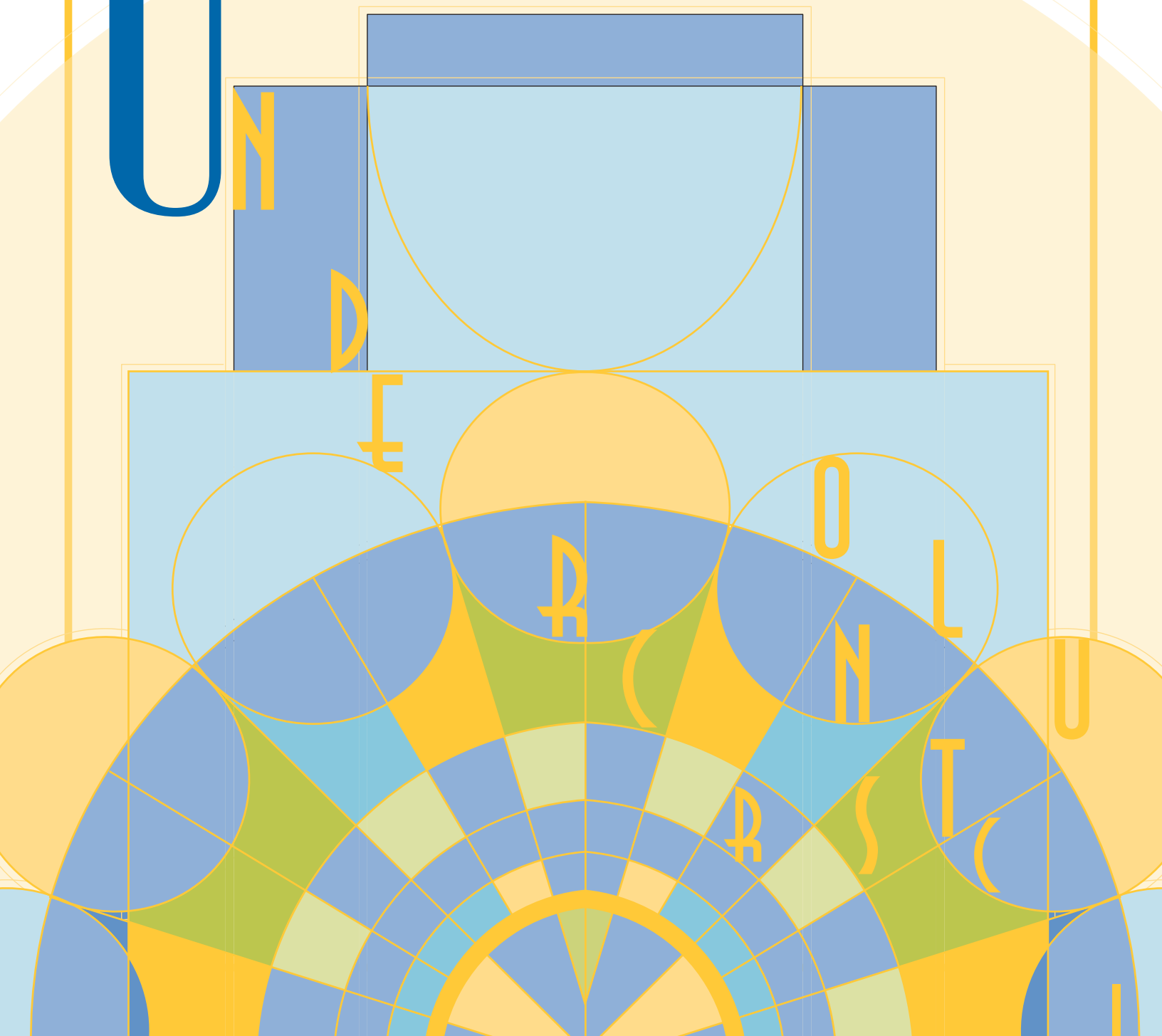
MwALT 2014

Midwest Association of Language Testers
Ann Arbor, Michigan, **Oct 3–4**

Under Construction

BUILDING ARGUMENTS, ASSESSMENTS, AND EXPERTISE



Contact Information

CaMLA Office *(on the third floor)*
Argus 1 Building
535 West William St., Suite 310
Ann Arbor, Michigan
48103-4978 USA

Tel: +1 866.696.3522 or +1 734.615.9629

Questions?

At any time:
mwalt2014@cambridgemichigan.org

During the conference:
Text or phone MwALT 2014 conference organizers Mark Chapman or Jessica O'Boyle:

- 734-255-7580 (Mark)
- 407-491-1188 (Jessica)

#mwalt14

TABLE OF CONTENTS

Find out why so many organizations trust CaMLA for their language assessments

Cambridge Michigan Language Assessments—CaMLA—offers a complete range of English language assessment and learning tools, including the well-known suite of University of Michigan exams. The MELAB and ECPE are accepted English language admissions credentials at colleges and universities worldwide.

CaMLA exams offer high-quality, objective proof that candidates have the American English language skills to succeed.

CAMBRIDGE ENGLISH Language Assessment
Part of the University of Cambridge

UNIVERSITY OF MICHIGAN

CambridgeMichigan.org

Welcome to MwALT 2014!

Welcome to the sixteenth annual conference of the Midwest Association of Language Testers. All of us here at Cambridge Michigan Language Assessments are delighted to host this year's MwALT conference at the University of Michigan in beautiful Ann Arbor. We would especially like to extend a warm welcome to our presenters who have traveled from across the Midwest and from further afield. This year we have speakers from Northern Arizona University, University of Maryland, Columbia University, and Georgia State University, in addition to several Midwest schools. These are exciting times for the MwALT organization as our group continues to grow and reach beyond the "traditional" boundaries of our community.

This year our conference theme is "Under Construction: Building Arguments, Assessments, and Expertise," which provides a forum for the research in language testing that is being conducted to develop increasingly robust theoretical and practical approaches to our important work in language assessment. We are confident that the talks and posters being presented at this year's conference will create a stimulating environment for the exchange of ideas between graduate students, faculty, and language testing professionals across our field. We look forward to having you as our guests.

Thank you for spending your valuable time with MwALT and enjoy the show!

Casey Marks
Chief Executive Officer, CaMLA

Conference Organizers

We'd like to recognize the dedication and hard work of all of those who helped make MwALT 2014 a success.

Jayanti Banerjee
Kate Boyd
Chris Burnett
Mark Chapman
Crystal Collins
Ed Cormany
Barbara Dobson
Heather Elliott
Ericka Finley
Caitlin Hirsch
Nabila Khan
Alexis Kielwasser
Eric Lagergren
Amanda McConville
Robert McCord
Natalie Nordby Chen
Jessica O'Boyle
JP Pagán
Renee Saulter
Katie Weyant

CAL Salutes MwALT

The Center for Applied Linguistics is proud to support
the Midwest Association of Language Testers
and its ongoing commitment to language testing.

Visit CAL's Redesigned Website

With a clean, modern look and user-friendly navigation, our new website provides the latest information about our research, projects and resources.

We encourage you to browse our new website on your desktop, laptop, tablet or phone.

Our URL below remains the same.

Learn about the CAL Internship Program

The internship program at CAL provides an opportunity for qualified individuals interested in language, linguistics, language testing, and related areas to gain experience working with CAL professional staff.

To learn more, visit
www.cal.org/who-we-are/career-opportunities/interns

Browse the CAL website for the latest updates and information.

www.cal.org

CAL
CENTER FOR APPLIED LINGUISTICS

Sponsors

A big thanks to the organizations and departments that helped sponsor MwALT 2014:

Welcome Reception
CaMLA

Plenary Speech
College of Literature, Science, and the Arts
University of Michigan

Refreshment Break
Linguistics
University of Michigan

Boxed Lunch
College of Literature, Science, and the Arts
University of Michigan

Refreshment Break
English Language Institute
University of Michigan

Awards Presentation
Educational Testing Service (ETS)

Closing Reception
Cambridge English Language Assessment

WORKSHOP

Workshop

Training speaking test examiners: The devil is in the detail

Instructors: Jessica O'Boyle and Mark Chapman

Friday, October 3 • 1:00–4:00 p.m. • CaMLA Office

The importance of the speaking examiner has been well documented in the second language speaking assessment literature. Speaking examiners are responsible for creating a comfortable environment for the test taker, administering the speaking test according to documented protocols, and awarding scores consistently from an established scoring rubric (Brown, 2005; Fulcher, 2003; O'Sullivan, 2008; Underhill, 1987). Although the literature has much to offer on examiner behavior and scoring, there is relatively little guidance regarding the specifics of how speaking examiners should be trained.

Participants in this workshop will have the opportunity to learn about the practical details of speaking examiner training provided in two different test programs; a speaking test used to screen international teaching assistants at The University of Michigan, and a multi-task, multi-level test of general spoken English language proficiency. After completing the workshop, participants will have a better understanding of how to effectively train speaking examiners in their own local context.

