Midwest Association of Language Testers
18th Annual Conference

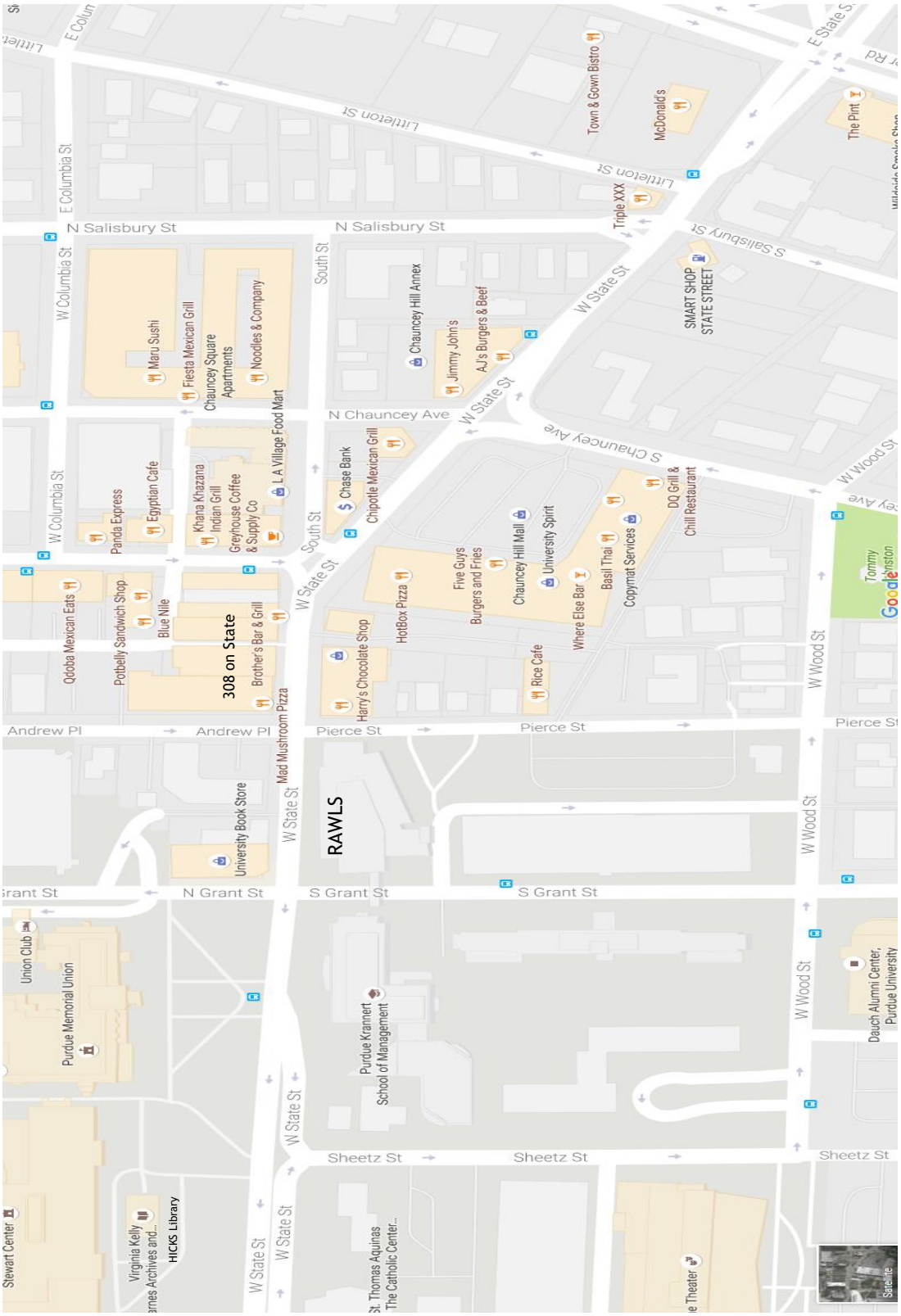
September 30 – October 1, 2016

Purdue University

Oral English Proficiency Program

West Lafayette, Indiana

Purdue/Chauncey Village area



Welcome to the 2016 Meeting of the

Midwest Association of Language Testers

Hosted by the

Oral English Proficiency Program

Purdue University

West Lafayette, Indiana

We thank the following for their generous support

of this year's conference:

