Ready, Steady, Grow!

Measuring Language Development from Pre-K through College and Beyond

MwALT 2018 Midwest Association of Language Testers
Madison, Wisconsin | September 21–22

Congratulations to MwALT 20 from the Assessment and Evaluation Language Resource Center (AELRC) at Georgetown University!

The AELRC provides **research, resources, workshops,** and **online courses** to support foreign language educators' capacity to engage in useful language assessment and program evaluation. Visit us online to learn more!

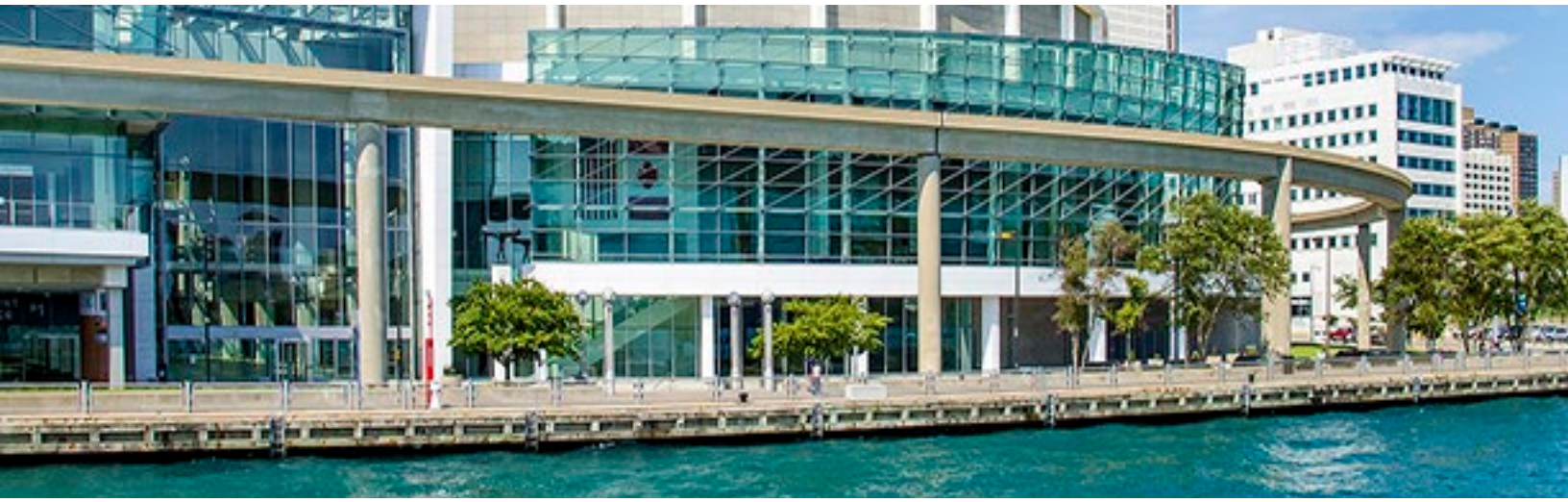
AELRC
Assessment and Evaluation Language Resource Center

aelrc.georgetown.edu

@AELRCDC

/AELRC

The AELRC is a Title VI Language Resource Center funded by the U.S. Department of Education.



WIDA™ Annual Conference

Schools and Communities Empowering One Another

October 23-26, 2018 | Cobo Center | Detriot, Michigan

wida.wisc.edu/conference

Contents

Questions?

Contact mwalt2018@wcer.wisc.edu

Welcome to MwALT 2018!

Welcome to the 20th annual conference of the Midwest Association of Language Testers. WIDA and the University of Wisconsin–Madison are delighted to host the MwALT conference for the very first time here in beautiful Madison, Wisconsin.

This year, we come together for two days on the UW-Madison campus to share research and ideas through workshops, a plenary session, more than ten posters, and more than twenty papers. We very much hope that the ensuing conversations and collaboration will be fruitful and enjoyable. The conference organizing committee also hopes you will find time to explore Madison, visit the lakes, and take a stroll along State Street!

As always, the MwALT conference brings together graduate students, researchers, faculty, and test developers from across the Midwest and beyond. This year, for our 20th gathering, we focus on a specific population and assessment context—the academic language of students in Kindergarten through 12th grade in U.S. public schools. Our conference theme is "Ready, Steady, Grow! Measuring Language Development from Pre-K through College and Beyond."

The academic achievement and language development of this highly diverse population of multilingual learners has been WIDA's mission since 2003. We are excited to provide a forum for research in this important field at this year's MwALT conference.

Mark Chapman

Director of Test Development, WIDA
2018 MwALT Conference Chair

Conference Organizers

Thank you to everyone who participated in making MwALT 2018 a success.

Conference Organizing Committee
Mark Chapman
Heather Elliott
Ahyoung Alicia Kim
David MacGregor
Matthew Freid
Houa Moua
Marcy Olson
Dale Erlandson

Abstract Reviewers

David Crouch
Parva Lazarjani
Kye-Gon Lee
Xiaorui Li
David MacGregor
Mayu Miyamoto
Stephen O'Connell
Kyongson Park
Ji-young Shin
Sharareh Vahed
Fang Wang
Yangting Wang
Xiaowan Zhang

Volunteers

Pauline Ho
Jason Kemp
Kyongson Park
Fang Wang
Xiaowan Zhang
Ji-Young Shin
Xiaorui Li

Sponsors


AELRC
Assessment and Evaluation
Language Resource Center


CAL
CENTER
FOR APPLIED
LINGUISTICS


DATA RECOGNITION
DRC
CORPORATION


LANGUAGE LEARNING
AND TESTING FOUNDATION


UNIVERSITY OF
MICHIGAN
MICHIGAN
LANGUAGE ASSESSMENT


MICHIGAN STATE
UNIVERSITY


WCEPS
WISCONSIN CENTER for
EDUCATION PRODUCTS & SERVICES

Workshops

Workshops will be held at UW-Madison's Educational Sciences Building, 1025 W. Johnson St.

Workshop I: K-12 ELLs and the Development of Academic Spoken Language

Facilitators: Meg Montee, Center for Applied Linguistics & Mark Chapman, WIDA

Friday, September 21, 9:30 a.m.–12:30 p.m., Educational Sciences, Room 259

This workshop will focus on assessing and scoring students' academic spoken language. The workshop will begin with an overview of the ACCESS 2.0 test construct and task design, including the use of model responses on the test that exemplify target response characteristics. Then, participants will review and analyze student responses at a variety of grade and performance levels and discuss approaches to developing a rubric and scoring scale descriptors that reflect differences in performance levels. Finally, participants will review the WIDA ACCESS 2.0 Speaking Scoring Scale and practice applying it to student language.

Workshop II: K-12 ELLs and the Development of Academic Written Language

Facilitators: Jing Wei, Center for Applied Linguistics & Heather Elliott, WIDA

Friday, September 21, 1:30 p.m.–4:30 p.m., Educational Sciences, Room 259

This workshop will focus on assessing and scoring students' academic written language. The workshop will begin with an overview of the ACCESS 2.0 test construct and task design. Then, participants will review and analyze student responses at a variety of grade and performance levels, and discuss approaches to developing a rubric and scoring scale descriptors that reflect differences in performance levels. Finally, participants will review the WIDA ACCESS 2.0 Writing Scoring Scale and practice applying it to student language.

Plenary

Redefining the Construct of English Language Proficiency to Support English Learners in the Content Areas

Lorena Llosa, New York University

Saturday, September 22, 8:45–9:45 a.m., Alumni Lounge, Pyle Center

English learners, the fastest growing population in US schools, face a unique challenge: they have to learn content (e.g., math, science) at the same time as they develop their English language proficiency. Traditionally, English learners' language and content learning needs have been addressed separately both in instruction and in assessment. Over the past few decades, however, there has been growing recognition that language and content overlap in significant and consequential ways, leading to English language proficiency standards and assessments that link language to content areas and content standards and assessments that include language-intensive practices (e.g., argue from evidence).

In this talk, I will argue that, despite increasing efforts to integrate content and language, current definitions of English language proficiency may not be as useful when the goal is to support English learners' content learning. Using science as an example, I will propose an alternate conceptualization of English language proficiency that embraces and leverages the overlap between content and language by focusing attention on the aspects of language that are most critical to communicating disciplinary meanings. I will provide examples of how we can operationalize this language construct in the context of science tasks, illustrate how this approach affords unique opportunities for rich formative assessment practices in the content classroom, and discuss implications for large-scale language assessment.

Lorena Llosa is an Associate Professor of Education in the Steinhardt School of Culture, Education, and Human Development at New York University. Her work addresses the teaching, learning, and assessment of content and language of English learners in schools. Her studies have focused on standards-based classroom assessment of language proficiency, assessment of academic writing, and the integration of language and content in instruction and assessment. She is currently Co-Principal Investigator on two projects funded by the National Science Foundation to develop science instructional materials that support English learners' science learning, computational thinking, and language development.

Conference Schedule

Friday, September 21
The Wisconsin Center for Education Research, 1025 W. Johnson St., Room 259

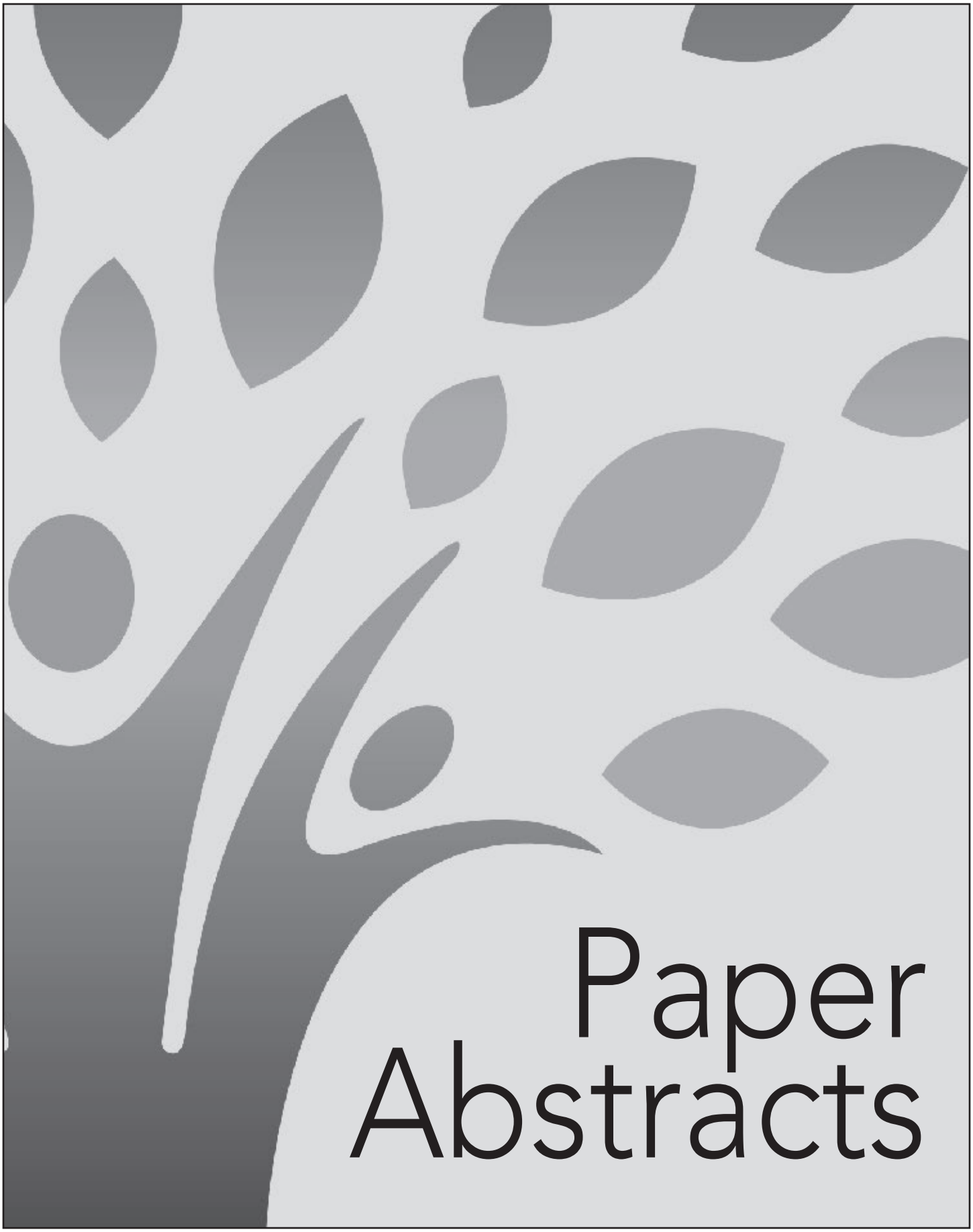
9:30–12:30	Workshop I: Spoken Language
1:30–4:30	Workshop II: Written Language