Workshop Schedule of Activities

Part 1 1:00–2:00

- Introduction to workshop and participants
- Questionnaire activity to establish participants' experience with speaking tests and their familiarity with speaking examiner (SE) training.
- Ranking activity to encourage participants to think about the key skills required to be a capable SE.
- Presentation of principles of SE training that help ensure SEs will develop the skills established in the previous activities

Part 2 2:10–3:15

- Present overview of SE training materials for the MET Speaking Test
- Watch two videos of MET Speaking test takers and review the importance of highlighting the main features of the performance that justify the score awarded

Part 3 3:20–4:20

- Present overview of SE training materials for ITASA
- Watch video of two test takers performing different test tasks; participants complete activities that emphasize the connections between the performances, the rating scale wording, and the commentary on the performances in the training materials

Part 4 4:30–5:00

- Discuss how scoring tolerances and certification criteria are set
- Summarize main points raised in workshop.

References

Brown, A. 2005. *Interviewer variability in oral proficiency interviews*. Frankfurt Am Main, Germany: Peter Lang.

Fulcher, G. 2003. *Testing second language speaking*. Harlow, UK: Pearson Education Limited.

O'Sullivan, B. 2008. *Modelling performance in tests of spoken language*. Frankfurt Am Main, Germany: Peter Lang.

Underhill, N. 1987. *Testing spoken language: A handbook of oral testing techniques*. Cambridge, UK: Cambridge University Press.

PLENARY

Plenary

A corpus linguist's view on speech and speaking assessment: Searching for patterns

Ute Römer, Georgia State University uroemer@gsu.edu

Saturday, October 4 • 8:45–9:30 a.m. • Forum Hall

The past few decades have witnessed a massive increase in corpus research activity in a range of linguistic subfields, including strands within Applied Linguistics. Corpora are increasingly accepted as powerful tools that help us gain insights into language structure and use, and help inform language teaching and testing practice (see, for instance, Flowerdew 2012, Hawkins & Filipovic 2012, Reppen 2010, and Römer 2011). This paper discusses the importance of considering corpus evidence in highlighting central aspects of spoken language and addresses the question "How can corpus tools and techniques help us shed light on the concept of speaking?" It also looks at speaking tests from a corpus perspective to see how well they reflect central patterns of speech. Since spoken language is not a uniform phenomenon but varies considerably depending on the context of use, the paper does not attempt to describe speech "in general." Instead, it focuses on one particular, more specialized type of language: spoken English produced in a US research university setting. This type of language is captured in MICASE, the Michigan Corpus of Academic Spoken English (Simpson et al. 2002), a collection of 152 transcripts and 1.8 million words, based on 200 hours of recordings of speech events from across the University of Michigan in Ann Arbor.

The paper starts out with a brief analysis of frequency word and keyword lists of academic speaking (compared to academic writing) and then focuses on phraseological items (variably referred to as n-grams, formulaic sequences, lexical bundles, clusters, etc.) that are particularly common in speaking and carry important discourse functions. Software packages for corpus access and analysis are used to extract lists of contiguous word sequences (n-grams, e.g. *you know*, *a lot of*) and non-contiguous word sequences (phrase-frames, e.g. *a * of*, *I don't * so*) of different lengths from MICASE. The resulting lists are filtered manually for items that play a central role in academic speech and appear to have a particularly high communicative value. The final section of the paper reviews rubrics of a selection of high-stakes speaking tests and discusses in how far these rubrics capture central aspects of spoken language as highlighted by corpus analysis. It then discusses implications of our MICASE-based findings for (academic) speaking assessment. Overall, the paper provides evidence for the interrelatedness of vocabulary and grammar in academic speech and stresses the importance of phraseology as a core, rather than a peripheral aspect of language (cf. Ellis 2008), adding to a growing body of existing work in corpus research on phraseology (see e.g. Biber 2009; Hoey 2005; O'Donnell, Römer & Ellis 2013; Römer 2009, 2010; Sinclair 2008). It demonstrates how corpus analysis can contribute to a better understanding of core aspects of speech and how it helps us uncover the patterned nature of speaking.

Ute Römer's bio is available at
CambridgeMichigan.org/mwalt/speaker

References

Biber, D. 2009. A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics* 14(3): 275–311.

Ellis, N. C. 2008. Phraseology: The periphery and the heart of language. In F. Meunier & S. Granger (Eds.), *Phraseology in Language Learning and Teaching* (pp. 1–13). Amsterdam: John Benjamins.

Flowerdew, L. 2012. *Corpora and Language Education*. London: Palgrave Macmillan.

Hawkins, J. A. & L. Filipovic. 2012. *Criterial Features in L2 English*. Cambridge: Cambridge University Press.

Hoey, M. 2005. *Lexical Priming. A New Theory of Words and Language*. London: Routledge.

O'Donnell, M. B., U. Römer & N. C. Ellis. 2013. The development of formulaic sequences in first and second language writing: Investigating effects of frequency, association, and native norm. *International Journal of Corpus Linguistics* 18(1): 83–108.

Reppen, R. 2010. *Using Corpora in the Language Classroom*. Cambridge: Cambridge University Press.

Römer, U. 2009. The inseparability of lexis and grammar: Corpus linguistic perspectives. *Annual Review of Cognitive Linguistics* 7: 140–162.

Römer, U. 2010. Establishing the phraseological profile of a text type: The construction of meaning in academic book reviews. *English Text Construction* 3(1): 95–119. [Reprinted in: Biber, D. & R. Reppen (eds.). 2012. *Corpus Linguistics. Volume I: Lexical Studies*. London: SAGE Publications. 307–329.]

Römer, U. 2011. Corpus research applications in second language teaching. *Annual Review of Applied Linguistics* 31: 205–225.

Sinclair, J. M. 2008. The phrase, the whole phrase, and nothing but the phrase. In S. Granger & F. Meunier (Eds.), *Phraseology: An Interdisciplinary Perspective* (pp. 407–410). Amsterdam: John Benjamins.

Simpson, R. C., S. L. Briggs, J. Ovens & J. M. Swales. 2002. *The Michigan Corpus of Academic Spoken English*. Ann Arbor, MI: The Regents of the University of Michigan.

CONFERENCE SCHEDULE

Conference Schedule

Poster Sessions
Location: Atrium 4

Examining the English reading section of the college entrance exams in East Asian countries

Renka Ohta, University of Iowa
Jui-Teng Liao, University of Iowa
Han Gil Kim, The Ohio State University

One Task Fits All? Constructing a Writing Assessment for a Wide Range of Abilities

Edward Cormany, CaMLA
Rachele Stucker, CaMLA
Mark Chapman, CaMLA
Ummehaany Jameel, CaMLA

Test stakeholders' perceptions of a post-placement assessment in ESL reading classes

Shannon McCrocklin, Iowa State University
Zhi Li, Iowa State University

Towards a Computer-Adaptive Rasch-Based Test of Productive Academic Vocabulary Knowledge (AVK)

Victor D.O. Santos, Iowa State University

A Validation Study of the Reading Section of Young Learners Tests of English (YLTE)

Shinhye Lee, Michigan State University
Irene Ahn, Michigan State University
Ina Choi, Michigan State University
Yaqiong Cui, Michigan State University
Hyung-jo Yoon, Michigan State University
Paula Winke, Michigan State University

Consequences of Violating the Monotonicity Assumption in Dichotomous IRT Models

Ryan Lidster, Indiana University-Bloomington

Beyond Assessment Literacy: ESL Teachers' Practices on Constructing Cognitive Diagnostic Assessment

Junli Wei, University of Illinois at Urbana-Champaign

Exploring the Content Structure of IELTS Speaking Topics: A Multidimensional Scaling Analysis

Abdolvahab Khademi, University of Massachusetts-Amherst
Akram Nayernia, Iran University of Science and Technology

Teacher Candidates' Evaluation of English-as-a-Second Language Writing

Hyun-Sook Kang, Illinois State University

Rater's Perceptions of Evaluative Criteria in an English Writing Placement Test

Nancy Ngan Vu, Iowa State University

Conference Sessions

Friday, October 3
University of Michigan — CaMLA Office

12:30–5:00	**Registration Open: CaMLA Office**
1:00–5:00	**Workshop: CaMLA Office** *Training Speaking Test Examiners: The Devil Is in the Detail* Jessica O'Boyle, *CaMLA* Mark Chapman, *CaMLA*
5:00--7:00	**Welcome Reception: CaMLA Office** Sponsored by: CaMLA

The CaMLA office is located at 535 West William St. in the Argus 1 Building on the third floor.
See the maps on p. 29 for more information about conference locations and addresses.