College of Liberal Arts, Purdue University

Office of International Students & Scholars, Purdue University

Purdue Language and Cultural Exchange (PLaCE)

Detailed Conference Schedule			
Friday, September 30			
12:00 - 4:45 p.m.	**Registration (Rawls Lobby)**		
1:00 - 4:00 p.m.	**Workshop** – *Cluster Analysis in Language Testing Research Dr. Xun Yan, University of Illinois* (pre-registration required) Hicks G959		
5:00 - 6:00 p.m.	**Welcome Reception – 308 on State.** (308 State Street, 1 block from Rawls Hall)		
Saturday, October 1			
7:30 - 11:45 a.m.	**Registration (Rawls Lobby)**		
8:30 - 8:45 a.m.	**Welcome and Introduction (Rawls 1086)**		
8:45 - 9:45 a.m. Concurrent Sessions	***Session 1 (Rawls 1086)*** **Paula Winke, Koen Van Gorp, Susan Gass, & Bill VanPatten (Michigan State University)** Using Different Carrots: How Incentivization Affects Proficiency-Testing Outcomes **Xiaowan Zhang (Michigan State University)** The Role of Released Test-Specifications in Test Preparation	***Session 2 (Rawls 1062)*** **Suthathip Thirakunkovit (Mahidol University, Thailand)** The Use of C-Test and Cloze-Elide in a Post-Entry Test **Jo-Kate Collier & Becky Huang (University of Texas at San Antonio)** Investigating the Validity of TELPAS as Predictor of STAAR Success	***Session 3 (Rawls 1071)*** **Kyongson Park (Purdue University)** University Policy for International Students: Self-Assessment on Campus Adaptation **Sharon Pearce & Stephen O'Connell (Cambridge Michigan Language Assessments)** The CEFR in Practice: Defining the "Just Qualified" Speaker
9:45 - 10:00 a.m.	**Break (Rawls Lobby)**		
10:00 - 11:30 a.m. Concurrent Sessions	***Session 4 (Rawls 1086)*** **YunDeok Choi & Sonca Vo (Iowa State University)** Approaches in Recent Validation Research: A Still Evolving Story **Ahmet Dursun, Catherine Baumann, James McCormick, Nicholas Swinehart, & Jason Merchant (University of Chicago)** Building a Validity Argument for a Graduate-Level L2 Reading Comprehension Exam: A Case of Domain Analysis	***Session 5 (Rawls 1062)*** **Laura Ballard (Michigan State University)** Young ELLs and Computerized Reading Assessment: Are Kids Ready? **Wayne E. Wright, Trish Morita-Mullaney, Rudy Rico, Marquetta Straight, Chen Li (Purdue University)** Introducing the Purdue English Language Learners Language Portraits (Purdue ELLLPs)	***Session 6 (Rawls 1071)*** **Lixia Cheng (Purdue University)** Pragmatic Appropriateness in an ESL/EFL Oral Discourse Production Test **F. Scott Walters (University of Illinois at Urbana-Champaign)** Widening the Scope of CA-Informed L2 Pragmatics Testing **Aysenur Sagdic (Indiana University)** The Effects of Task Mode on Assessing Pragmatic Inferential Skills