Saturday, September 22
The Pyle Center, University of Wisconsin-Madison

8:00–12:30	Poster Setup: Lee Lounge		
8:00–12:00	On-site registration: Main Lobby		
8:00–9:00	Breakfast: Alumni Lounge		
8:30–8:45	Opening Remarks: Alumni Lounge		
8:45–9:45	Plenary: Alumni Lounge Lorena Llosa, New York University		
Paper Sessions	Room 209	Room 213	Room 226
9:50–10:20	**Where Are the Values in Academic English Language Tests?** Carol Chapelle *Iowa State University*	**Exploring profiles of Long-term English Learners using large-scale language proficiency assessment data** Narek Sahakyan *WIDA at the Wisconsin Center for Education Research* Sarah Ryan *WIDA at the Wisconsin Center for Education Research*	**Investigating variability and dependability in an ITA speaking test: A generalizability theory study.** Ji-young Shin *Purdue University*
10:20–10:50	**Use of ECD to develop the speaking portion of a university English placement test** Roz Hirch *Iowa State University* Haeyun Jin *Iowa State University* Gary Ockey *Iowa State University*	**Assessing Incoming Kindergarten Dual Language Learners: A Multilingual Approach** Florencia Tolentino *Educational Testing Service* Danielle Guzman-Orth *Educational Testing Service* Alexis Lopez *Educational Testing Service*	**The effects of rating contexts in evaluating ELL children's spoken responses: A case of the decision-making processes of local administrators and trained raters** Shinhye Lee *Michigan State University* Kyoungwon Bishop *WIDA at the University of Wisconsin-Madison* H. Gary Cook *WIDA at the University of Wisconsin-Madison*

10:50–11:00	Midmorning refreshments: Alumni Lounge		
Paper Sessions	Room 209	Room 213	Room 226
11:00–11:30	**Extending the Validity Argument for an In-house ESL Proficiency Test Through Test Score Gains** Lixia Cheng *Purdue University*	**Leveraging Language Performance Assessment Implementation and Outcomes Data for Pedagogical Accountability and Innovation: A K-12 Evaluation Study** Reuben Vyn *University of Iowa*	**Language Instruction in a Distant Classroom—Does distance really matter for oral proficiency gains?** Mary Jo DiBiase-Lubrano *Yale University*
11:30–12:00	**Using an Argument-based Validation Framework to Guide the Refinement of a High-stakes Writing Test for K-12 English Language Learners** Jing Wei *Center for Applied Linguistics* Tanya Bitterman *Center for Applied Linguistics* Ruslana Westerlund *WIDA at the Wisconsin Center for Education Research* Jennifer Norton *Center for Applied Linguistics*	**Examining the Cognitive Dimension of L2 Academic Speaking Ability through a Scenario-Based Assessment Approach: A Pilot Study** Yuna Seong *Teachers College, Columbia University* **Winner of 2018 Best Student Paper Award**	**Measuring the Development of L2 Spanish Accentedness and Comprehensibility** Ziwei Zhou *Iowa State University*
12:00–1:00	Lunch: Alumni Lounge		
12:15–1:30	MwALT Business Meeting: Room 209		
1:00–2:00	Poster Session: Lee Lounge See list of poster presentations on page 28		
Paper Sessions	Room 209	Room 213	Room 226
2:00–2:30	**Investigating Raters' Decision-making Process when Rating Responses to an Integrated Writing Task** Phuong Nguyen *Iowa State University*	**Item Development in Practice: Examples from a K-12 EFL Context** Laura Ballard *Center for Applied Linguistics*	**An Evaluation of Elicited Imitation (EI) Task: An Item Analysis Using Classical Test Theory (CTT)** Xiaorui Li *Purdue University*

Paper Sessions	Room 209	Room 213	Room 226
2:30–3:00	**How Does Rater Performance Change over the Course of Rater Training? Insights Gained from Many-facet Rasch Modeling in Second Language Writing Assessment** Hyunji (Hayley) Park *University of Illinois at Urbana-Champaign*	**Collaboration for ELL Stem Literacy: Beyond Language Development** Kyongson Park *Purdue University*	**Developing a Web-based Academic English Placement Test using Authentic Text** Senyung Lee *Indiana University* Yena Park *Indiana University* Sun-Young Shin *Indiana University*
3:00–3:30	**The Use of Outside Judges to Independently Verify Essay Ratings from 2 CEFR-aligned Exams** Daniel Reed *Michigan State University* Koen Van Gorp *Michigan State University* Aaron Ohlrogge *Michigan State University*	**Assessing K-12 Teachers' Language Assessment Literacy** Michelle Stabler-Havener *Teachers College, Columbia University*	**Using Standardized and Locally Developed Placement Tests Complementarily for Determining the Need for University Support** Sonca Vo *Iowa State University* Shireen Baghestani, *Iowa State University* Gary Ockey *Iowa State University*
3:30–3:45	Afternoon Refreshments: Alumni Lounge		
Paper Sessions	Room 209	Room 213	Room 226
3:45–4:15	**Longitudinal Development of Second Language Fluency in Writing and Speaking** David Crouch *Purdue University*	**Self-assessment in the Foreign Language Classroom: How to Engage Students in the Target Language** Stephanie Gaillard *University of California San Diego* Dorian Dorado *Louisiana State University*	**The Effect of Delivery Mode on L2 Students' Listening Comprehension Skills: Content versus Animated Videos** Leyla Karatay *Iowa State University*
4:15–4:45	**The Effect of Time Pressure on the Use of Linking Adverbials in ESL Writing** Yasin Karatay *Iowa State University*	**Self-assessment: A Feisty or Reliable Tool to Assess the Oral Proficiency of Chinese Leaners?** Wenyue Ma *Michigan State University* Paula Winke *Michigan State University*	**Does the Ability to Comprehend Conversational Implicature Predict Higher-level Language Ability?** Stephen O'Connell *WIDA at University of Wisconsin-Madison*
5:00–5:15	Awards Presentation and Closing Remarks: Alumni Lounge		
5:15–7:00	Closing Reception: Patio (Alumni Lounge if raining)		



Paper Abstracts

Laura Ballard
Center for Applied Linguistics

Item Development in Practice: Examples from a K-12 EFL Context

The purpose of this paper is to describe the test-development process of a large-scale primary- and secondary-school language assessment used in Central America. This test is used in public schools to monitor and evaluate progress in attaining nationwide curricular language goals. I will describe the test specifications and the process for developing items aligned to the Common European Framework of Reference (CEFR). This process includes item-writer training, item writing, item review, graphics development, bias and sensitivity review, field testing, item selection, and operational form creation. I will also highlight real-world challenges in the item-development process, including incorporating cultural considerations in to the item development, creating engaging items for low-proficiency levels, and deciding which items should be included in operational test forms. In the discussion, I will show how this process exemplifies evidence-centered test design (Mislevy, Behrens, Dicerbo, & Levy, 2012) and the importance of situating test design within the larger realm of test validation. I will conclude by showing test results from 900 primary- and secondary-school English Language Learners in Mexico.

Carol Chapelle
Iowa State University

Where Are the Values in Academic English Language Tests?

At the 2018 Language Assessment Research Conference, Tim McNamara's plenary asked what values underlie the resistance of the field to testing English as a lingua franca (ELF). The assumption that the field of language testing resists testing ELF was supported by a 2016 chapter by Jenkins and Leung that states, "tests of ELF do not currently exist." Their paper claims that theory and practice in language testing are not up to the task of developing tests of ELF. McNamara proposed a more insidious explanation stemming from the value driven and politically motivated nature of language testing. He questioned Jenkins and Leung's claim that the field lacks the technical capacity to create tests of ELF, but he seemed to accept the assertion that no tests of ELF exist. Furthermore, he suggested that all English language tests, including tests of academic English, reflect the high value the field places on native-speaker varieties of English.

My paper examines EAP testing practices in higher education in view of all three claims: that no tests of ELF exist, that the field is technically incapable of producing tests of ELF, and that the construct-related values of language testers are responsible for refusal to create tests of ELF. Through an analysis of how oral EAP is tested for admissions and placement at one university in the United States, the paper exposes values with respect to ELF that underlie test development, administration, decision-making, and institutional policy for test use. In doing do, the paper identifies specific areas of EAP testing practice that can be examined as a basis for identifying construct- and policy-related values. I will argue that testers need to able to identify the values implicit in their practices in order to attempt alignment of the values of testing with those of society as well as to engage in academic discussion with other applied linguists, who are not always aware of actual testing practices and how they align with local contexts.

Lixia Cheng
Purdue University

Extending the Validity Argument for an In-house ESL Proficiency Test through Test Score Gains

Longitudinal tracking of ESL test performances not only contributes to program evaluation, but it also extends the validity argument for the language test used. A validity argument is supported when test scores reflect appropriate change as a function of construct-related teaching or learning (Cronbach, 1971; Messick, 1989).

The Assessment of College English – International (ACE-In), a locally developed, computer-mediated, semi-direct English proficiency test has been used as embedded assessment in an English for Academic Purposes (EAP) program at a large public university for purposes such as providing diagnostic information to language teachers, tracking students' language development, and gathering information for program effectiveness. This EAP program provides language and culture support to matriculated international undergraduate students with relatively lower TOEFL iBT or IELTS scores.