Saturday, October 4
University of Michigan — Palmer Commons, 4th Floor

7:30–8:00	**Exhibits Setup: Main Foyer**	
7:30–10:00	**Poster Setup: Atrium 4**	
8:00–12:00	**Registration Open: Main Foyer**	
8:00–5:00	**Exhibits Open: Atrium 4**	
8:00–9:00	**Breakfast: Main Foyer**	
8:30–8:45	**Welcome: Forum Hall** *CaMLA*	
8:45–9:30	**Plenary Speech: Forum Hall** *A corpus linguist's view on speech and speaking assessment: Searching for patterns* Ute Römer, *Georgia State University* Sponsored by: University of Michigan (LSA: College of Literature, Science, and the Arts)	
	Paper Session 1: Forum Hall Session Chair: Robert McCord	Paper Session 2: Great Lakes Central Session Chair: Suthathip Thirakunkovit
9:35–10:05	*Investigating the Multi-faceted Nature of Speaking Performance with a Multivariate Method* Shelley Staples, *Purdue University* Geoffrey LaFlair, *Northern Arizona University* Jesse Egbert, *Brigham Young University*	*Utilization and Ramification Inferences: Building a validity argument beyond score interpretations* Yoo-Ree Chung, *Iowa State University*
10:05–10:35	*How Can Speech Content Contribute to Integrated Speaking Scoring Rubrics?* Huong Le, *Iowa State University*	*Construct Definition: The First Step in Building Validity Arguments—The Case of a Persian Reading Comprehension Test* Payman Vafaee, *University of Maryland*
10:35–11:30	**Poster Session: Atrium 4** *See list of poster presentations on page 4* **Refreshment Break: Main Foyer** Sponsored by: University of Michigan (Linguistics)	

#mwalt14

Saturday, October 4

University of Michigan — Palmer Commons, 4th Floor

	Paper Session 3: Forum Hall Session Chair: Sarah Goodwin	Paper Session 4: Great Lakes Central Session Chair: Nicolas May
11:30–12:00	*Expanding Bias Investigations: The Influence of Reasons for Taking the MELAB* Jayanti Banerjee, *CaMLA* Geoffrey LaFlair, *Northern Arizona University*	*L2 Learners' Use of Interaction Features in Different Paired Speaking Tasks* Linxiao Wang, *Northern Arizona University*
12:00–12:30	*Task Features and Discriminant Validity of Elicited Imitation in L2 Research: A Systematic Review* Xun Yan, *Purdue University* Yukiko Maeda, *Purdue University* Jing Lü, *Purdue University*	*Assessment and feedback: Examining the relationship between self-assessment and blind peer- and teacher-assessment in TOEFL writing* Meghan Odsliv Bratkovich, *Teachers College, Columbia University*
12:30–2:00	**Boxed Lunch: Great Lakes North** Sponsored by: University of Michigan (Literature, Science, and Arts)	
12:30–2:00	**MwALT Business Meeting: Great Lakes South**	
	Paper Session 5: Forum Hall Session Chair: Kristin Graw	Paper Session 6: Great Lakes Central Session Chair: Xun Yan
2:00–2:30	*What's in a score: Integrated skills assessment and language learners' ability in reading, writing, and source use?* Lia Plakans, *University of Iowa*	*The influence of second language knowledge on written proficiency rating: An investigation into raters' behavior* Lorena Valmori, *Michigan State University*
2:30–3:00	*Using Circle-Arc to Equate Placement Tests in Small Language Programs* Maria Nelly Gutierrez Arvizu, *Northern Arizona University* Geoffrey LaFlair, *Northern Arizona University* Daniel Isbell, *Northern Arizona University*	*Automated writing evaluation for formative assessment: Investigating accuracy and efficiency as part of argument-based validation* Jim Ranalli, *Iowa State University* Stephanie Link, *Iowa State University* Evgeny Chukarev-Hudilainen, *Iowa State University*
3:00–3:15	**Refreshment Break: Main Foyer** Sponsored by: University of Michigan (English Language Institute)	
	Paper Session 7: Forum Hall Session Chair: Ed Cormany	Paper Session 8: Great Lakes Central Session Chair: Ummehaany Jameel
3:15–3:45	*A Multi-Facet Rasch Analysis Comparing Essay Rater Behavior on a Test Used for Two Purposes* Sarah Goodwin, *Georgia State University*	*Situating standard setting within argument-based validity* Spiros Papageorgiou, *Educational Testing Service* Richard J Tannenbaum, *Educational Testing Service*
3:45–4:15	*Investigating the rater bias patterns in reading-to-write tasks using FACETS* Sun-Young Shin, *Indiana University*	*Developing a Post-Admissions Business Language Assessment* Geoffrey LaFlair, *Northern Arizona University*
4:15–4:45	*Outcomes-based Assessment for Learning in an EAP Program* Joan Jamieson, *Northern Arizona University* L. D. Nicolas May, *Northern Arizona University*	
4:45–5:15	**Awards Presentation: Forum Hall** Sponsored by: Educational Testing Service (ETS)	
5:30–8:00	**Closing Reception: The Original Cottage Inn Restaurant** 512 E. William St. in Ann Arbor Sponsored by: Cambridge English	*Let's celebrate a successful MwALT 2014 conference! Join us for a dinner buffet and drinks at The Original Cottage Inn (walking distance from Palmer Commons) following the awards presentation.*

PAPER ABSTRACTS

Jayanti Banerjee
CaMLA

Geoffrey LaFlair
Northern Arizona University

Expanding Bias Investigations: The Influence of Reasons for Taking the MELAB

One component of fairness is ensuring a lack of bias in the process, records, and interpretations of a test (Bachman & Palmer, 2010). Differential item functioning (DIF), and differential test functioning (DTF) provide statistical evidence of the presence or lack of bias in items or a test. An important feature of a DIF/DTF study is the grouping characteristic of the test takers. In language testing research most of the research investigating DIF up through 2005 has grouped test takers by language background, gender, or ethnicity (Ferne & Rupp, 2007). To date MELAB DIF investigations have grouped test takers by gender (Wang, 2006; Aryadoust, Goh, and Kim, 2011).

The purpose of this study is to extend the bias research on the MELAB. A recent technical review showed that test takers who use the test for professional certification perform better on the exam. This study seeks to confirm that this pattern is due to a genuine difference in language proficiency rather than an inherent bias in the test. The study groups test takers by reason for taking the test in order to investigate both uniform and non-uniform DIF/DTF. The data comprised four MELAB administrations, and the DIF and DTF analyses were conducted using the Rasch model in Winsteps. The paper will report the analyses: items show statistically significant bias toward one of the subgroups but very few of the items show substantial DIF. It will reflect upon the implications of these findings for MELAB item design.

Meghan Odsliv Bratkovich
Teachers College, Columbia University

Assessment and feedback: Examining the relationship between self-assessment and blind peer and teacher-assessment in TOEFL writing

This study investigated the nature of self-assessment, and blind peer- and teacher-assessment in L2 writing. The type of feedback students gave to themselves and peers, the type of feedback used in the revision process, and the source of the feedback used were all analyzed. Additionally, student perceptions of self- and peer-assessment, feedback, and their relationships to perceived writing improvement were also studied. Findings revealed that students in this study did not use teacher feedback significantly more than feedback from themselves or their peers, but they did give different types of feedback than the teacher, and favored using feedback related to language use in the revision process. Students perceived their writing abilities to have increased due to self- and peer-assessment, but responded more positively to peer-assessment than self-assessment. Surprisingly, students also perceived their abilities to have increased in rubric areas in which the feedback they received was not used and not regarded as useful, and the highest perceived gains in writing ability were in areas which accounted for the lowest amounts of feedback given.

Yoo-Ree Chung
Iowa State University

Utilization and Ramification Inferences: Building a validity argument beyond score interpretations

Since Messick's (1989) seminal paper, it has often been stressed that the validation of test score interpretations and uses should pay sufficient attention to value implications and social consequences of test use (Kane, 2013; McNamara, 2006). However, few empirical studies have been conducted to meet this demand. Kane's (2006) validation framework also fails to address the intended and unintended consequences of a test use as one independent aspect of a validity argument. On the other hand, while paying full attention to test uses and consequences in validation, Bachman and Palmer's (2010) Assessment Use Argument (AUA) framework is not as effective as Kane's approach in addressing validity issues pertaining to score interpretations. Given these issues, the paper will first suggest a hybrid validation framework, taking strengths of Kane's and AUA frameworks, and propose a new term for an inference pertinent to consequences of test use (i.e., Ramification Inference). Second, the paper will demonstrate an attempt to validate a productive grammatical writing test in academic English (an academic grammar test, in short), particularly focusing on building a validity argument around the two inferences on test use and consequences (i.e., Utilization and Ramification). Through this attempt, the study will also show that a validation attempt, even made on a hypothetical situation, can help test practitioners anticipate both intended and unintended consequences of a test use and make appropriate adjustments to mitigate negative consequences.

Sarah Goodwin
Georgia State University

A Multi-Facet Rasch Analysis Comparing Essay Rater Behavior on a Test Used for Two Purposes

Many second language (L2) writing researchers note that scores assigned by human raters can be impacted by various characteristics of raters' background (Cumming, 1990; Vaughan, 1991; Weigle, 1994, 1998, 2002; Cumming, Kantor, & Powers, 2001; Lumley, 2002; Barkaoui, 2010). Weigle (1998) and Lumley (2002) both note the complex nature of the rating process. However, little is known about how raters use the same rubric for different examinee groups.

The present paper focuses on an integrated reading and writing test of academic English used for both admissions and placement purposes for U.S. undergraduate and graduate students. Raters are trained to interpret the analytic scoring rubric similarly no matter which test type is scored, but checking whether raters are indeed behaving consistently is crucial for the validity of the assessment. Using multifaceted Rasch measurement (Linacre, 2014), I investigated 11,392 ratings assigned to 1,512 examinees over seven semesters. The four facets were examinees, raters, test type (admissions or placement), and scales (four analytic ratings of Content, Organization, Accuracy of Grammar/Vocabulary, and Range of Grammar/Vocabulary). Results indicated that some of the 25 raters behaved significantly differently when scoring the two test types, falling into about three distinct rater groups (separation index = 2.8). Four raters were misfitting while one rater was overfitting. Additionally, there were six instances of statistically significant bias shown by five unique raters on admissions or placement tests. The findings suggest that raters may be attributing scores to a wider range of writing ability levels on admissions than on placement tests.

