learning and Testing Foundation*, cstansfield@lltf.net
Daniel Reed, *Michigan State University*, reeddan@msu.edu
Heekyoung Kim, *Michigan State University*, kimheek@msu.edu

The Modern Language Aptitude Test (MLAT) has been delivered in paper and pencil (PPT) format since 1958. It is widely used by government agencies in English-speaking countries. This poster will discuss the recent completion of a three-year project to adapt the MLAT for on-line presentation (OLT). In a test of mental abilities, such as the MLAT, any change in the task can affect timing, the cognitive abilities required, performance, and attitudes toward the task. We will focus on problems and obstacles encountered in moving to an online format without significantly changing the cognitive demands of the item formats.

In order to increase the probability of a successful adaptation to the OL format, an extensive review process was employed. The process was iterative, and each iteration produced minor changes in approach, directions or layout. Because the MLAT is used in different countries, it was reviewed in different countries. Eventually, reviewers' comments indicated that the test was ready to undergo a comparability study.

During May and June of 2019, 75 students at the Foreign Service Institute took the online MLAT. All had previously taken the PPT version five or more months earlier. The purpose was to determine the comparability of scores on the two versions of the same test form as well as examinee attitudes toward them. After taking the online version, they completed a short questionnaire online in which they indicated which delivery format (PPT or OLT) they liked most, for each of the five parts and for the test as a whole. Examinees were also invited to write an opinion of either delivery format. We compared their part and total scores on the PPT and OLT versions in order to determine if one delivery format produces higher scores than the other.

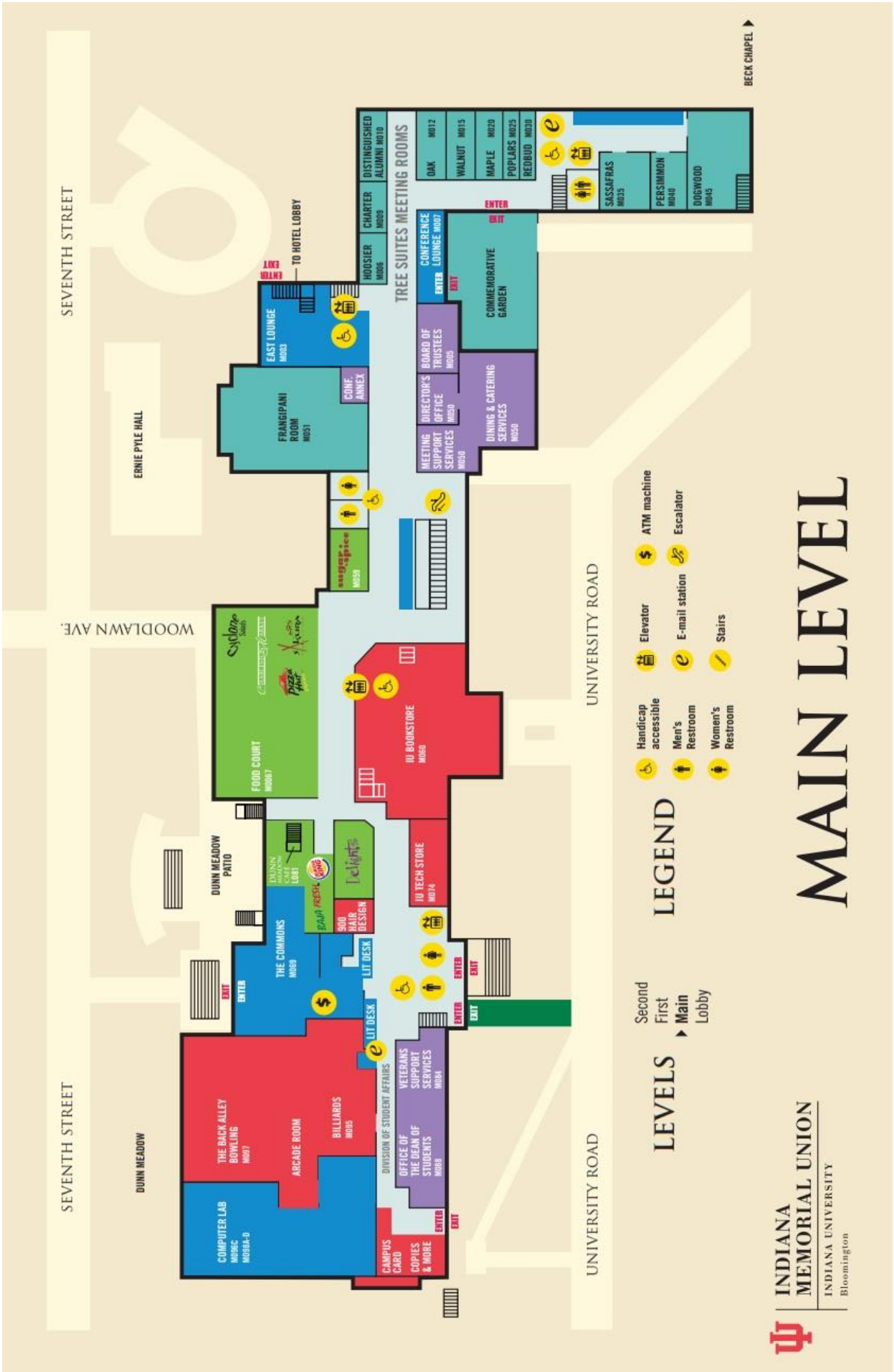
The results of the comparability study will be reported and related to examinee feedback on each part of the test.