This study focused on a machine-scored cloze-elide section and a human-rated elicited imitation section on the ACE-In. Test items in both sections were developed based on well-defined test specifications and pilot-tested before operationalization. Cloze-elide assesses vocabulary, grammar, and silent reading whereas elicited imitation assesses listening comprehension, information retention, and grammatical accuracy in oral production. Both sections have high internal consistency estimates and small standard errors of measurement. The cloze-elide score has a significant, moderate correlation with the TOEFL iBT writing score (r = .42, p = .04) while the elicited imitation section score has a significant, strong correlation with the TOEFL iBT speaking score (r = .74, p < .0001).

Test data gathered at the beginning of a two-semester sequence, at the end of Semester 1, and at the end of Semester 2 indicates that the international students in this EAP program made gains in each semester (Cloze-elide: $\chi 2$ = 217.07, p < .001; Elicited imitation: F = 41.28, p < .001, $\eta\rho 2$ = .40). This evidence of score gains helps to extend the validity argument for ACE-In.

David Crouch
Purdue University

Longitudinal Development of Second Language Fluency in Writing and Speaking

This study analyzed the longitudinal development of oral and written fluency in the test responses of 99 first year L1-Chinese undergraduate students over two semesters. The author collected responses to a post-entry, semi-direct, computer-administered language proficiency test required of all first-year international students with TOEFL scores at 100 or below at a large university in the US. The students took the test at the beginning of the first semester and again at the end of the second semester of a required two course ESL sequence. For both the written and the spoken task, each student responded in support or opposition to a statement of opinion.

The author analyzed the written responses automatically to calculate mean length of sentence and mean length of T-unit using Lu's (2010) L2 Syntactic Complexity Analyzer and the oral responses for speech rate (syllables/second) and mean syllables per run (syllables/run) using a proprietary system specially designed to measure oral fluency.

Results showed that the test-takers increased their oral fluency significantly but not their written fluency. When the mean syllables per run of the oral pre-tests (M=7.11, SD=1.60) was compared to that of the oral post-tests (M=7.66, SD=1.84), there was a statistically significant difference, t(98)=-3.74, p=.00031. In a paired sample t-test the speech rate of the oral pre-tests (M=2.88, SD=0.40) was compared to that of the oral post-tests (M=3.00, SD=0.45), and there was also a significant difference, t(98)=-3.69, t=.00037. A Wilcoxon signed rank test compared the mean length of sentence of the written pre-tests (M=19.86, SD=5.23) to that of the written post-tests (M=20.29, SD=4.77), showing no significant difference, Z(98)=-1.501, p=.133. Another Wilcoxon test compared the mean length of T-unit of the written pre-tests (M=17.05, SD=3.85) to that of the written post-tests (M=17.57, SD=3.79), showing no significant difference, Z(99)=-.949, p=.342. Theoretical and practical implications will be discussed.

Mary Jo DiBiase-Lubrano
Yale University

Language Instruction in a Distant Classroom—Does Distance Really Matter for Oral Proficiency Gains?

Language instruction in virtual classrooms has become more and more popular thanks to the affordances that advances in technology provide. However, research is still scanty as to how student learning outcomes are effectively attained in this environment and if these gains are comparable to those within traditional settings. In particular, how do oral proficiency outcomes vary in virtual classrooms when the learner is in a distant environment and how do these gains compare to those of learners physically present in traditional classrooms? To address this question, a longitudinal research project was conducted within a shared courses initiative established among three institutions which offers grant funded, credit bearing courses in less commonly taught languages. Within this initiative, language instruction is delivered via videoconferencing technology in synchronously connected campuses allowing institutions to expand their language offerings while offsetting the limitations of local resources (Van-Deusen-Scholl and Charitos, 2017).

This study aims to address the issue of oral language proficiency gains for geographically distant learners and how they compare to those who are physically present with the instructor. Oral proficiency outcomes were measured in pre and post ACTFL OPIs for students enrolled in the research project. Additionally, biodata and post courses surveys responses which revealed students' motivations largely relating to linguistically and culturally diverse heritages also provided further scope in the interpretation of oral proficiency gains.

This paper will present the results of the OPI tests triangulated with qualitative data relating to students' motivations and perceptions in a virtual classroom and offer some tentative conclusions as to how proficiency gains may vary when learners are in virtual classrooms compared to those in traditional settings.

Stephanie Gaillard
University of California San Diego

Dorian Dorado
Louisiana State University

Self-assessment in the Foreign Language Classroom: How to Engage Students in the Target Language

The use of the communicative language teaching approach (Celce-Murcia 2007; Ellis 2008; Brown 2015; Shrum & Glisan 2016) bases its principles on encouraging the active engagement of L2 learners in order to facilitate acquisition of the target language (TL), yet getting students to actively engage in the classroom remains a challenge for many teachers at the university level. To address this concern, we explore the use of a self-assessment (Oscarson 1989; Bachman 1990; Fulcher 2010; Shepard 2000; Thompson 2015; De Saint Léger 2009) device – the participation log – in a French language program at a large tier 1 university at the undergraduate levels as a means of enhancing students' language engagement in the classroom.

The participation log consists of a set of criteria which students use at the end of each class period for self-assessment and reflection on their classroom engagement. The instructors then use the same criteria to agree with the students' self-evaluation or readjust the self-rating accordingly. To observe the effects of the participation log on students' engagement and motivation, we administered 709 online surveys for the students to complete anonymously and 20 surveys to the teachers of these courses.

The results demonstrated positive effects for using the participation log in the classroom from both the students' and the teachers' perspectives. More specifically, teachers expressed that students' motivation increased due to the use of the participation log in class. They also specified that it helped them increase their students' enthusiasm and interest in language learning. From the student's viewpoint, they acknowledged that this tool motivated them to speak French in class and were very satisfied with the criteria provided to them because the methodology of the participation log values their participation rather than their accuracy in the TL.

Roz Hirch
Iowa State University

Haeyun Jin
Iowa State University

Gary Ockey
Iowa State University

Use of ECD to Develop the Speaking Portion of a University English Placement Test

The speaking construct is complex, consisting of many different potential factors, including vocabulary, grammar, pragmatics, and pronunciation, among others. Creating an effective speaking assessment that addresses all of these elements while considering the needs of stakeholders is difficult. For this reason, test developers have found evidence centered design (ECD) useful, since this normative model is effective for richer tasks (Mislevy et al., 2006). ECD gives test developers a step-by-step approach that coordinates all elements, often by answering questions such as "What are we measuring?" (proficiency model) and "How do we measure it?" (evidence model) (Almond et al., 2016). In sum, ECD allows test developers to describe "what we want to say about students and what kind of evidence we need to see" (Mislevy et al., 2002, p. 479).

This presentation describes the ECD process used to develop the speaking portion of a university English placement test (EPT) at a Midwestern US university. According to the test specifications, the intended construct – oral communication skills necessary for success in academic contexts – consists of four elements (proficiency model: What are we measuring?): fluency, interactional competence, comprehensibility/pronunciation and grammar/vocabulary. The evidence model, which answers "How do we measure the construct?" comprises two tasks: an individual scripted-interview task and a paired-discussion task. Each task has a slightly different task model (How do we measure the construct?), with the first task being conducted and rated for each of the four elements by one administrator, while two test-takers do the second task together and are rated by two administrators who observe. This presentation will discuss how layers of the ECD framework (Mislevy et al., 2005; Almond et al., 2016) were used to guide the development of the speaking test and will conclude with a discussion of the helpfulness of ECD for this test development project.

Leyla Karatay
Iowa State University

The Effect of Delivery Mode on L2 Students' Listening Comprehension Skills: Content versus Animated Videos

This study investigated the effect of different delivery modes on L2 students' listening comprehension. A mixed-methods explanatory research design was used to identify whether animated videos or content videos facilitated better listening comprehension. Fifty-seven ESL undergraduate/graduate students were recruited from an academic speaking and pronunciation course offered at Iowa State University. The effectiveness of the delivery modes was assessed by the participant's ability to orally summarize the videos. Two trained raters were recruited to rate participants' audio-recorded summaries. A one-way ANOVA was used to test for a significant difference between mean scores of the two groups. Also, students' perceptions of the delivery modes were elicited through a survey. Results showed that students in the animated video group outperformed those in the content video group. In addition, all students in the animated group preferred animated videos. However, one third of the students in the content group reported preferring audio-only listening since they did not find the content videos helpful. Also, the results indicate that the students' performance on the first video about superstitions was did not differ much in both group while it differed on the second video standardized tests. This finding makes implications about how different L1 backgrounds of the students might affect their listening comprehension and oral performance.

Yasin Karatay
Iowa State University

The Effect of Time Pressure on the Use of Linking Adverbials in ESL Writing

This study investigates whether writing under time pressure and without any time pressure affect second language writers' (L2) writing productions. Since the idea of language tests that measure learners' production skills is to elicit test-takers' performances that are expected in target language use (TLU) domain, the study primarily intends to compare the use of linking adverbials (LAs), one type of cohesive devices, in essays produced by second language learners in a placement test and for in-class assignments to elicit any differences between two different time conditions. In order to achieve this, two corpora were used: Iowa State University English Placement Test Corpus of Learner Writing (2017) for timed essays (n=990) and International Corpus of Learner English (ICLE) (Granger et al. 2002) for untimed essays (n=3823). The results indicated that L2 writers tend to use LAs more in all categories under time pressure in testing condition compared to untimed writings. This tendency towards overall use was particularly noticeable for additive and sequential LAs while qualitative data revealed that they occasionally misused these LAs. The findings of this study are suggestive in that test developers should take writing under time pressure into account when they interpret test takers' writing performances. Considering that both corpora consist of culturally and linguistically diverse students (i.e., mostly Asian students in timed corpus and mostly European students in untimed corpus), this study also makes implications about how this variety affects the use of LAs in both timed conditions.

Senyung Lee
Indiana University

Yena Park
Indiana University

Sun-Young Shin
Indiana University

Developing a Web-based Academic English Placement Test using Authentic Text

This presentation reports the design and development process of a new web-based Academic English Test, an ESL placement test at a large university in the U.S. The purpose of the test is to identify the incoming international students who need English language support for undergraduate or postgraduate studies and to place them into appropriate language courses. The faculty of the university has been reporting inconsistencies between international students' scores on standardized academic English proficiency tests and their actual English language skills in academic settings. In order to address this issue, we made three substantial changes from the earlier test, with regard to the listening section, the writing section, and the medium of delivery. The previously scripted, audio-only listening test has been replaced with an audio-visual academic listening test using authentic lectures video-recorded from undergraduate classes at the university. The previously independent, single-task writing test has been replaced with a combination of an independent narrative writing task and a read-to-write integrated argumentative task. Both listening and writing tests have been changed from paper-based to web-based tests, where not only test-taking but also scoring is entirely done on computers. The test development procedure thoroughly followed the principles of designing language tests (Bachman 1990; Carr, 2011; Davidson & Lynch, 2002), focusing on the authenticity of the test (Lewkowicz, 2000) and incorporating the input from the stakeholders (Bachman & Palmer, 2010). Feedback from 219 examinees and 14 instructors supports that the new test measures the target constructs as intended. We argue that the new test using authentic text has greater construct representation, greater practicality of scoring, and less measurement errors than the earlier test. We will also discuss the benefits and challenges of using authentic lecture videos and authentic written text for developing a large-scale language test.