Maria Nelly Gutierrez Arvizu
Northern Arizona University

Daniel Isbell
Northern Arizona University

Geoffrey LaFlair
Northern Arizona University

Using Circle-Arc to Equate Placement Tests in Small Language Programs

Equating is a statistical procedure that adjusts for differences in difficulty between test forms, allowing scores on both forms to be used interchangeably (Kolen & Brennan, 2004). Traditionally used in large scale testing administrations, for programs with small sample sizes the Circle Arc (CA) equating method has been suggested (Livingston & Kim, 2009). CA accounts for a curvilinear relationship between scores on both forms using three points: highest possible score, midpoint, and lowest meaningful score. A step-by-step explanation of how to conduct CA equating will be presented using data from a study conducted in an intensive English program. A non-equivalent anchor test design was used to equate two placement test forms administered in Fall 2011 (N=173) and Fall 2012 (N=88). The listening and reading sections of the tests shared a set of multiple choice, dichotomously scored items (9 listening, 11 reading). The forms were equated using two CA methods—one with a low point of zero and one with a low point of chance. These methods were compared to results of no equating (i.e., identity equating) by analyses of error (random, systematic, and total) and changes in placement decisions. Results indicated that both CA methods were better than no equating, and that the most accurate method was CA with a low point of zero. Small scale testing programs that design tests to make placement decisions could benefit from conducting equating that is practical and appropriate for their contexts.

Joan Jamieson
Northern Arizona University

L. D. Nicolas May
Northern Arizona University

Outcomes-based Assessment for Learning in an EAP Program

Arguments for the validity of assessment programs focus not only on the psychometric properties of a test, but also on the intended use of the test and the involvement of stakeholders (Bachman & Palmer, 2010; Norris, 2013). In an English for academic purposes (EAP) program, important instructional decisions involve monitoring learner progress during instruction (a formative purpose) and deciding whether learners have achieved particular instructional goals (a summative purpose). These purposes can be served with an outcomes-based approach to classroom assessment (Brown and Hudson, 2002; Hughes, 1989; Tyler, 1934). This approach requires classroom teachers to reflect on what their outcomes should be, and then to develop tasks and assess students in relation to those outcomes. By integrating learning and assessment, we are in a position to profit from formative assessment (Black & Wiliam, 1998). This paper presentation will describe the test development process used in our EAP program, beginning with teachers' statements of outcomes that are then subdivided into which are best measured by tests or other means. Teachers, course coordinators, and assessment team members meet to design and develop achievement tests. Each item is coded by its outcome to facilitate score reporting. Automated score reports give total score but more importantly pass/fail decisions for each outcome. Three types of reports are generated—student, class, and level. Information from these reports is used by students, teachers, coordinators, and program administrators to determine strengths and weaknesses for individuals, teachers, and the curriculum. This assessment program puts theory into practice.

Geoffrey LaFlair
Northern Arizona University

Developing a Post-Admissions Business Language Assessment

Over 20% of international students enrolled in U.S. institutions major in business and management; however, many students struggle with the language demands of the university (Andrade, 2006; Institute of International Education, 2012). This presentation will report on the design, piloting, and revision of the Business Language Assessment (BLA), a post-admissions language assessment to identify these students and provide language support (Dunworth, 2009). A needs analysis with input from subject matter specialists informed the development of the BLA as a measure of basic and advanced reading and listening comprehension skills. Textbook passages and lectures were drawn from the business law and ethics course. This step provided a framework to analyze whether the input and tasks were representative of the target language use domain. The BLA was administered in Spring and Fall 2013, and Spring, 2014. Forms were analyzed using classical test theory and the Rasch one-parameter IRT model. Results of the analyses of the first administration of the test indicated that the design of the test did not maximize information about test takers at lower ability levels. Revisions included eliminating certain item types (e.g., text organization) and items that exceeded the desired thresholds of item difficulty and discrimination. Steps were taken to calibrate the BLA through the use of test information functions. The results of the analyses of the redesigned test indicated that the second version return much more information about test takers at lower ability levels. Evidence for the usefulness of the BLA will be provided.

Huong Le
Iowa State University

How Can Speech Content Contribute to Integrated Speaking Scoring Rubrics?

Speech content has taken many forms, such as topic development and sophistication of ideas, in the scoring rubrics of integrated speaking tests (Brown, 2007; Brown, Iwashita, & Mc Namara, 2005; Sato, 2012). The inclusion of speech content as a part of rating scales for grading integrated speaking performance sparks a debate on its validity in assessing oral language proficiency skills (Douglas, 1997; Jamieson et al., 2008; Lee, 2006; Luoma, 2004, Brown, 2007; Brown et al., 2005; Eckes, 2009; Frost et al., 2012; Sato, 2012). In light of the validity argument framework in language testing and assessment (Chapelle, 2012), this paper presentation will identify problems in the development of speech content rating scales for integrated speaking tasks by reviewing current practice and research. First, an illustrative analysis is conducted using two speech content rating scales used for integrated speaking tasks in two widely used English proficiency tests, the Oxford English (OE) language pilot test and the Internet-Based Test of English as a Foreign Language (TOEFL iBT). The analysis then opens an in-depth discussion on sources of threats to construct validity and reliability in content-rating for integrated speaking tasks, leading to several implications for future research on developing speech content rating scales for integrated speaking tasks.

Spiros Papageorgiou
Educational Testing Service

Richard J. Tannenbaum
Educational Testing Service

Situating standard setting within argument-based validity

In recent years, argument-based validity theories (Chapelle, 2008; Kane, 2006, 2013) have been particularly influential in the field of second language assessment. Language testers have also started showing an increased interest in standard setting (Cizek & Bunch, 2007), the decision-making process of setting minimum scores (cut scores) to classify examinees into a number of levels or categories. This increased interest is primarily because of the publication of the Common European Framework of Reference (CEFR; Council of Europe, 2001), which defines learning objectives at different proficiency levels. However, it is not always clear how standard setting fits into argument-based validity, in particular in the context of language assessment. In this paper we first argue that standard setting is an essential part of the test development and validation process because of the important consequences cut scores might have. We then explain how evidence from standard setting can support claims about consequences, decisions and interpretations presented in the Assessment Use Argument (AUA) framework (Bachman & Palmer, 2010). We finally identify several challenges in setting cut scores in relation to the CEFR levels and argue that despite these challenges, standard setting is a critical component of any claim focusing on the interpretation and use of test scores to classify test takers in relation to the CEFR levels. The main point of this paper is that standard setting should be an integral part of the validity argument supporting score use and interpretation and not be treated as an isolated event that occurs between the completion of test development and the reporting of scores.

Lia Plakans
University of Iowa

What's in a score: Integrated skills assessment and language learners' ability in reading, writing, and source use?

The use of integrated skills assessment (ISA) tasks has blossomed in large scale testing of English as an academic language as well as in assessment of task-based and content-based language learning. Along with this trend, research in language testing has been investigating the issues surrounding these test tasks. The first wave of studies since 2000 centered on comparing integrated with non-integrated assessment tasks; however, research questions have advanced to concerns about ISA task design, raters, and score interpretation. Score interpretation research is directly linked to validation study and considers the relationship between processes, products, and score. This presentation will discuss research on score interpretation in three areas: reading, writing, and source use. In each area, general findings will be presented that summarize research on ISA; then more detail will be given on a focal study to illustrate ISA in reading, in writing, and in source use. In reading, the focal study investigated the reading strategies test takers' used in composing writing for an ISA. In writing, the focal study looked at the relation between ISA score and organizational features such as organization pattern, coherence, and cohesion. Lastly for source use, the focal study investigated the relation of selecting important ideas, the use of reading and listening source materials, and verbatim source use to ISA score. The findings from this research can provide test developers, researchers, and users with a view on how test takers approach these tasks and what inferences can be made from integrated writing scores.

Jim Ranalli
Iowa State University

Stephanie Link
Iowa State University

Evgeny Chukarev-Hudilainen
Iowa State University

Automated writing evaluation for formative assessment: Investigating accuracy and efficiency as part of argument-based validation

Increasing use of tools for automated writing evaluation (AWE) in composition classrooms demonstrates growing interest in these tools' potential for formative assessment. A recent interpretive argument (Chapelle, Cotos & Lee, 2013) outlined the evidence needed to support use of the Criterion AWE tool in a college-level ESL writing course. The present research contributes to a critical review of classroom applications of AWE by investigating evidence for the assumptions underlying two key inferences in this argument. In the first of two studies, we investigated backing for the evaluation inference addressing the claim that Criterion provides students with accurate feedback. We collected all writing submitted to Criterion by students in courses for lower- and higher-proficiency students over a semester, identified ten of their most common error types, and had two expert judges rate a 600-error sample. We found high accuracy for some error types (e.g., ill-formed verbs) and low accuracy for others (e.g., those related to article usage). The second study focused on the utilization inference, which assumes that learners can use AWE feedback effectively in improving their written work. The students' submissions to Criterion were used to develop an error-correction task featuring multiple instances of the 10 common error types. In addition to correcting the errors, participants were also asked to rate each item for the amount of mental effort expended in completing it. The task scores and mental effort ratings were then used to calculate a coefficient of instructional efficiency (van Gog & Paas, 2008). Learners in the lower-level course were found to be more efficient in using the Criterion feedback, possibly because of the way AWE had been incorporated into their course. Overall, the findings provide neither clear support nor refutation for the interpretive argument but instead offer insights into the conditions under which it might apply.