	Elena Cotos & Yooree Chung (Iowa State University) Investigating Functional Language Use for Domain Description	**Jui-Teng Liao (University of Iowa)** Effects of Testing Format on Second Language Reading Performance	
11:30 - 11:45 a.m.	**Break (Rawls Lobby)**		
11:45 a.m. – 12:45 p.m.	***Plenary Panel Discussion (Rawls 1086)*** ***Selection Policy and Practice for International Admissions*** • Panelists: Mike Brzezinski, Dean, International Students and Scholars, Purdue University • Panelists: Thomas Atkinson, Associate Dean of the Graduate School, Purdue University • Moderator: April Ginther, Director, Oral English Proficiency Program & Purdue Language and Cultural Exchange, Purdue University		
12:45 - 1:45 p.m.	**Lunch (Rawls Lobby)**		
1:30 - 2:30 p.m.	***Posters (Rawls Lobby)*** • Mariam Alamyar, Purdue University: An Interpretative and Comparative Review on Writing Assessment in First and Second Language • Matthew Allen, Purdue University: Measuring ESL Reading Fluency Development with Assisted Repeated Reading • Guangyan Chen, Texas Christian University: Is it Fair for English-Speaking Learners in the U.S. to Take the New HSKs? • Jean Young Chun & Claire Fields, Indiana University: Use of Hedges and Intensifiers by NNS in Oral Discussion • Jung Han & Hyun Jin Cho, Purdue University: Challenges of Elementary School ELLs in Mainstream Classes and Effective Use of Classroom-based Assessments of ELLs • Hyun-Ju Kim, Danook University, S. Korea: Implementation of Criterion-Referenced College Scholastic Ability Test in Korea • Susie Kim & Ok-Sook Park, Michigan State University: Validation of a Korean Speaking Proficiency Assessment • SungAe Kim, Purdue University: Variables that Affect English Reading Difficulty for English Language Learners • Xiaorui Li, Purdue University: What does *it* Mean? Predicting Oral Proficiency by Using Referential *it* • Yi Li, Southwest University, China: Automated Essay Feedback Generation and Its Impact on EFL Writing Revision • Mayu Miyamoto, Purdue University: A Scale Development Project for Performance-Based Test (PBT) • GoMee Park, University of Iowa: Use of TOEIC: de facto Language Policy in South Korea • Sockwun Phng, Iowa State University: Man/Machine: Human and Computer Raters in L2 Oral Assessments • Ji-young Shin, Purdue University: The Use of "Actually" in Korean College Aptitude Scholastic Test • Evan Simpson, Iowa State University: Shifting Baselines: How Changing from NRT to CRT Affects Placement • Magda Tigchelaar, Michigan State University: Using Self-Assessments to Predict Spoken French Proficiency • Taichi Yamashita, Iowa State University: The Significance of Triangulation in Written Corrective Feedback Research		
1:45 - 2:30 p.m.	**Business Meeting (Rawls 2077)**		

2:30 - 4:00 p.m. Concurrent Sessions	***Session 7 (Rawls 1086)*** **Saerhim Oh (Teachers College, Columbia University)** Investigating L2 Learners' Use of Linguistic Tools in an Online Writing Test **Lia Plakans & Atta Gebril (University of Iowa & American University in Cairo)** Integration in Academic Assessment: Lexical Diversity, Textual Borrowing and Proficiency **Phuong Nguyen (Iowa State University)** Linguistic Analysis of Essays from Integrated Writing Tasks	***Session 8 (Rawls 1062)*** **Aleksandra Swatek & Aleksandra Kasztalska (Purdue University & Southern Arkansas University)** "Write an Email to Your Friend in England": Assessment of the Written English Matura Exam in Poland **Rongchan Lin (Teachers College, Columbia University)** A Multi-Contextual Approach to Investigating Chinese Language Assessment in Singapore **Ananda Muhammad (Iowa State University of Science and Technology)** Alignment of Target Language Use and the High School English National Examination in Indonesia	***Session 9 (Rawls 1071)*** **Ha Ram Kim (University of Illinois at Urbana-Champaign)** Towards a Profile-Based Rating Scale for Post-Admission Writing Placement Tests **Senyung Lee (Indiana University)** A Data-Driven Rating Scale for Assessing Coherence in L2 Writing **Reuben Vyn (The University of Iowa)** A Validity Argument for a Foreign Language Writing Performance Assessment Rubric
4:00 - 4:15 p.m.	**Break (Rawls Lobby)**		
4:15 - 5:15 p.m. Concurrent Sessions	***Session 10 (Rawls 1086)*** **Ziwei Zhou (Iowa State University)** Exploring the Relationship Between Fluency and Speaking Performance of ITAs **Daniel Isbell (Michigan State University)** Assessing Comprehensibility and Accentedness in L2 Korean Speech	***Session 11 (Rawls 1062)*** **Elnaz Kia & Valeriia Bogorevich (Northern Arizona University)** Linking Rating Behavior to Criteria: Evaluation of Paired Speaking Task **Anna Mikhaylova (University of Iowa)** The Effects of French Teaching Assistants' Backgrounds on Writing Assessment	***Session 12 (Rawls 1071)*** **Jie Gao (Purdue University)** A Comparison of Two Automatic Writing Evaluation Tools in China: Friendly to Use and Easy to Accept? **Virginia David (Western Michigan University)** Timed-Writing and Process-Based Writing Exams: Comparing Learners' Performance and Perceptions
5:15 - 5:30 p.m.	**Break (Rawls Lobby)**		
5:30 - 5:45 p.m.	**Best Presentation & Best Student Paper Awards & Closing (Rawls 1086)**		