References

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford University Press.
Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice.* Oxford University Press.
Carr, N. T. (2011). *Designing and analyzing language tests.* Oxford University Press.
Davidson, F., & Lynch, B. K. (2002). *2002: Testcraft: a teacher's guide to writing and using language test specifications.* New Haven, CT: Yale University Press.
Lewkowicz, J. A. (2000). Authenticity in language testing: some outstanding questions. *Language Testing, 17,* 43-64.

Shinhye Lee
Michigan State University

Kyoungwon Bishop
WIDA at the Wisconsin Center for Education Research

H. Gary Cook
WIDA at the Wisconsin Center for Education Research

The Effects of Rating Contexts in Evaluating ELL Children's Spoken Responses: A Case of the Decision-making Processes of Local Administrators and Trained Raters

In this study, we explore whether a particular rater group exhibits behavioral variance when situated in a specific rating context of the spoken performance of ELL children. Particularly, we compare the decision-making processes demonstrated by local administrators (primarily current ELL educators) in an immediate, face-to-face testing context versus by external, trained raters engaged in post-hoc rating of audio-recorded spoken responses. We situate the study in the unique assessment contexts of the speaking domain of ACCESS for ELLs 2.0 (henceforth ACCESS Speaking) in which the local administrators simultaneously serve the role of scorers as the ELL children speak their responses to them. While test administration is contingent upon extensive training, it is of empirical interest whether the multiple roles taken up by the local administrators as well as the immediacy in scoring bring about observable differences in the evaluation process compared to centrally-trained raters. The question we raise is: Do contextual differences in rating as well as the agency of raters factor in in how they arrive at a particular judgement? As part of a larger ongoing study, we report the interview responses of 23 local administrators and 3 centrally trained scorers, who each rated 150 speech samples produced in the face-to-face ACCESS Speaking. After the rating session, each scorer participated in a post-hoc interview in which they were asked to (1) provide descriptions on how they assigned a particular score; and (2) elaborate the advantages or disadvantages they had found about the given rating contexts. Our thematic analysis in NVivo implied how each group of scorers oriented to the here-and-now aspects of rating settings they were engaged in for making their final judgments. We further address the need to better ascertain both cognitive and environmental aspects involved in the rating processes in the context of child language assessment (Lim, 2011).

Xiaorui Li
Purdue University

An Evaluation of Elicited Imitation (EI) Task: An Item Analysis Using Classical Test Theory (CTT)

Elicited imitation (EI) has been widely used to assess second language (L2) proficiency. In an EI task, examinees are provided with a series of sentence stimuli and are expected to repeat the sentences as accurately as possible (Larsen-Freeman, 1991). EI task design has been frequently investigated because the appropriate design of the task is essential to the validity of the instrument (Erlam, 2006). The present study analyzes the performance on the EI task of the Assessment of College English-International (ACE-In) by performing thorough item analysis on pre-test and post-test exams using Classical Test Theory (CTT). One hundred undergraduate students who speak English as a Second Language took a pre-test at the beginning of the semester and a post-test at the end of the semester. The present EI task has four forms, and each form comprises 12 items/sentence stimuli. The item analysis results indicate that all four forms have shown similar item difficulty levels. Although a few items in each form have close item difficulty, the overall item difficulty and the spread of item difficulty are suitable for the current assessment purpose. When comparing pre-test and post-test exams, the item difficulty of most items has decreased. As for item discrimination, a few items that have low discrimination index may need modification or deletion, but the majority of the items have discrimination index above 0.6. Item discrimination index has mostly remained stable in pre-test and post-test. This study offers insights into the features of well-designed EI items and suggestions of future modification of the current EI task. In addition, the comparison between pre-test and post-test item analysis provides information regarding items that are suitable for different stages of assessment.

Wenyue Ma
Michigan State University

Paula Winke
Michigan State University

Self-assessment: A Feisty or Reliable Tool to Assess the Oral Proficiency of Chinese Leaners?

In this study, we took a close look at the results of oral proficiency self-assessment tests and OPIc (Oral Proficiency Interview-computer) tests taken twice by the same group of students. We did this to explore the role of self-assessment in Chinese language programs. The data were collected as part of a Language Flagship Proficiency Assessment Project at a large mid-western U.S. university. In this study, we use data from 80 college students who were taking Chinese. During the spring of two subsequent years, the students took a self-assessment (with NCSSFL-ACTFL Can-do Statements, 2015) as part of the project, and then immediately took an official ACTFL OPIc with a level of difficulty that was matched to their self-assessment outcome. We analyzed the self-assessment results on both the test and item level. In general, we investigated whether self-assessment can reliably indicate students' language gains over time, with the benchmark of true gain being (in this study) their OPIc scores. The findings revealed that most students' language trajectories were reflected by the results of the self-assessment. In addition, the accuracy rate of self-assessment was positively correlated with students' proficiency levels. After a close examination of the items that were misidentified by the students regarding the difficulty level, students tended to under-assess rather than over-assess their oral proficiency. The comparison of the scores of repeated self-assessments and OPIc tests showed that there was no significant difference in how accurately students could self-assess themselves before and after an academic year in a language program.

Phuong Nguyen
Iowa State University

Investigating Raters' Decision-making Process when Rating Responses to an Integrated Writing Task

The process of rating examinees' responses has been considered a critical area of research in performance assessment, yielding practical implications for language test developers and administrators (Fulcher, 2003; McNamara, 1996; Weigle, 2002) in terms of rater training, rating scale development, and rating procedures. Studies on raters' behaviors in second language (L2) writing in the past decades have, to some extent, illuminated what goes on in the rater's mind when they assess test takers' written responses. However, it remains unclear what raters do when they rate responses to source-based writing tasks and how they react to test takers' integration of information from the stimuli into their responses.

Therefore, the goal of this study was to investigate raters' cognitive processes when assessing responses to an English read-to-write task. The research questions this study aimed to address include:

1. What aspects of examinees' responses did raters report attending to?
2. Did experienced and novice raters differ in their attention to aspects of examinees' responses?
3. How did raters attend to the rubric categories?
4. Did experienced and novice raters differ in their attention to the rubric categories?

Stimulated recalls and eye-tracking data were collected from five raters who were asked to rate 10 essays on the computer with an eye-tracker and to verbalize their thinking processes after finishing rating each essay. Transcriptions of the verbal reports were analyzed qualitatively to address the first two research questions. To answer the last two research questions, eye-tracking data for the raters were analyzed quantitatively and compared. The analyses showed that raters paid more attention to organization, arguments, and language use compared to conventions and that the strategies used by raters were relatively idiosyncratic. Findings from the study could contribute to the knowledge about rater variability and carry implications pertaining to rater training, rating scale development, and task selection for L2 writing tests.

Stephen O'Connell
WIDA at the Wisconsin Center for Education Research

Does the Ability to Comprehend Conversational Implicature Predict Higher-level Language Ability?

Understanding conversational implicature is widely believed to be an important part of second-language (L2) listening ability (Buck, 2001; Rost, 2011), and is included in most communicative competence models (e.g., Bachman & Palmer, 1996; Canale & Swain, 1980). The ability to understand conversational implicature is also often linked to the upper levels of language proficiency frameworks (e.g., the CEFR, ACTFL), but despite a fairly substantial literature looking at whether implicature and inferencing increases difficulty for listening test items (e.g., Freedle & Kostin, 1996; Kostin, 2004; Rupp, Garcia, & Jamieson, 2001), there is less research that has targeted this question directly.

To further investigate the subskill separability of implicature from general listening and the topic of differential difficulty by subskill, a 60-item listening test was created. Specifications for all items were identical except for the subskill classification of implicature (k=30) and general listening (k=30). The items were developed following standardized testing best practices (i.e., three independent content reviews) and were administered in two formats (multiple choice and constructed response) to 255 learners of English whose proficiency was in the low intermediate to lower advanced range (i.e., B1–C1 on the CEFR). After summarizing the overall listening test results by subskill and format (which were scored using Rasch analyses), this talk will focus on logistic regression analyses of a subset of the participants (n=84) whose proficiency was confirmed by an independent CEFR-linked assessment, the Michigan English Test (MET).

The logistic regression analyses showed that general listening ability was a better predictor of membership in the CEFR C1 category than implicature listening ability was. Discussion of these results will center on how separable listening for implicature is from general listening and the implications for how conversational implicature can be assessed with L2 learners, including the role of degree and type of implicature.

Hyunji (Hayley) Park
University of Illinois at Urbana-Champaign

How Does Rater Performance Change over the Course of Rater Training? Insights Gained from Many-facet Rasch Modeling in Second language Writing Assessment

In language testing, rater studies tend to examine rating performance at a single time point. Research on the effect of rater training either adopts a pre- and post-training design or compares rating performance of novice and experienced raters. While those studies provide insights into the end results of rater training, little has been known about how rater performance changes during the process of rater training. This study examined how rater performance develops during a semester-long rater training and certification program for a post-admission English as a second language (ESL) writing placement test at a large US university.

The certification program aims to align raters to a newly developed rating scale that provides both placement recommendations and diagnostic information regarding students' writing proficiency level and skill profile. The training process employed an iterative, three-stage approach, consisting of face-to-face group meetings, individual rating exercises, and scale re-calibration based on rater performance and feedback. Using many-facet Rasch modeling (Linacre, 1989, 2006), we analyzed rating quality of 17 novice raters across four rounds of rating exercises. Rating quality was operationalized in terms of rater severity and consistency, rater consensus, and raters' use of rating scale. These measurement estimates of rater reliability were compared across time and between certified and uncertified raters.

At the start of the training program, all raters were inconsistent, varied widely in severity, and achieved low exact score agreement. Over time, certified raters improved on multiple indices of rating quality and became more indistinguishable from one another in the application of the rating scale. However, rater performance did not improve in a linear fashion but instead followed a U-shaped developmental pattern. In contrast, uncertified raters' performance remained inconsistent across rounds. Findings of this study have implications for the effectiveness of rater training and developmental patterns of rating behavior over time.

Kyongson Park
Purdue University

Collaboration for ELL Stem Literacy: Beyond Language Development

English language learners (ELL) students have a double burden or responsibilities as they need to learn content such as STEM as well as language. The purpose of this study is to design and deliver effective STEM (science, technology, engineering, and mathematics) integrated lessons for 7th and 8th grades in a public school. Both domestic and English language learners are included in this study in their regular classroom settings. The researcher observed the teachers of science and ELL classes, co-designed a new lesson plans, co-taught some lessons in the classes to all students in the classes for ten weeks. At the end of the semester, the subjects (n=140) participated in a short quiz and survey as a formative assessment. The main ESL and science teachers focused on providing both high challenges and high supports classroom for ELL students to develop their STEM literacy. Based on the findings, the results indicate that students valued linguistic, sensory, and interactive supports and group work based on teachers' clearly directed lessons. This research suggests that the collaboration between content and ELL teachers be recommended for ELL's academic as well as language development. This research can contribute to teacher education in global educational era and provide the plausible ways of cooperation between content and language teachers.