Sun-Young Shin
Indiana University

Investigating the rater bias patterns in reading-to-write tasks using FACETS

Reading-to-write tasks have recently become more common in university settings as more valid and authentic means to assess academic writing skills than a traditional, impromptu composition test (Plakans, 2009; Weigle, 2004). Nevertheless, little is still known about a rater's severity or leniency pattern in reading-to-write tasks. In this vein, this paper aims to explore rater severity and consistency, and scoring domain difficulty and consistency along with rater bias patterns towards examinee levels and domains in the reading-to-write tasks using FACETS. In the study, six experienced raters scored 83 ESL learners' written responses to two texts about the same topic, but with different points of view, using analytic rating rubrics comprising five domains: 1) Recognition; 2) Text Engagement; 3) Organization; 4) Development; and 5) Language Use. Findings revealed that raters were not equally severe despite their high inter-rater reliability, and "Recognition" category was found to be the most harshly rated compared to other scoring domains. Additionally, raters tended to score more harshly high-level test takers and more leniently low-level ones, which is in line with the findings of previous research on rater bias patterns in independent writing tasks (Kondo-Brown, 2002; Schaefer, 2008). The analysis also showed that reading-related domains, such as "Recognition" and "Text Engagement", are significantly either harshly or leniently scored by different raters. This study will provide an insight into analytic rating rubric development and rater training issues in high-stakes academic writing tests using reading-to-write tasks.

Shelley Staples
Purdue University

Geoffrey LaFlair
Northern Arizona University

Jesse Egbert
Brigham Young University

Investigating the Multi-faceted Nature of Speaking Performance with a Multivariate Method

Performance data from speaking assessments have the potential to provide useful evidence of the linguistic features that discriminate between different levels of achievement (Iwashita, Brown, McNamara, & O'Hagan, 2008). However, test takers who receive the same score level often produce different lexical, grammatical, and fluency features (Douglas, 1994). It is possible that this lack of consensus exists because we are trying to use individual performance features to explain a multifaceted phenomenon that is best characterized not by individual features but by co-occurrence patterns among them. A method that has been used to investigate co-occurrence patterns of lexico-grammatical features is Biber's Multi-Dimensional (MD) analysis (see, e.g., Biber, 1988; 2006). MD analysis relies on corpus linguistic techniques and factor analysis to reduce a large set of linguistic variables down to a much smaller set of functionally interpretable linguistic factors. The purpose of this study is to investigate relationships between scores awarded to the oral proficiency performances and linguistic factor scores from MD analysis. The corpus in this study comprises 98 transcribed MELAB speaking tests which were scored by trained MELAB examiners using a holistic rubric. We will perform an MD analysis on the rates of occurrence of various linguistic features related to clausal elaboration (e.g., complement clauses and adverbials), vocabulary, and fluency. Bivariate correlations will be used to investigate relationships between speaker performance and linguistic factor scores. Developing a clearer picture of the relationship between performance level and linguistic features has implications for how speaking tasks are designed and assessed.

Payman Vafaee
University of Maryland

Construct Definition: The First Step in Building Validity Arguments — The Case of a Persian Reading Comprehension Test

In the contemporary validity investigation approaches, two types of arguments should be laid out: First, an interpretive argument which entails defining a construct and laying out all the inferences and assumptions that underlie a test score's interpretation and use; second, a validity argument which is developed through a critical analysis of the plausibility of the theoretical rationales and empirical data which support the inferences in the interpretive argument.

The purpose of the current paper is to document the effort by which the theoretical construct of the reading sub-test of a Persian proficiency test was defined. This paper also explains how the defined construct will be used in the process of test construction, and how it is linked to the other inferences in the interpretive argument which is being laid for this test.

Unlike many test development efforts in which either an L2 proficiency level descriptor (e.g., ILR) or a theoretical model is employed as the construct of a test, in the current paper, it is explained that how by adopting a unifying approach towards construct definition, different features of both of these two approaches was utilized. Moreover, by endorsing the interactionist view of construct definition, the contextual features of the TLU domain were incorporated in the construct of this reading test.

The current study will show how building validity arguments is an on-going process which is more complex than typical post-hoc test-data analyses.

Lorena Valmori
Michigan State University

The influence of second language knowledge on written proficiency rating: An investigation into raters' behavior

Recent research (e.g., Carey, Mannell & Dunn, 2011; Winke, Gass & Myford, 2013; Xi & Mollaum, 2011; Zhang & Elder, 2011) has raised concerns about the impact of raters' native-language backgrounds and language familiarity on oral-performance ratings. Building on Winke at al. (2013), in which language familiarity was operationalized as "having learned the test takers' first language (L1)" (p. 231), in this mixed-methods study I investigated whether raters' knowledge of the test takers' L1 influenced the scores raters gave when rating L2 essays. Sixteen trained novice raters with a second language (L2) of Mandarin Chinese or Arabic rated 20 English essays written by 10 learners each from these two L1 backgrounds. Afterwards, in individual interviews raters looked at 4 sample essays (2 Chinese L1, 2 Arabic L1) and identified the authors' L1s. The raters discussed essays' structural and discourse aspects. Mixed-effects factorial ANOVAs on total and analytic scores showed that raters did not exert more leniency toward test takers with a familiar L1. However, the statistical main effect of test takers' L1 showed that the two groups of essays were rated differently. Raters overall gave higher scores to Chinese writers. Analysis of interview data through recursive readings (Richards, 2003) revealed that when raters could identify the writers' L1 through structural and discourse cues, they understood more. I discuss how the mechanisms triggered by language familiarity (e.g., better understanding, empathy, positive attitude) may influence raters' attitudes, compromising test reliability. I conclude that these issues should be addressed in rater training.

Linxiao Wang
Northern Arizona University

L2 Learners' Use of Interaction Features in Different Paired Speaking Tasks

Paired speaking tasks elicit test-takers' performance while communicating with others using various interaction features (Brooks, 2009; Csépes, 2009; Dimitrova-Galaczi, 2013). However, little is known about how to categorize interaction features, how they are related to performance, or how they are affected by task types (Gan, 2010; May, 2010; Nakatsuhara, 2004). This paper presentation will report on a study that collected responses of 35 pairs of intermediate adult ESL learners on four paired speaking tasks: spot the difference, complete a story, decision-making, and free discussion. All interactions were transcribed and coded for 19 features in four hypothesized interaction categories: interactive listening, topic management, turn-taking management, and the use of questions. The interactions were also scored on an interaction scale. The data is being analyzed this summer using factor analysis and multiple regression analysis. Drawing on the findings in two pilot studies, the researcher expects that results from factor analysis will support the hypothesized model of interaction features. Furthermore, it is anticipated that results from the multiple regression analysis will suggest that features of turn-taking management and using questions are effective predicators of interactive task scores. Finally, the researcher expects that the distribution of interaction features will differ by task types. The presentation will conclude with implications of using interaction features as indicators for L2 interactive task performance.

Xun Yan
Purdue University

Yukiko Maeda
Purdue University

Jing Lü
Purdue University

Task Features and Discriminant Validity of Elicited Imitation in L2 Research: A Systematic Review

Elicited imitation (EI) has been a popular measure of L2 proficiency, especially in the 1970s and 1980s. However, there has been debate over its psychometric properties, specifically issues related to validity of EI scores. To avoid the construct-irrelevant variance in EI scores, scholars have recommended control over four EI tasks features: stimuli sentence length, delayed repetition, grammatical features of the stimuli, and scoring method.

This systematic review examines the control of selected EI task features and/in relation to the discriminant validity of EI scores in L2 research. We first synthesized 72 published and unpublished studies, in the period of 1970-2013, to investigate the status quo of the use of EI regarding the research/assessment context (i.e., targeted construct, language, and proficiency levels). Next, we conducted a meta-analysis of 10 EI group-comparison studies (with 13 independent effect sizes) to examine the discriminating power of EI, by observing effect sizes that represent mean score differences on EI tasks across L2 learners of different proficiency levels.

Overall, there was a great degree of variation in the control over the four characteristics of EI tasks across studies. However, EI tasks in general showed a strong ability to discriminate learners across L2 proficiency levels. The weighted average effect size for EI mean score differences was 1.39, with a large variance in the magnitudes across studies. Moderating effect of three EI tasks features on the discriminating power of EI is also discussed. Findings of this study inform appropriate development and use of EI tasks in L2 assessment.