Language proficiency literacy (LPL) development of admissions decision makers

Sharareh Taghizadeh Vahed, *Purdue University*, staghiz@purdue.edu

The growing literature on the assessment literacy of educators and language teachers reflects the importance of testing in today's educational settings. However, exploring the language assessment literacy literature quickly bares the truth about the extent to which specific stakeholders were excluded from research studies. A very diverse group of individuals who are frequent users of the most renowned standardized language tests, such as Test of English as a Foreign Language

(TOEFL), are university graduate admissions committee. The task of admitting students to various graduate programs is usually undertaken by the professors in each specific program and research area. However, there is little research about the extent to which this group of stakeholders are aware of the meaning of language test scores, university cut-off scores, and sub-scale scores (O'Loughlin, 2018). Yet, the little research that exists fails to report on any action taken to bridge the gap between this group of stakeholders' lack of language assessment knowledge and the need for professional development in this area so that more informed decisions are made about graduate students academic futures based on their language proficiency scores (Baker 2016).
This poster showcases the researchers' plan for the language proficiency literacy (LPL) development of a large university's graduate admissions committee. After making a case for the importance of LPL development among this group of stakeholders, the researchers discuss how a LPL development plan will lead to professionalism, performativity, and empowerment in the specific context of the study. The researchers adopt the Continuing Professional Development (CPD) framework to explain how the LPL development plan will lead to change through the skills development cycle. After presenting a general overview of the three phases of the research, the poster will be concluded with a detailed plan for LPL development workshops and a LPL development website that can be used by university admissions officers.


MAIN LEVEL
INDIANA MEMORIAL UNION
INDIANA UNIVERSITY
Bloomington
LEVELS
Second
First
Main
Lobby
LEGEND
Handicap accessible
Men's Restroom
Women's Restroom
Elevator
E-mail station
Stairs
ATM machine
Escalator
SEVENTH STREET
UNIVERSITY ROAD
WOODLAWN AVE.
ERNIE PYLE HALL
DUNN MEADOW
DUNN MEADOW PATIO
BECK CHAPEL
TO HOTEL LOBBY
TREE SUITES MEETING ROOMS
DISTINGUISHED ALUMNI M010
CHARTER M009
HOOSIER M008
OAK M012
WALNUT M015
MAPLE M020
POPLARS M025
REDBUD M030
SASSAFRAS M035
PERSIMMON M040
DOGWOOD M045
CONFERENCE LOUNGE M007
COMMEMORATIVE GARDEN
BOARD OF TRUSTEES M005
DIRECTOR'S OFFICE M050
MEETING SUPPORT SERVICES M058
DINING & CATERING SERVICES M030
EAST LOUNGE M003
CONF. ANNEX
FRANGIPANI ROOM M031
FOOD COURT M006/7
IU BOOKSTORE M060
IU TECH STORE M074
THE COMMONS M009
DUNN MEADOW CAFE LOBBY
900 HAIR DESIGN
LIT DESK
LIT DESK
DIVISION OF STUDENT AFFAIRS
VETERANS SUPPORT SERVICES M004
OFFICE OF THE DEAN OF STUDENTS M068
CAMPUS CARD COPIES & MORE
THE BACK ALLEY BOWLING M007
ARCADE ROOM
BILLIARDS M005
COMPUTER LAB M096C M088A-D
ENTER
EXIT

Indiana Memorial Union Map

Index

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™

WIDA is housed within the Wisconsin Center for Education Research at the University of Wisconsin-Madison.

Innovation in assessment

duolingo english test

Designed with AI

Employs machine learning and natural language processing to develop items, deliver, and score the computer adaptive test.

Accessible administration

Delivered online, on demand. Students can take the test from their computer in under 1 hour for $49. Results can be shared with unlimited institutions.

Secure Certification

Robust proctoring process. Remote human proctoring with the help of artificial intelligence certifies the authenticity of each test session.

englishtest.duolingo.com/research