Daniel Reed
Michigan State University

Koen Van Gorp
Michigan State University

Aaron Ohlrogge
Michigan State University

The Use of Outside Judges to Independently Verify Essay Ratings from 2 CEFR-aligned Exams

Some of the most obvious backing data needed for exam validation is either never collected or not given the consideration merited. For example, benchmark essays are often generated in-house based on strict agreement criteria but with no external verification. Further validation typically focusses on the accurate application of the rubric and sets of sample essays in the grading of other essays. Given the critical uses of benchmarks, test developers should empirically confirm agreement on their rank order and independently verify the levels that benchmark samples are supposed to represent in a larger framework. This presentation includes one such small-scale, external validation study of benchmarks as well as a second study on another rating issue, each respectively associated with a high-stakes B2 and C2 exam developed in the U.S. and administered in Europe. Two groups of 9-10 testing experts familiar with the CEFR and from diverse testing bodies across Europe participated.

The first study focused on a set of benchmarks intended for rater training and public examples. Results of this work confirmed the rank-order within calibrated sets with 100% accuracy but suggested that better exemplification of one of the performance levels was needed. The second study examined essays from a small group of students who received very high marks on the exam overall but received unusually low essay scores. This was part of a wider look into whether skill-level discrepancies were as pronounced as was suggested by skill-level profiles for this population, which had been a point of skepticism among stakeholders. Results indicated that the skill-level discrepancies were indeed real and, therefore, that candidates with similar profiles should focus on writing skills in their future English studies. Both studies confirmed the value of involving independent testing experts through small-scale studies as part of the ongoing validation process of a large-scale exam.

Narek Sahakyan
WIDA at the Wisconsin Center for Education Research

Sarah Ryan
WIDA at the Wisconsin Center for Education Research

Exploring Profiles of Long-term English Learners using Large-scale Language Proficiency Assessment Data

Despite the sizable and growing presence of English Learners (ELs) in K-12 schools, these students continue to experience uneven access to high-quality educational opportunities and significantly poorer educational outcomes than their English-fluent and never-EL peers. Educational outcomes are especially troubling for students who remain classified with an EL status for an extended period of time. In the literature and policy documents, these students are increasingly referred to as long-term English learner students, or LTELs. Evidence suggests that as many as one quarter to one half of ELs will remain classified as such for six or more years making so–called LTELs a group with growing national significance.

In this paper, we use longitudinal analyses methods and large-scale assessment data from WIDA Consortium states that administer the ACCESS for ELLs test as a part of their accountability systems, and explore the relationships between students' language growth trajectories and educational practices, systems and policies. Our findings shed light on some of the potential mechanisms, including students' initial English proficiency, states' reclassification criteria and student mobility, which affect the LTEL population in different ways across states. We find substantial heterogeneity across states in the size and characteristics of the population of students referred to as LTELs, suggesting a substantial role of local policies and other contextual factors in shaping this student subgroup. We identify a pressing need for research that rejects an overly simplistic understanding of the long-term English learner designation as something located within students, and focuses on how educational and social systems, practices, and policies structure the unique educational experiences and diverse growth trajectories of all ELs, including students identified as LTELs.

Yuna Seong
Teachers College, Columbia University

Examining the Cognitive Dimension of L2 Academic Speaking Ability through a Scenario-Based Assessment Approach: A Pilot Study

Winner of 2018 Best Student Paper Award

In the field of second language (L2) assessment, the conceptualization of L2 proficiency has evolved and broadened over the past few decades (Purpura, 2016). Accordingly, there has also been a shift in how L2 speaking ability is defined and measured. Alternative approaches, such as scenario-based assessment (e.g., Sabatini & O'Reilly, 2013) are being sought to measure language proficiency and factors that have traditionally not been assessed. In the academic domain, L2 speaking ability calls for students' oral communication of disciplinary knowledge and their integrated use of (meta)cognitive strategies to process, synthesize, and utilize the information to perform complex tasks. Therefore, assessing the L2 learner's ability to use relevant strategies is an integral part of measuring L2 academic speaking ability. The purpose of the study was to examine the cognitive dimension of L2 academic speaking ability, specifically the L2 learner's ability to use cognitive strategies, while performing on an online scenario-based academic English speaking test (SBAEST). In addition to speaking tasks, the test included strategy tasks designed to specifically elicit the test taker's use of cognitive strategies. Thirty-two participants took the test, and the results were analyzed using many-facet Rasch measurement. Relationship between the performance on strategy tasks and speaking tasks was analyzed using correlational analyses. Results indicated that strategy use is indeed an integral component of the academic speaking ability construct. Assessments such as the SBAEST may lead to better evaluation of student learning needs, specifically with respect to distinguishing between linguistic and cognitive aspects of speaking ability.

Ji-young Shin
Purdue University

Investigating Variability and Dependability in an ITA Speaking Test: A Generalizability Theory Study.

The present study investigates the sources of variance in different types of tasks on an oral English proficiency test for international teaching assistants (ITAs), and the reliability of the sources. Identifying the structure of the various sources of variance on a speaking test is important since it provides critical evidence on construct-related validity (Cronbach, Rajaratnam, and Gleser, 1963). Using generalizability theory (G-theory), previous studies examined the sources that account for variabilities in general English proficiency tests for second language speakers (In'nami & Koizumi, 2016; Lee, 2006). However, few studies explored English tests for ITAs for the interplay that test takers, raters, and tasks creates. Thus, the current study addresses this issue with two questions: 1) How much do three different categories of sources (person-, task-, and rater-related factors) influence three different scores on an oral English proficiency test for ITAs (item scores, rater's overall scores, and final overall scores) 2) How reliably rated are three different scores?

The data were gathered from a computer-based, oral English proficiency screening test that was administered to 789 ITAs over 5 semesters at a large public university in the US. The test includes both 3 independent tasks and 5 integrated tasks with additional input sources (reading passage, graph, conversation, and lecture). The collected data include individual rater's rating (item scores, overall scores, and final scores), meta-data about test takers (college, first-language, sex) and raters (first language, confidence level on rating, experience). First, using G-theory, variance decomposition was conducted and then reliability of each component was calculated using D-study. Findings indicated that the largest sources are test takers' performance, confirming previous literature (Kunnan, 1992) while interestingly, the interaction between individual rater, test types, and test taker's first language was responsible for 12 percent of variance. The result will be interpreted with implication for differences between Hindi and Chinese, and engineering and non-engineering students.

Michelle Stabler-Havener
Teachers College, Columbia University

Assessing K-12 Teachers' Language Assessment Literacy

Assessment literacy (AL) empowers teachers (Grabowski & Dakin, 2014) by helping them make better decisions about the development, administration, and use of assessments (Harding & Kremmel, 2016). This is crucial because when teachers make erroneous interpretations, which lead to incorrect decisions, students can suffer unintended, negative consequences (Purpura, 2016). As essential as AL is to providing quality teaching and learning opportunities, assessment education remains inadequate (Lam, 2015; Vogt & Tsagari, 2014). According to Cheng (2001), teachers spend up to a third of their time in assessment related activities; however, most "do so with little or no professional training" (Bachman, 2000, p. 19-20). Therefore, if school administrators believe "student assessment is an essential part of teaching and…good teaching cannot exist without good student assessment" (AFT, NCME, & NEA, 1990, p. 2), how do they assess whether pre- and in-service teachers possess the knowledge and skills to effectively execute assessment-related responsibilities?

To answer this question, K-12 teacher certification tests and performance assessments were examined to determine to what degree they measure teachers' language assessment literacy (LAL). Research was conducted by reviewing practice test materials and literature published by test developers on the tests/performance assessment under investigation. Some tests reviewed included the edTPA, Praxis ESOL teacher tests, the National Board [teacher] Certification process, and Cambridge University Press's Teaching Knowledge Test. Study findings indicated that language teacher certification tests and performance assessments purported to measure several language assessment competencies, but due to the limited number test items and restricted scope of performance assessments, it was unlikely that all competencies tests/assessments claimed to measure were assessed. Even if all competencies were assessed, it appeared unlikely that the competencies were measured in much depth. Of the assessments examined, the National Board Certification process appeared to provide the most complete measure of teachers' LAL levels.

Florencia Tolentino
Educational Testing Service

Danielle Guzman-Orth
Educational Testing Service

Alexis Lopez
Educational Testing Service

Assessing Incoming Kindergarten Dual Language Learners: A Multilingual Approach

Young dual language learners (DLLs) represent one of the fastest growing populations in the United States. An enhanced understanding of the construct of multilingualism as an integrated system may enable teachers to better understand these students' diverse needs. Moreover, it is critical that language assessments result in a valid representation of what test takers know and are able to do. Thus, it is important to develop multilingual assessments that allow test takers to use dynamic and fluid language practices within a single assessment context so they can select languages and multisemiotic features from their repertoire in ways that fit their communicative needs (García & Li, 2014; Otheguy et al., 2015; Shohamy, 2011). Translanguaging, or using one's entire linguistic repertoire to communicate meaning, may enable learners to effectively demonstrate what they know and can do. This presentation uses a case study approach with data from a cognitive lab study in which kindergarten English-Spanish DLLs were presented with a dual language task that allows students to translanguage (DLT-TL). For this study, we focused on two students who scored very low (0) on the WIDA Access Placement Test (WAPT). These two DLLs completed computer-mediated tasks designed to assess both receptive and expressive skills for a narrative retell task. A trained, bilingual examiner presented the task and illustrations on the computer to the student and followed a standardized interview protocol. Results showed that students who scored 0 on their WAPT were fully capable of comprehending and producing meaningful language when allowed to use their entire linguistic repertoire. Sample responses will be shared and contextualized with background data about the linguistic environments the DLLs are exposed to at school and at home. Implications for designing DLT-TLs and the potential to address some challenges associated with assessing incoming kindergarten DLLs will be discussed.

Sonca Vo
Iowa State University

Shireen Baghestani,
Iowa State University

Gary Ockey
Iowa State University

Using Standardized and Locally Developed Placement Tests Complementarily for Determining the Need for University Support

Various approaches are used to help determine the need for language support courses after an L2 English user is admitted to a university. Some programs use standardized test scores, which are available from the admissions process while others use locally developed placement tests, which are aligned with the local language needs and ESL course curricula. The latter option may be more defensible, but sufficient resources are often not available. A third approach is to use standardized tests for initial screening followed by locally developed placements tests for those students who need to be further evaluated. This study used scores on the oral communication portion of the English placement test (EPT OC) at a US Midwestern university to determine what level of TOEFL iBT scores students would need to almost certainly pass the locally developed placement test. The study included 332 students who joined the university during the academic year 2017-2018. Logistic regression was used to establish a cut-score on the TOEFL iBT speaking section, which could be used to exempt matriculated students from taking the locally developed test to determine if they needed to take an ESL oral communication support course. The dependent variable was dichotomous: pass/fail of the EPT OC. The independent variable was TOEFL iBT scores. The preliminary analysis showed that a defensible cut-score of 22 out of 30 on the TOEFL iBT speaking test should be used for decision-making. Students who had TOEFL iBT scores of 22 or above would be very likely to pass the locally developed placement test and having these students take the local test may be a waste of resources. The implications of the study are that standardized test scores can be used to aid in placement decisions, but should be used in conjunction with locally developed placement tests. How to use standardized test scores should be based on empirical data in relation to the locally developed test.