Pre-Conference Workshop, September 30, 2016
Hicks Undergraduate Library, Room G959
(Pre-Registration Required)

Cluster Analysis in Language Testing Research

In both language testing research and real practice of test use, there has been a growing interest in examining subscale scores and score profiles of language learners. This workshop will introduce cluster analysis as a useful statistical procedure to identify language score/skill profiles and explore their relationships with language learning and academic performance. Workshop participants will learn about the basic principles underlying cluster analysis and have the opportunity to apply cluster analytic procedures on language testing data for admission and placement considerations.

During the workshop, we will first discuss when, what, how, and why to cluster language test scores and testing data in general. Then, we will use cluster analysis to identify score/skill profiles of language learners from different test data sets. We will also examine the relationships among subscale scores, score profiles, and other criterion measures of language learning and academic achievement, to better understand how cluster analysis can be used to assist admission policies, placement decisions, and language pedagogy in institutional contexts. As such, the workshop will mainly cover the following components:

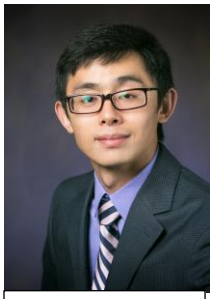
Part I: Basic concepts

Language score/skill profiles
Statistical principles underlying cluster analysis

Part II: Data analysis

Correlational analysis
Cluster analysis for admission policies
Cluster analysis for placement decisions

This workshop will be offered in a computer lab at Purdue University. We will use SPSS to perform the correlational and cluster analyses. Participants are expected to have some understanding of basic statistics, although this is not required.



Dr. Xun Yan

Presented By: Dr. Xun Yan, University of Illinois , Supervisor, English Placement Test (EPT); Assistant Professor of Linguistics; Affiliated Faculty, Asst. Professor of Second Language Acquisition and Teacher Education Dr. Yan's research interests include the development and quality control of post-admission language assessments, assessment literacy for language teachers, formulaic language acquisition and lexical development by L2 speakers, L2 pronunciation and intelligibility, and test score use in educational settings. Education: Ph.D., English, Purdue University; M.A., TESOL, The Ohio State University; B.A., English, Wuhan University, Wuhan, China

Plenary Panel Discussion (Rawls 1086)

Selection Policy and Practice for International Admissions

Panelists: Mike Brzezinski, Dean, International Students and Scholars, Purdue University
Thomas Atkinson, Associate Dean of the Graduate School, Purdue University
Moderator: April Ginther, Director, Oral English Proficiency Program & Purdue Language and Cultural Exchange, Purdue University

Dr. Brzezinski began his career at Purdue University in 1993. In addition to directing the International Students and Scholars office for 16 years, Brzezinski has led International Programs as dean since 2009. Under his leadership 9,300 international students from more than 125 countries are now enrolled. A new study abroad initiative has increased student participation rates to their highest levels ever. Nearly 2,200 Purdue students study around the world on an annual basis. Furthermore, a new intercultural learning initiative has been established focused on promoting and assessing intercultural learning that occurs during overseas study experiences as well as during co-curricular activities both on and off campus. Under Brzezinski's leadership, a student integration subunit was created to promote learning between domestic and international students.