POSTER ABSTRACTS

Edward Cormany
CaMLA

Rachele Stucker
CaMLA

Mark Chapman
CaMLA

Ummehaany Jameel
CaMLA

One Task Fits All? Constructing a Writing Assessment for a Wide Range of Abilities

This work reports the task design process of the Michigan English Test (MET) writing section. The MET targets the A2–C1 levels of the CEFR (Council of Europe, 2001), but it is challenging to assess writing across this range with a single task. An essay task alone cannot differentiate levels below B1, since A2-level writers can only produce "phrases and sentences linked with simple connectors" (Council of Europe, 2001: 61). Conversely, the task must simultaneously target low- and high-proficiency writing goals to "[allow] better writers to demonstrate their best writing" (Weigle, 2002: 90).

Two tasks, each combined with a traditional essay, were piloted. Task A asked test-takers to recall a life experience, express an opinion about it, and elaborate on the situation in chunks of 1–3 sentences. Task B presented two photos and asked test-takers to compare them and state which they prefer. We show that responses to task B demonstrated far less syntactic complexity and propositional content, even for writers who demonstrated high ability on the essay task. Since these flaws in Task B show no correlation with writer L1 or overall ability, Task A was selected.

Task A elicits sentence-level language reflecting the writer's ability, from beginner to advanced. The elicited language — despite being produced in stages — reads like a single, connected text, even for low-level writers. Finally, Task A prompts are simple to produce and can cover a wide range of topics familiar to test-takers of diverse backgrounds.

Hyun-Sook Kang
Illinois State University

Teacher Candidates' Evaluation of English-as-a-Second-Language (ESL) Writing

This study investigated whether the cultural/ethnic identity of a writer and error gravity would play a role in secondary-level teacher candidates' evaluation of ESL writing, using a matched-guise protocol. A one-page essay about transportation in a foreign city was elicited from an ESL learner enrolled in an intensive ESL program, and was manipulated in spelling, word order, and place names to lead teacher candidates to believe it was produced by an ESL learner whose first language is either Spanish or Chinese. One-hundred fifteen education majors (42 male, 73 female: 110 White, 1 black, 2 Hispanic, 1 Native American, 1 interracial) at a U.S. Midwestern university participated. All the participants, except for one who reported to have learned English at the age of three, reported that English is their first/dominant language. Nearly half of the participants (N=60) were made to believe the writing sample was produced by a Spanish-speaking ESL learner and the other half (N=55) were led to believe it was by a Chinese-speaking ESL learner. The participants evaluated the writing holistically on a scale from 1 to 10 and were invited to correct the five most troublesome errors and make suggestions about how to improve writing for ESL learners. Results show that the ethnic guise did not influence holistic scores or the types of feedback. The teacher candidates tended to suggest reading an essay out loud and reading the English literature as ways to improve writing skills for ESL learners, which does not necessarily reflect ESL learners' needs.

Abdolvahab Khademi
University of Massachusetts-Amherst

Akram Nayernia
Iran University of Science and Technology

Exploring the Content Structure of IELTS Speaking Topics: A Multidimensional Scaling Analysis

A principal argument in test validation is that a test assesses the language ability of individuals independent of general cognitive abilities and other construct-irrelevant factors. Dimensions other than the construct under examination can adversely affect the interpretation of scores and also affect individuals with different attributes that may interact with those construct-irrelevant dimensions. Therefore, ensuring that the test assesses what it purports to assess and inheres an acceptable level of fairness becomes a priority in measurement. The present study attempts to explore the content structure of IELTS Speaking topics for the purpose of investigating different likely structures. Variable performance on a given topic may be attributed not only to the speaking proficiency of the examinees (construct-relevant) but also because of other possible facets in the topics, such as cognitive complexity, task specificity, topical familiarity or culture-specific factors. Multidimensional Scaling provides a powerful statistical tool to explore different dimensions of a stimulus through analyzing the perceived judgments of raters and producing interpretable dimensions. For this study, IELTS speaking topics are randomly selected from official past examination papers. Experienced and trained judges will rate the topics on a Likert scale on the basis of similarity of topics to develop a proximity space. MDS will be applied on the data to identify possible dimensions in the data space. It is hypothesized that different topics will not scatter or cluster differentially. The implications of the study primarily apply to test validation and fairness.

Shinhye Lee
Michigan State University

Irene Ahn
Michigan State University

Ina Choi
Michigan State University

Yaqiong Cui
Michigan State University

Hyung-jo Yoon
Michigan State University

Paula Winke
Michigan State University

A Validation Study of the Reading Section of Young Learners Tests of English (YLTE)

We investigated the validity of the CaMLA Bronze and Silver Young Learners Tests of English (YLTE) reading tests. Seven native speakers of English and 12 learners of English (with Korean or Chinese native languages), all ages 7 through 9, took the tests (Bronze first, Silver second, with a break in between). We videotaped the children as they took the tests, had each child draw a picture of how he or she felt during each test, and asked them questions about the tests. We used Excel and Nvivo to analyze the data. For the Bronze test, the average item facility was .802 (.880 for the native speakers, and .757 for the non-native speakers), and the average item discrimination was .123 (with the cut score at the native, non-native-speaker divide). For the Sliver test, facility was .718 (.858 for the native speakers, .641 for the non-natives); discrimination was .217. Given the relative ease of the tests for our participants, the tests appear to be reliable and consistent in discriminating learners from native speakers. However, analyses of the responses show that some items were more difficult for native speakers than for learners of English. We showcase those items and use the interview data to give explanations as to why the items were inversely discriminating. We summarize issues in layout and formatting that were problematic for the children. We conclude by discussing how and why children are not ideal test takers: they sometimes think and perform differently than expected.

Ryan Lidster
Indiana University-Bloomington

Consequences of Violating the Monotonicity Assumption in Dichotomous IRT Models

The development of Item Response Theory (IRT) has enabled transformational advances to language testing theory and practice, but the utility of IRT models depends on whether or not the assumptions of the model are met (Orlando & Thissen, 2003). The monotonicity assumption—that the probability of a correct response is a non-decreasing function of ability level—is a fundamental assumption of the most commonly used IRT models. However, there is considerable empirical evidence from longitudinal and cross-sectional SLA research suggesting that non-monotonic development, often called "U-shaped" growth, is commonplace across many linguistic domains (Siegler, 2004). Additionally, research from artificial intelligence and cognitive science suggests that non-monotonic growth may be inevitable and even desirable in abductive reasoning conditions such as language learning (Carlucci & Case, 2013).

In 2012, Wang demonstrated that it was in principle possible to estimate non-monotonic growth curves reliably using non-parametric IRT methods, and that non-monotonic curves probably exist in extant test data. To date, however, it is unknown whether such deviations from monotonicity actually affect test-taker ability estimates or item parameter recovery. Using simulated data covering 26 conditions, this poster explores the consequences of different amounts, types, and distributions of non-monotonic items. Results indicate that underlying non-monotonicity causes biased ability and item parameter estimates, but the nature of the errors varies by condition depending primarily on the spread of non-monotonic items throughout the latent trait continuum. Moreover, item fit indices only inconsistently identify non-monotonic items. Recommendations for possible corrections and future research are discussed.

Shannon McCrocklin
Iowa State University

Zhi Li
Iowa State University

Test stakeholders' perceptions of a post-placement assessment in ESL reading classes

This study aims to investigate test stakeholder's perception of a post-placement assessment in English as second language (ESL) reading classes at a large Midwestern university in the U.S. The post-placement assessment, locally called "diagnostic test", is an in-class reading test used to double check the placement decisions based on the test scores on the English Placement Test at the university. The current diagnostic test, however, only provides a composite score as a proof for ESL course waivers. A qualitative approach was employed in this study. 81 ESL students in the reading courses responded to an online survey in the spring semester, 2014. A focus group interview was conducted with five ESL reading teachers to explore their views on the test. The survey responses were analyzed with descriptive statistics and the focus group interview was transcribed and coded inductively. It is found that overall both ESL teachers and students had a positive view of the diagnostic test because it helped identify potentially misplaced students. The score information had not been fully used by students or teachers to help diagnose reading strengths or weaknesses. The ESL teachers also expressed their concerns about the task authenticity and construct representiveness in the test and shared their suggestions on how to improve the test and make it a real diagnostic test. The qualitative findings will help develop a new diagnostic test featuring more authentic reading tasks, better match with the course objectives, and more effective profiling of students' strengths and weaknesses in score reports.

Renka Ohta
University of Iowa

Jui-Teng Liao
University of Iowa

Han Gil Kim
The Ohio State University

Examining the English reading section of the college entrance exams in East Asian countries

East Asian countries tend to require high-stakes college entrance exams to measure test takers' English competence for academic and communicative purposes. The results of the college entrance exams are usually required for university admission. Across East Asian countries, English is a required subject for elementary and secondary school students (Warden & Lin, 2000). The national curriculum and testing policies have brought about competitive and norm-referenced tests. While the English section of the college entrance exams have similar purposes, the exams are not easily comparable.

This study analyzes the difficulty levels of English reading passages of the college entrance exams in East Asian countries, including South Korea, Taiwan, and Japan, using Coh-Metrix (Graesser, McNamara, Louwerse &Cai, 2004). Coh-Metrix is a computer-based tool that examines text cohesion, discourse characteristics, and text difficulty levels. This study highlights the English reading section of the college entrance exams because reading performance is especially crucial in academic success as printed materials, such as textbooks and articles, are overtly used in higher education (Stoynoff, 1997). The research questions guiding this study are: 1) What are the profiles of reading passages in these tests in terms of text cohesion and difficulty? 2) How do text cohesion and reading passage difficulty on college entrance exams in East Asian countries differ? Based on the results of the Coh-Metrix analysis, implications and suggestions are discussed.