Reuben Vyn
University of Iowa

Leveraging Language Performance Assessment Implementation and Outcomes Data for Pedagogical Accountability and Innovation: A K-12 Evaluation Study

In the wake of the accountability era, foreign language programs are under increasing pressure to provide evidence of student learning outcomes, often in the form of large-scale performance assessments (Malone & Sandrock, 2016). However, little is known about how the introduction of such assessments and the outcomes data they produce influence language teaching and learning, particularly in the K-12 context (Donato & Tucker, 2010). To fill this gap, this mixed methods evaluation study measures the extent to which foreign language students in one urban K-12 school district are able to achieve the program's stated goal of Intermediate levels of proficiency. Additionally, it investigates the potential washback effect of the American Council on the Teaching of Foreign Languages' (ACTFL) Assessment of Performance toward Proficiency in Languages (AAPPL) on teachers' instructional and assessment-related practices. Preliminary findings from an online survey (n=47) and semi-structured interviews (n=10) suggest that teachers do intentionally modify their curricula and assessments to align with the content and format of the AAPPL. Furthermore, teachers who elected to administer the AAPPL also consistently reported increasing their use of the ACTFL Proficiency Guidelines, Can-Do Statements, and World Readiness Standards. However, several teachers also reported limitations to the washback effect including the lack of immediate results for real-time student feedback and changes to instruction, as well as a need for professional development focusing on assessment literacy. These initial qualitative findings will later serve to compliment or problematize students' AAPPL outcomes data, which will be analyzed for potential differences within the district and also compared to national averages. The synthesis of these multiple data sources will lead to a more rich and accurate description of the program's overall performance, as well as practical recommendations for how its stakeholders may improve curricula and instruction, ultimately leading to increased student proficiency development over time.

Jing Wei
Center for Applied Linguistics

Tanya Bitterman
Center for Applied Linguistics

Ruslana Westerlund
WIDA at the Wisconsin Center for Education Research

Jennifer Norton
Center for Applied Linguistics

Using an Argument-based Validation Framework to Guide the Refinement of a High-stakes Writing Test for K-12 English Language Learners

Argument-based validation frameworks have been used extensively to guide test development (Li, 2016). However, most validation work has been conducted as a one-time endeavor in the initial test development stage rather than as a continuous evaluation process after the test has been put in use (Cummins, 2010).

This study fills in the gap by illustrating how an argument-based validation framework—the CAL validation framework—can be used to guide the refinement of the WIDA ACCESS 2.0 Writing Test, a high-stakes test designed to assess the academic English development of K–12 English Language Learners (ELLs) in the United States (Kenyon, 2018). This paper focuses on domain analysis and domain modeling phases of writing test refinement. Our research questions are: (1) What language demands do ELLs encounter in content area classes in writing? (2) How do language demands in writing vary by grade level and content area? We conducted a multi-phase mixed-methods study to investigate language demands in the target language use (TLU) domain. In Phase I, we reviewed conceptualizations of academic language development by national standards and literature. We decided to take a genre-based approach in our review, because the WIDA standards that serve as the basis of the test development are informed by a functional view of language in context (Halliday & Hasan, 1985). Next, we collected survey data from educators on the relative importance of each of the genres generated from Phase I. In Phase III, we conducted focus group interviews with educators, discussing genres, task topics and task characteristics specific to each grade. Finally, we triangulated data from all three phases to establish task features that are consistent with the TLU domain and therefore support to appraise the claims about the validity of inferences drawn about ELLs' academic language development. This study sheds light on the trajectory of academic written language expectations in grades K–12 and illustrates the benefits of applying an argument-based validation framework to navigate the complex process of test refinement.

Ziwei Zhou
Iowa State University

Measuring the Development of L2 Spanish Accentedness and Comprehensibility

When it comes to evaluating L2 pronunciation, it is not unusual for "many educators, researchers, policy makers, and members of the general public [to] equate non-native speakers' accents with their ability to communicate effectively" (Trofimovich & Isaacs, 2012, p. 905). However, studies have consistently shown that non-native accent does not necessarily impede successful communication (e.g. Rubin, 1992; Derwing & Munro, 2009; Kang & Rubin, 2009). In other words, as key construct components in L2 pronunciation assessment, accentedness and comprehensibility, while related, are *partially independent* dimensions (Derwing & Munro, 2005a, 2005b; Thomson, 2018). While many studies have demonstrated that rating scores measuring these construct components can be predicted by a variety of speech features (e.g. Kang, 2010; Kang, Rubin, & Pickering, 2010; Isaacs & Trofimovich, 2012; O'Brien, 2014; Saito, Trofimovich, & Isaacs, 2015, 2016), few have investigated the measurement of L2 Spanish accentedness and comprehensibility across time, as well as the relationship between automatically extracted acoustic features and the rating scores. Using a sample of 20 native speakers of Spanish, who rated both accentedness and comprehensibility on 9-point Likert scales over 31 sentences spoken by 26 English speaking learners of Spanish across five sessions in an academic year, this study first shows substantial variance due to rater severity/leniency and that the 9-point falls short in representing interval-scale measurement. Then, this study adopts mixed-effects ordinal logistic regression to map automatically extracted features (including segmental features from Google ASR engine and suprasegmental features from Prosogram) and ratings on accentedness and comprehensibility across time. Results indicate that both accentedness and comprehensibility improved over time, while the former was consistently rated higher than the latter (e.g. speech is accented while understandable). Moreover, accentedness and comprehensibility ratings share a common subset of acoustic correlates – speech time, speech rate, pitch, and rhythm metric – in comparable magnitude and direction.

Poster Abstracts

Laurene Christensen
WIDA at the Wisconsin Center for Education Research

Vitaliy Shyyan
Smarter Balanced Assessment Consortium

James Mitchell
WIDA at the Wisconsin Center for Education Research

Characteristics of English Learners with Significant Cognitive Disabilities

Little is known about English language learners with the most significant cognitive disabilities as defined by the Individuals with Disabilities Education Act (IDEA) 2004. This paper aims to fill that gap by presenting results from the Individual Characteristics Questionnaire (ICQ), developed under an enhanced assessment grant from the U.S. Department of Education. The purpose of the ICQ is to gather key information about the characteristics of ELs with significant cognitive disabilities in K-12 settings in to create a foundational knowledge base of this population of students. The ICQ collects information related to languages across multiple settings, disabilities, communication preferences including augmentative and alternative communication (AAC) systems, participation and performance on alternate assessment in ELP, English language arts, math, and science, and receptive and expressive communication and engagement in English and/or language other than English. The ICQ was distributed to teachers across the U.S., who completed the survey on behalf of their students. Over 600 responses were collected. Data is currently being collected and analyzed using a grounded theory approach.

This session will emphasize the importance of knowing the characteristics of the students being tested. Preliminary findings indicate that many students have intellectual disabilities (mild, moderate, and profound) as well as autism. Findings also indicate that the majority of students have Spanish a home language, but a number of students are exposed to a variety of other languages such as American Sign Language, Arabic, and Somali. Furthermore, many students who were given an alternate ELP assessment scored low. These students primarily used extended time, read aloud, and directions repeated as testing accommodations. Findings also indicate that almost half of these students do not have an English language acquisition specialist on their individual education plan (IEP) team. Such findings give insight to the development of a future alternate ELP assessment and implications for accountability.

Jie Gao
Purdue University

Linguistic Profile Analysis of Higher-Level English Speakers - Chinese and Indian Examinees from Oral English Proficiency Test as the Case

While rating an oral language proficiency test, a broad spectrum of factors influencing examinee's final score in combination. Although research shows that raters are achieving better agreement results when speakers of higher proficiency are involved (Yan, 2014), an increase in language sophistication may still lead to diversified profile characteristics with greater delicacy. Accurate description of this information would benefit from more detailed explication of the construct of fluency, along with the co-functioning of lexical complexity factors.

This profiling study is conducted based on a local oral English proficiency test of a public university, which is designed for prospective international graduate teaching assistants with screening purposes. The test has L1 Hindi speakers and L1 Mandarin speakers as the two largest groups of examinees. On a five-level scale from 35 to 55, test responses scored 50 are in possession of proficiency issues that are challenging to problematize explicitly. With 45 being the cutting score for teaching assistantship eligibility, the score of 50 is sometimes dangling on the borderline and creating rater disagreement. For examinees rated as 50 from L1 Hindi or Mandarin background, the difference between mean syllables per run was only 0.34 (Park, 2016). However, the D-measurement result for lexical diversity differs greatly (M=58.55 for Hindi speakers and M=65.73 for Mandarin speakers). Also, You (2014) found that L1 Hindi speakers (M=8.04) made more use of words from the Academic Word list (Nation, 2001) than L1 Mandarin speakers (M=6.60).

To accomplish a descriptive analysis of oral proficiency profiles within examinees rated as 50, this study will conduct a cluster analysis as the next step, presenting more detailed classification information about emergent profiles. Future analysis will also help clarify possible connection between profile information and rating practices, meanwhile validating representative measurement indices for each construct.

Nazlinur Gokturk
Iowa State University

Online Video-Conferencing for Delivering Interactive Speaking Tasks: Opportunities and Challenges

Despite the importance of interactional competence as part of one's oral language ability, assessing interactional competence in a valid and practical way remains as a challenge for many large-scale speaking assessments. While paired/group oral tasks have often been reported to measure the construct of interactional competence in a valid way (Ockey, Koyama, Setoguchi, & Sun, 2015; Brooks & Swain, 2014), the use of these tasks in large-scale face-to-face speaking tests is logistically demanding and complex. In this work-in progress research, the potential usefulness of online video-conferencing for delivering interactive speaking tasks in an ESL placement test has been explored. Following the principles of Evidence-centered assessment design (Mislevy & Haertel, 2006), two prototype interactive speaking tasks have been designed. In the first task, test takers are expected to make a short presentation on a disciplinary topic, ask and answer questions appropriately in a follow-up Q/A session. In the second task, each test taker is first asked to summarize a short text about a possible solution for a problem and then participate in a group discussion to decide on the most appropriate solution for the problem given. The test is delivered using Zoom, an online video conferencing software, and test sessions are video-recorded for subsequent rating. To assess dimensions of oral performance, analytical rating scales have been developed for each task type. To gather test taker's perceptions of the tasks and technology used in the session, a follow-up survey has been administered during piloting. In this presentation, I will present the rationale of this test development project, describe the tasks as well as preliminary rating scale used in the test, and discuss potential benefits and challenges of using online-video conferencing for delivering a speaking test based on data collected during piloting.