Tom Atkinson is Associate Dean of the Graduate School at Purdue University. A native Hoosier, Atkinson received B.S. and M.S. degrees from Purdue and a Ph.D. degree from University of Pennsylvania. Before coming to the Graduate School in 2001, he served for ten years as Assistant to the Director of Academic Programs in Purdue's College of Agriculture. At the Graduate School, Atkinson provides leadership for recruitment, admissions, and student records, serves as ombuds, and advises the Purdue Graduate Student Government. He also serves as the Graduate School's International Programs Officer.

Graduate Student Award

Saerhim Oh is a doctoral candidate in the TESOL & Applied Linguistics Program at Teachers College (TC), Columbia University specializing in second language assessment under the supervision of Dr. James E. Purpura. Her current research focuses specifically on the use of linguistic resources in second language writing assessment. Her work on language assessment has been presented at professional conferences such as the Language Testing Research Colloquium, the East Coast Organization of Language Testers, the Second Language Research Forum, and the Symposium on Second Language Writing. She has also been working as a TC-ETS doctoral graduate fellow for three years. Besides research, Oh also has taught ESL/EFL learners for more than 10 years at English language programs at the University of Hawaii at Manoa, Seoul National University, and Columbia University. She has also taught Second Language Assessment courses at TC and at Hunter College, City University of New York. In addition to teaching, she is also an English test item writer and a grammar and reading textbook writer.

ABSTRACTS

SESSION 1 (RAWLS 1086)
8:45-9:45 A.M.

Paula Winke, Koen Van Gorp, Susan Gass, & Bill VanPatten (Michigan State University)
Using Different Carrots: How Incentivization Affects Proficiency-Testing Outcomes

In both teaching and testing for proficiency, an important issue in interpreting outcomes is the motivation that students have for achieving any particular level of ability. In most university settings, the only language requirement for students is seat time. At our university, the Spanish program has a stated outcome goal of Intermediate-Mid in speaking, listening, and reading on the ACTFL proficiency scale, although the goal is not a requirement. In investigating a larger question (Could we require intermediate-mid?); we investigated the proficiency-test results from 257 students completing second-year Spanish courses at our university. The students took the ACTFL OPIc and the ACTFL listening and reading tests in proctored computer labs. We incentivized two groups of students: for one group (n = 96) 10% of their final grade was based on meeting the pre-established program goals: for a second group (n = 73), 5% of their grade was based on the same criteria. With a comparison group (n = 88), there was no incentivization: they only received 5% for taking the three proficiency tests. We compared the proficiency measures on the standardized listening, reading and speaking tests for the three groups taking into account the effect of gender. Our findings show that the 10% group performed significantly higher than the two other groups and that there was an effect for gender. These results suggest that programs should consider proficiency-based exit-tests if there are requirements tied to them that articulate the programs' stated goals (see Chalhoub-Deville, 1997).