Victor D.O. Santos
Iowa State University

Towards a Computer-Adaptive Rasch-Based Test of Productive Academic Vocabulary Knowledge (AVK)

Knowledge of individual lexical items is perhaps the most important aspect of learning a foreign language (Richards, 2000). Although knowledge of general English vocabulary is an essential component of communicative success in English-medium universities, it might not suffice. Knowledge of vocabulary that is more specific to the academic target language use domain becomes necessary for successful communication and performance (Biemiller, 2010; Nagy & Townsend, 2012).

This poster makes use of Gardner and Davies' (2013) new core academic vocabulary list, which contains the 3,000 most frequent lemmas in the academic subset of COCA (Corpus of Contemporary American English) in order to answer the following questions: (1) Do more frequent academic words tend to be learned before less frequent academic words? (2) Does the relationship between Rasch difficulty and word frequency differ according to part of speech? (3) Can a 30-item productive AVK (Academic Vocabulary Test) test be psychometrically robust? (4) Can the results of the AVK test be used as an indication of general proficiency in English, as well as an indication of academic vocabulary proficiency? Results from a pilot sample of 48 test-takers at a Midwestern university in the USA indicate that the test has the potential to achieve high reliability (current alpha coeff. at 0.73), is uni-dimensional, and that there is a high correlation (currently at 0.69) between word frequency and difficulty of recall. Current limitations, implications, and future steps towards a computer-adaptive version of the AVK test will be discussed in this the poster.

Nancy Ngan Vu
Iowa State University

Rater's Perceptions of Evaluative Criteria in an English Writing Placement Test

In writing assessment, much research has been conducted into rater cognition involving in performance evaluations (e.g., differential focuses on performance features), focusing on decision-making behaviors of experienced raters when they grade ESL/EFL students' essays. However, little research into human rater's grading perceptions of scoring criteria for placement purposes has left a gap in the literature. The purpose of this study is to initially describe raters' perceptions of evaluative criteria that are taken into account in their decision-making processes in the context of Writing Section of an English Placement Test at a Midwest university (EPT Writing). Three experienced raters engaged in one-on-one semi-structure interviews, providing qualitative data for this study. The findings of this qualitative study revealed that the criteria the raters considered fell into criteria categories in the scoring rubric, except for the Mechanics category. Furthermore, during their decision-making processes, the raters employed different cognitive procedures with various criteria across the performance levels. However, the raters reached a consensus on decision-making criteria in general at each level, even though it remained controversial in their perceptual differentiation by writing quality (e.g., good/bad, serious/minor). Regarding influential factors in raters' placement decisions, rater training, the scoring rubric and teaching experiences were the most significant, suggesting correspondence of the ESL writing courses objectives and the EPT scoring rubric. Limitations of the study and outlook for future research will be discussed.

Junli Wei
University of Illinois at Urbana-Champaign

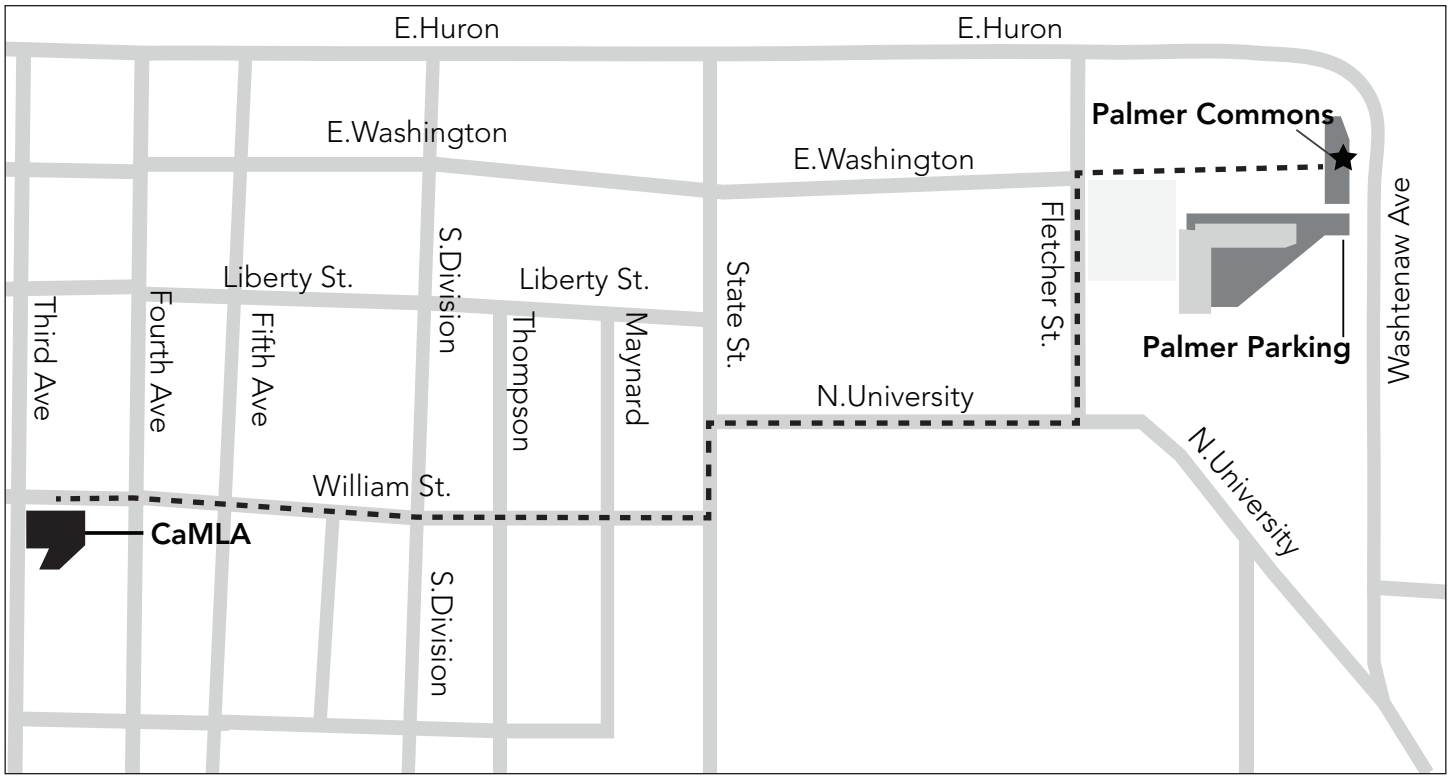
Beyond Assessment Literacy: ESL Teachers' Practices on Constructing Cognitive Diagnostic Assessment

Classroom assessment is both a cause of learning and a measure of effective learning. To certify students' learning and to promote that learning, teachers should develop high quality classroom assessments and gather accurate information about student achievement for making instructional decisions. Teachers thus should be able to develop cognitive diagnostic assessment (CDA) that provides information with more diagnostic value to identify individual students' strengths and weaknesses. Despite the importance of CDA in educational testing, its most current applications rely on a post-hoc approach to test design where exiting items are coded for cognitive attributes and then analyzed (Gorin, 2007). To make CDA apply to classroom assessments more meaningfully, the inherent limitations associated with retrofitting should be overcome. This study aims to fill in the gap by designing a CDA from an explicit cognitive framework at the beginning. Ten ESL teachers' survey responses on assessment (e.g. knowledge, self-efficacy) and teacher-made tests will be collected and analyzed. Factors affecting the test quality and the areas that teacher need to be trained will be identified. A workshop on CDA-focused test development will be provided to help the teachers understand the basic elements of CDA. Furthermore, the teachers, content experts and statisticians will collaboratively work together to develop a CDA and keep refining it. Finally, ESL students will take the refined CDA and their responses will be analyzed for evaluating the quality of the CDA. The CDA test development practices described here will have strong implications on ESL classroom assessment and teacher training.