Rosalie Grant
WIDA at the Wisconsin Center for Education Research

Cathlin Foy
WIDA at the Wisconsin Center for Education Research

Alaska Native Language, Yugtun, Proficiency Test: Early Reflections on Enacting a Community-Based Participatory Research Paradigm

Purpose

WIDA researchers are supporting the development of a Yugtun Proficiency Test (YPT) through building the capacity of educators and administrators in a large, off-road school district in Alaska. Community-Based Participatory Research (CBPR) principles and practices are being enacted to:

- Develop, and successfully implement, a culturally responsive, valid and reliable YPT PreK-6 system for the district.
- Provide critical data with which to examine relationships across Yugtun, academic English language proficiency, and academic content achievement.
- Use lessons learned to benefit Native and non-Native educators wishing to initiate and/or expand their heritage language programs.

Methodology

CBPR, as defined by Isreal, et al. (1998) is "[A] collaborative approach to research that equitably involves, for example, community members, organizational representatives and researchers in all aspects of the research process." For several decades partnerships among Tribal communities and university researchers have used CBPR principles to address health concerns in Native communities. The YPT project is applying CBPR practices in healthcare research to the field of language Assessment Development (CBPAD) by positioning Native leaders as the drivers of the project and accountability discussions.

Results

In the first two years of the five-year project WIDA researchers have been reflecting on enacting CBPAD in the partnership with the Yup'ik community. One of the unique features of this partnership is including Yup'ik Worldview and Non-verbal subtests in the YPT. Yup'ik members have identified the main features of the non-verbal language they want to assess. Features include non-verbal language critical to community life that support the desire of Yup'ik people for the assessment to stand on Yup'ik ground.

Reference

Israel BA, Schulz AJ, Parker EA, Becker AB. (1998) Review of community-based research: assessing partnership approaches to improve public health. *Annual Review of Public Health. 19:*173-202.

Elizabeth Lee
Iowa State University

The Effect of Using Diagnostic Tests for Placement Decision-making in Higher Education

Placement tests and diagnostic tests are typically intended for two different purposes, but it is not uncommon to observe institutions using placement tests for diagnostic purposes and diagnostic tests for placement purposes (Alderson, 2005). Because the consequences may not be severe, stakeholders may not revise their current policies. They may instead identify that students who failed to pass the diagnostic test are required additional ESL training. However, it is difficult to grasp the extent to which test users and relevant stakeholders benefit from using a diagnostic test for placement purposes. To that end, this study poses two research questions: (1) What is the relationship between students' performances on the reading diagnostic test and their course performance in an ESL reading course? (2) What factors contribute to students' performances on the diagnostic test and their coursework? Test scores and grades were collected from 30 students who took an ESL reading course at a large midwestern university. Additionally, semi-structured interviews were conducted in order to identify the factors that contributed to test and coursework performances. To answer the first question, descriptive statistics were used to identify relationships between students' performances on the diagnostic reading test and their final grades in the ESL reading course. To answer the second question, student interviews were conducted and thematic analysis was performed. Preliminary results suggest that there is no relationship between the distribution of diagnostic test scores and course grades. Upon interviewing students, several social, academic, and motivational factors appeared to contribute to students' overall performances on the test and their coursework. This suggests that the diagnostic reading test that is used to place students into the ESL reading course may not reveal how students will actually perform in the ESL reading course due to various factors that impact students' class performance.

Brent Miller
Michigan Language Assessment

Patrick McLain
Michigan Language Assessment

Renee Saulter
Michigan Language Assessment

Rachele Stucker
Michigan Language Assessment

Creating a Listening Test to Track Young Learners' Language Development

Constructing a listening test to track the development of young learners presents particular challenges, including theoretical construct, operationalization, target-language use situation, and target language ability (Buck, 2001; Gu, 2015). The development process is also complicated by two additional factors: (1) targeting multiple ability levels across varied L1 backgrounds and cultures with the intention of providing cut scores (Zieky et al., 2008) and (2) selecting the most appropriate task types and format for beginners (Field, 2015).

In this poster, we describe the development process and pilot results of a multi-level listening test to track young learners' development of English from the A1 to B1 levels on the Common European Framework of Reference for Languages (CEFR). We demonstrate how we applied Buck's "default listening construct" (2001, p. 112-115) to an EFL test for young learners from around the world, and developed task types that were appropriate and effective at assessing the ability of these learners: picture tasks, listener-directed questions, and non-listener-directed dialogues and talks. We share the results of our pilot study to highlight the performance of the different task types and show how our predictions of the CEFR levels assessed by each task type (i.e., A1, A2, B1, etc.) were reaffirmed in our pilot results (e.g., picture tasks, which were intended to be accessible to beginners, were found to be significantly less difficult than the non-listener-directed talks, which were intended to be accessible to intermediate learners). We also discuss adjustments that were made to the test construct based on our analysis of the pilot study results, such as the removal of an item type from the final operationalized construct and the increase in word count limits for dialogues and talks.

References:

Buck, G. (2001). *Assessing listening.* Cambridge: Cambridge University Press.

Field, J. (2015). *The effects of single and double play upon listening test outcomes and cognitive processing.* British Council.

Gu, L. (2015). Language ability of young English language learners: Definition, configuration, and implications. *Language Testing, 32*(1), 21-38.

Zieky, M. J., Perie, M., & Livingston, S. A. (2008). *Cutscores: A manual for setting standards of performance on educational and occupational tests.* Princeton, NJ: Educational Testing Service.

Jennifer Piotrowski
Michigan Language Assessment

Sharon Pearce
Michigan Language Assessment

Brent Miller
Michigan Language Assessment

Rachel Basse
Michigan Language Assessment

Approaches to Scoring Task Completion in Speaking Performances: Balancing the Need for Validity and Practicality

A variety of rating tools exist for assessing L2 speaking performances, including holistic and analytic rating scales, decision trees, and checklists. Stakeholders often raise validity and practicality concerns for each instrument, such as how its categories and levels are interpreted and used by raters (Galaczi et al., 2011), or the ease and efficiency of the rating process (Brown, 2006). The current study explores how these considerations are applied in scoring task completion in speaking performances, as part of a larger rating tool development project for a multi-level speaking test (CEFR levels A1-B1) for learners ages 11 to 15.

This poster compares three tools for rating speaking task completion, with varying levels of descriptor specificity: a holistic task completion scale (low specificity), a three-category analytic scale (moderate specificity), and a detailed checklist with multiple task-specific criteria (high specificity). For the initial pilot phase, speaking performances were scored for task completion using the three-category analytic scale. In the second phase, performances were scored using the holistic scale and the detailed checklist. In the third phase, a revised version of the checklist was used to evaluate speaking performances. At each phase, raters' scores and feedback on the rating instrument(s) were collected by means of surveys. Raters commented on ease of application, ambiguities between categories, and any difficulties encountered in applying the tools. The poster summarizes these quantitative and qualitative data.

In evaluating the viability of each rating instrument, we found benefits and costs to each format and their levels of descriptor specificity. Ultimately, the decision as to which tool to employ operationally was based on which best balanced ease of rating while effectively capturing task completion in a way that would provide comprehensive feedback to test takers. Beyond addressing these concerns, this study provides additional insights into less-traditional approaches to rating speaking performance.

References

Brown, A. (2006). An examination of the rating process in the revised IELTS Speaking Test. International English Language Testing System (IELTS) *Research Reports 2006: Volume 6,* 1.

Galaczi, E. D., French, A., Hubbard, C., & Green, A. (2011). Developing assessment scales for large-scale speaking tests: A multiple-method approach. *Assessment in Education: Principles, Policy & Practice, 18*(3), 217-237.

Suchada Sanonguthai
Indiana University-Bloomington

Detecting Differential Item Functioning Items in a University-Level English Vocabulary Subtest

The present study analyzed an English vocabulary subtest (25 dichotomously scored items) from a university admission test for differential item functioning (DIF) analysis for two groups of high school test-takers in Thailand. The main goal of the study is to investigate whether test takers with the same gender, nationality, first language, and curriculum but different in types of bilingual education programs show differing probabilities of success after matching on their underlying ability. The study was conducted using data from a large-scale English assessment, which is used to help inform undergraduate university admissions' decision making in Thailand. Data in this study were collected from 507 high school male students from two types of schools in Thailand. The first group (N = 211) came from a bilingual school where both Thai and English are used as mediums of instruction by Thai and English native speaker subject teachers. The second group (N = 306) came from an English immersion school where only English native speaker teachers teach all school subjects. The method used for DIF analysis was the logistic regression procedure (Swaminathan & Rogers, 1990), which is suitable for the present study due to the relative small sample size and to its capacity for detecting both uniform and non-uniform DIF effects. Following Zumbo's rule of thumb that R2 needs to be at least 0.13 for an item to be identified as displaying significant DIF, the magnitude of DIF effect in two non-uniform DIF items were moderate. Content analysis of DIF items suggested that both construct-irrelevant and construct relevant DIF might be potential sources. This study is responsive to major changes in K-12 education reform in Thailand and its increasing number of bilingual and English immersion schools.

Sonca Vo
Iowa State University

Effects of Task Types on Interactional Competence in Oral Communication Assessment: A Rater Cognition Study

Studies of interaction in speaking assessment have highlighted problems regarding the unequal distribution of interaction patterns in different task types (i.e., interviews versus paired/group formats; Van Lier, 1989; Young & He, 1998). However, little attempt has been made in these studies to include not only the verbal but also the nonverbal interaction features elicited in interviews versus paired/group formats. Therefore, the purpose of this study was to examine the effects of task types on verbal and nonverbal interaction features in speaking assessment by investigating which types of interaction features raters noticed when rating interaction. To achieve this goal, the two tasks to be studied were the individual scripted interview task and the paired discussion task. The study qualitatively analyzed the use of verbal and nonverbal interaction features in 20 verbal reports from five raters. The raters commented on interaction features that affected their judgment, using a stimulated verbal recall method. The findings of the study showed that raters attended to a wider range of interaction features when judging test takers' interactional competence in the paired discussion task than in the individual scripted interview task. Theoretically, the study suggested that not only verbal interaction, but also nonverbal communication, contributed to interactional effectiveness. Practically, it suggested that different task types affected the elicitation of interaction features, and that the paired discussion task seemed to be more effective in measuring test takers' interactional competence. For the research implication, this study suggested that there would be a need to compare between the two research methods, a stimulated verbal recall and a discourse analysis, to investigate if salient interaction features found in a discourse analysis approach are similar to salient features elicited in a rater's stimulated verbal recall.

Fang Wang
University of Iowa

Examining Developmental Measures and Proficiency Measures in the TOEFL iBT Integrated Writing Task

One type of developmental studies in second language acquisition literature is the developmental index studies (Wolfe-Quintero, Inagaki, & Kim, 1998). It aims to investigate the development of learners at known proficiency levels through the use of fluency, accuracy, and complexity (CAF) measures. However, very few studies have examined comparisons between developmental measures and proficiency measures (Young, 1995). In other words, whether the writing rating scales can detect real changes in linguistic features over time and which aspects of the rating scales are sensitive to writing development are lack of study (Polio, 2017).

This study aims to examine the essays written by ELLs at three different proficiency levels who are preparing for the TOEFL iBT independent writing task across three different writing occasions. The author will investigate the following research questions:

1. To what extent and how do the first essay written by different groups of participants differ in terms of complexity, accuracy, and fluency?
2. To what extent and how do the complexity, accuracy, and fluency changes among essays written by different groups of participants across time?
3. To what extent and how do the complexity, accuracy, and fluency of the essays relate to participants' writing scores across time?
4. To what extent and how are the changes in complexity, accuracy, and fluency reflected in the rubric descriptors?
5. What is the relationship between participants' initial L2 writing ability and the length of practice and their changes of complexity, accuracy, and fluency across time?

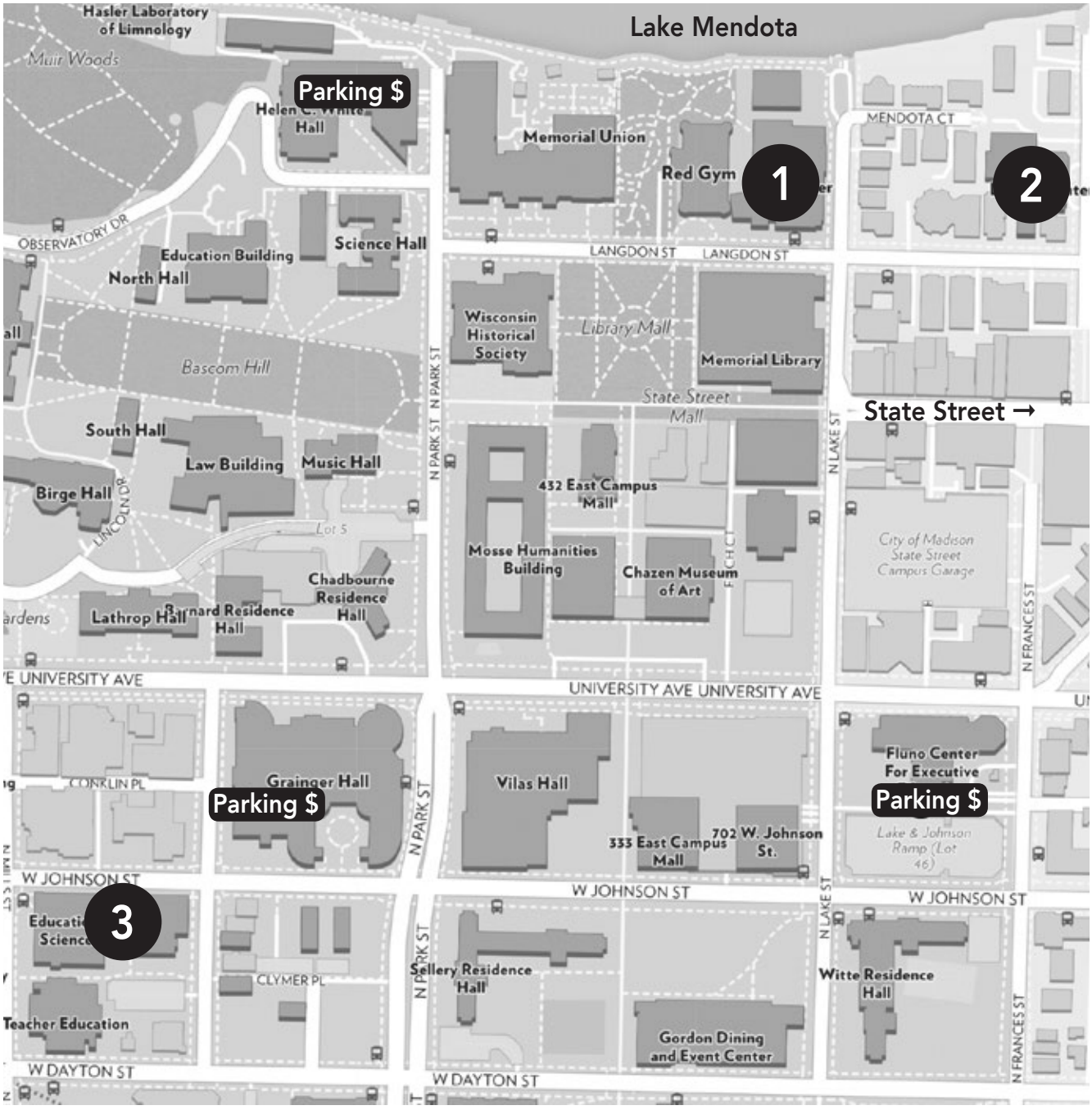
The study aims to understand ELLs' development of CAF variables over time and compare their development with proficiency measures on the rubric. Coh-Metrix and Criterion will be used to analyze CAF variables. ANOVA and Pearson Correlation will be employed to answer research questions.

Maps

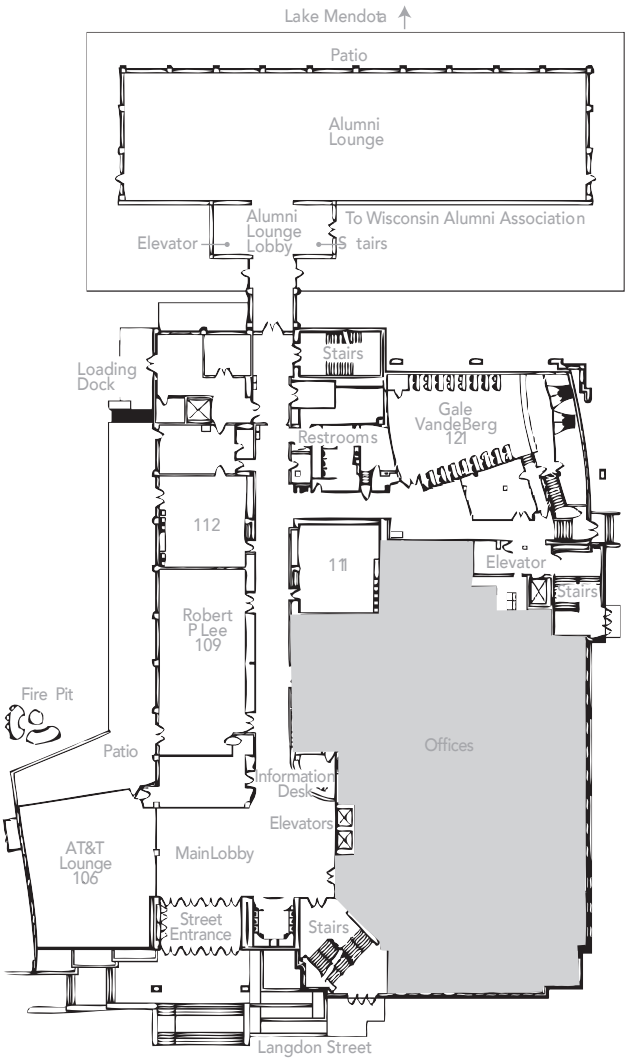
1 **The Pyle Center**
(conference location)
702 Langdon St.

2 **The Lowell Center**
(block of rooms held here)
610 Langdon St.

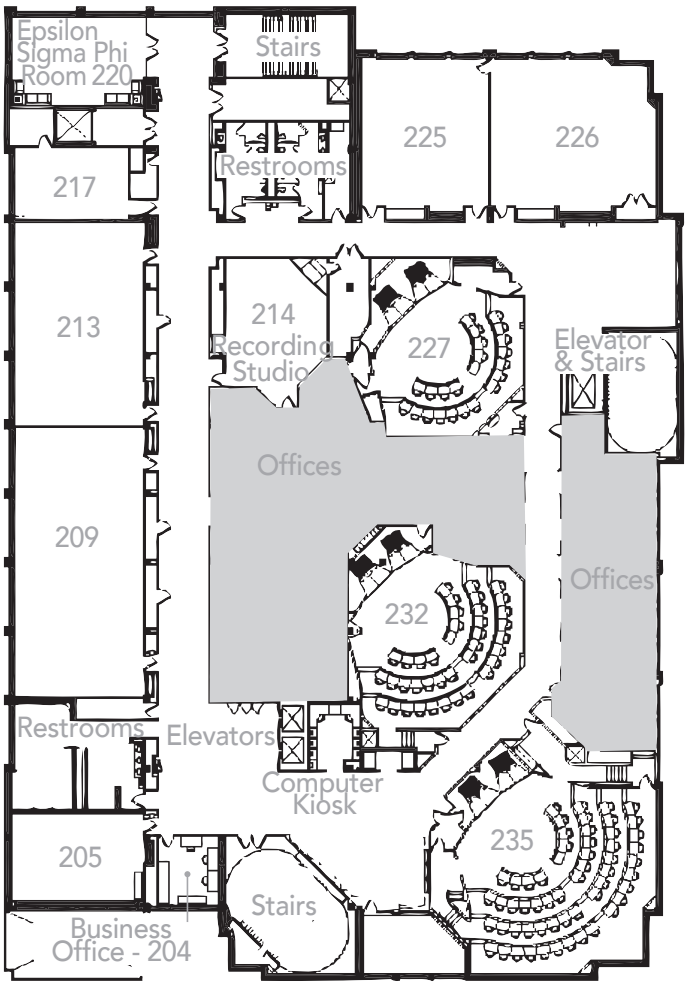
3 **Educational Sciences**
(pre-conference workshops)
1025 W. Johnson Street



Pyle Center First Floor



Pyle Center Second Floor



Index

Are you interested in Language Aptitude?

Are you interested in the variables associated with success in second language learning?

Then, consider using the

Pimsleur Language Aptitude Battery

or the

Modern Language Aptitude Test MLAT-E

for your research needs.

The MLAT-E can be used with children ages 8-12, while the PLAB can be used with students ages 12 - adult.

Each is a valid and reliable measure of language learning aptitude and associated cognitive skills.

LANGUAGE LEARNING AND TESTING FOUNDATION

For more information on these well-established instruments or an order form, check our website (www.LLTF.net).

Language Learning and Testing Foundation
Rockville, MD 20852

WIDA™

MISSION

WIDA advances academic language development and academic achievement for children and youth who are culturally and linguistically diverse, through high quality standards, assessments, research, and professional learning for educators.

VISION

To be the most trusted and valued resource in supporting the education of multilingual learners.

VALUES

Innovation: Drawing upon research and practice to create the best resources for children, youth, and educators.

Service: Exceeding expectations with trusted and knowledgeable support of our clients and stakeholders.

Can Do Philosophy: Recognizing and building upon the assets, contributions, and potential of culturally and linguistically diverse children and youth.

Collaboration: Facilitating interaction among educators, state and local educational agencies, researchers, policy-makers, and experts worldwide.

Social Justice: Challenging linguistic discrimination, cultural biases, and racism in education

Proud to host the 20th Annual MwALT Conference

Looking forward to next year at Indiana University!