MAPS

Maps

CaMLA to Palmer Commons



CaMLA Office
Argus 1 Building
535 West William St., Suite 310

Palmer Commons
100 Washtenaw Avenue

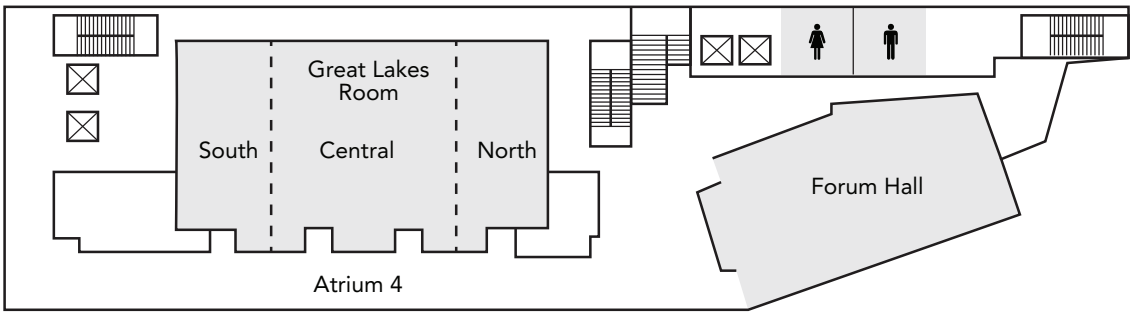
Need Help?
Lost your way? Just not sure? Phone or text conference organizers Mark or Jessica:

- 734-255-7580 (Mark) • 407-491-1188 (Jessica)

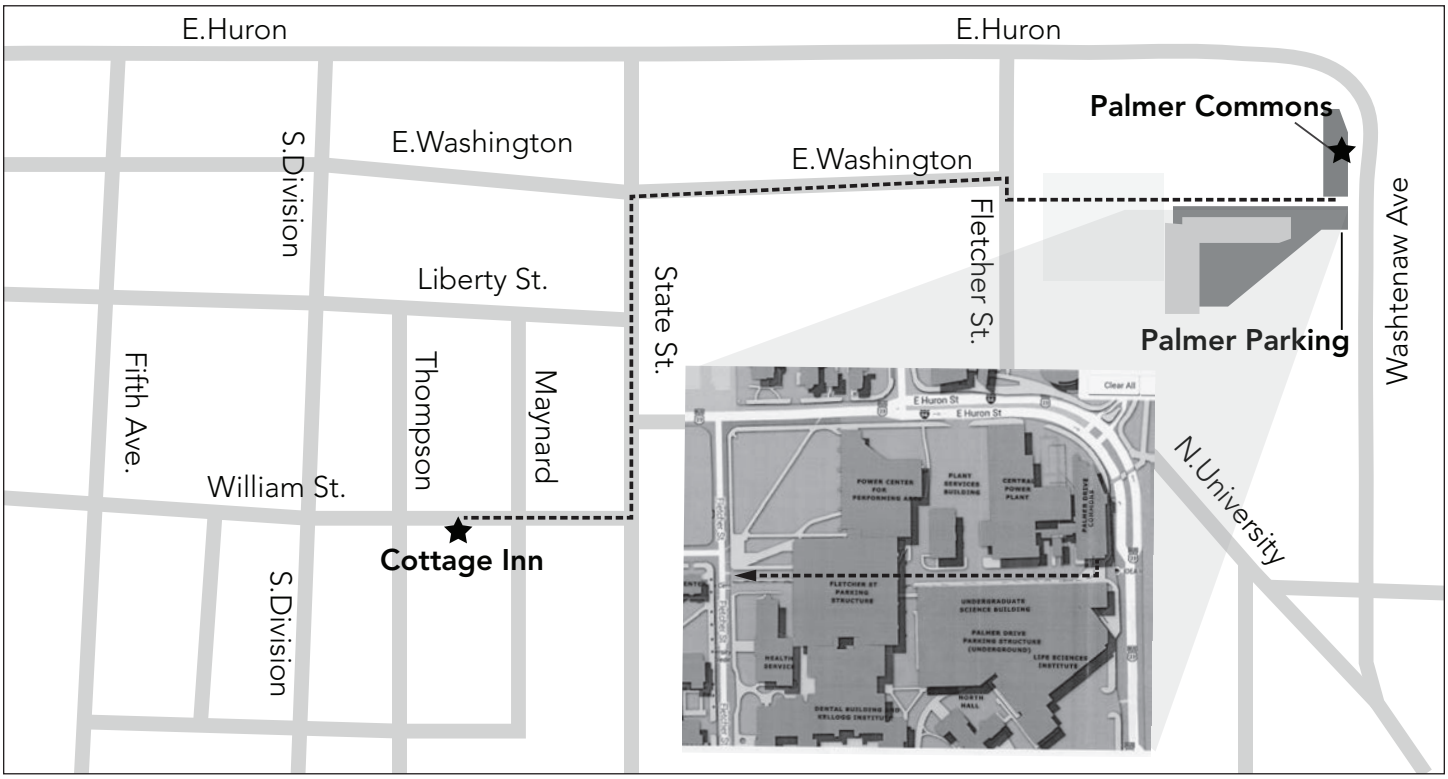
Walking Directions: CaMLA to Palmer Commons
approx 25 minutes

1. From the parking lot, turn left and walk towards William St.
2. Turn right onto W. William St. and keep walking (*about 10–15 minutes*)
3. Turn left on State toward N. University Ave
4. Turn right onto N. University Ave
5. Turn left onto Fletcher St. and continue straight toward E. Washington St.
6. Walk up the ramp (on right) to cross the Palmer Commons bridge.
7. Turn left into Palmer Commons.

Palmer Commons 4th Floor



Palmer Commons to Cottage Inn Restaurant



Palmer Commons
100 Washtenaw Avenue

Cottage Inn Pizza
512 E. William St.

Need Help?
Lost your way? Just not sure? Phone or text conference organizers Mark or Jessica:

- 734-255-7580 (Mark) • 407-491-1188 (Jessica)

Walking Directions: Palmer Commons to Cottage Inn
approx 12 minutes

1. Take third floor Palmer Commons bridge to Fletcher St.
2. Cross Fletcher and continue onto E. Washington St.
3. Turn left on State St.
4. Continue straight 2 blocks to William St.
5. Turn right onto William St.
6. Cottage Inn will be on your left.

#mwalt14

LSA ENGLISH LANGUAGE INSTITUTE
UNIVERSITY OF MICHIGAN

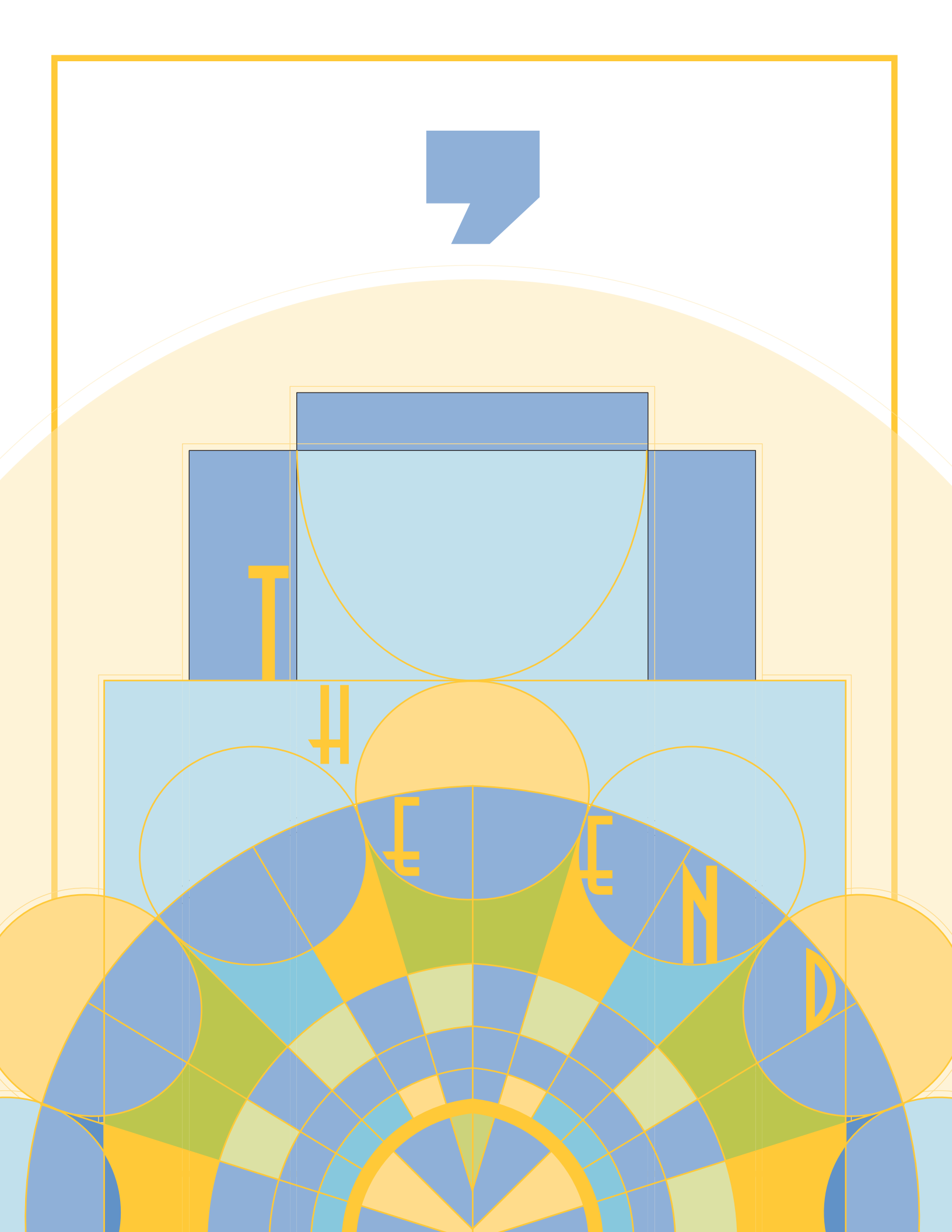
Serving International Students and Scholars at the University of Michigan since 1941

lsa.umich.edu/eli

INDEX

Index

MY NOTES


THE END