Xiaowan Zhang (Michigan State University)
The Role of Released Test-Specifications in Test Preparation

The role that released test specifications can play during test preparation is often neglected by test takers, policy makers, and even researchers. Focusing on the Test for English Majors-Band 4 (TEM4)—a nationwide, large-scale, high-stakes test for English majors in China—this study investigated the preparation effects associated with the use of the TEM4 Syllabus, that is, its released specifications. Data collection involved 48 test takers of the TEM4 recruited at a large university in central China, where the experimental group was given a tutorial session on the TEM4 Syllabus as the treatment. Specifically, preparation effects were measured using both a quantitative metric of score improvement and a qualitative metric informed by a framework adapted from Messick (1982) and Xie (2013). A variety of statistical procedures, including t-test, multiple linear regression, and expectancy graphing, were employed in the examination of score improvement. The study found that there was no significant score inflation associated with the TEM4 Syllabus, and that the tutorial on the TEM4 Syllabus failed to foster positive behavior changes in test takers from their old teaching-to-the-test practices. Along with the exploration of preparation effects, this paper also discussed the ethicality of different test preparation practices and proposed three possible solutions that test developers or policy makers could adopt to discourage unethical teaching-to-the-test practices from the perspective of specification releasability (Davidson, 2012).

SESSION 2 (RAWLS 1062)
8:45-9:45 A.M.

Suthathip Thirakunkovit (Mahidol University, Thailand)
The Use of C-Test and Cloze-Elide in a Post-Entry Test

In 2013, a large public university in the Midwest of the United States developed a post-entry English language proficiency test named the Assessment of College English-International (ACE-In) for

identifying incoming international students who may benefit from extra language support. This study only focuses on the C-tests and cloze-elide of the ACE-In. This study reports on reliability of these two tasks. The results of 232 international examinees were validated against 28 English native speakers' test performance. The data were analyzed for four traditional item characteristics: 1) Descriptive statistics, 2) Item facility, 3) Item discrimination, and 4) Cronbach's alpha and Pearson's correlation reliability coefficients.

The results of the pilot study showed that C-test items are considered too easy for the test population, while cloze-elide items are of medium difficulty. Even though C-test items have acceptable discrimination i.e., the average rpb index is 0.3, cloze-elide items are shown to have much better discrimination values on average i.e., rpb indices are higher than 0.5. The Pearson product-moment correlation between the C-test and cloze-elide is high (r = .66).

Even though the key results showed that C-test did not meet the standard of item difficulty and discrimination, it does not necessarily mean that C-test cannot sufficiently serve its intended purpose as a preliminary screening tool. After examining the score distributions of both C-test and cloze-elide scores, the scores of both tasks range widely. With fairly wide standard deviations, there is a potential to combine the scores of these two screening tasks to identify students who had a uniformly low performance across both tasks.

Jo-Kate Collier & Becky Huang (University of Texas at San Antonio)
Investigating the Validity of TELPAS as Predictor of STAAR Success

Federal law requires annual assessment and reporting of the English language acquisition of students classified as English Language Learners (ELL), a requirement that will continue under the recently enacted Every Student Succeeds Act (ESSA). To comply with this requirement Texas developed the Texas English Language Proficiency Assessment System (TELPAS). TELPAS was first implemented in the 2007-08 academic year with the intent of complying with the reporting requirement and providing an indicator of a student's ability to succeed on the state assessment, TAKS, in the area of language arts. In 2011-12 Texas implemented a new state assessment, STAAR, developed as a more rigorous assessment program making TELPAS a less than accurate predictor of student success on the new STAAR. Accordingly, TELPAS was revised and first implemented in 2013-14.

This article will present a study which examined assessment data for ELL students over the course of three consecutive academic years from 2013-2016, including the revised TELPAS assessment and STAAR data. The focus of the investigation is the claim by the state of Texas that high scores on TELPAS reading correlate to successful STAAR language arts scores. The project looked at the validity of TELPAS scores as predictors of student success on STAAR language arts. Three groups of students representing three different grade clusters from one urban school district were selected for study. The project used readily available data from district accountability reports and looked for correlations between TELPAS reading scores and STAAR language arts scores for the same students.

SESSION 3 (RAWLS 1071)
8:45-9:45 A.M.

Kyongson Park (Purdue University)
University Policy for International Students: Self-Assessment on Campus Adaptation

Despite an explosion in the number of international undergraduate students on the USA campuses, there is little empirical research into the relationship between academic and social integration of college students in the American universities. Even though international students are satisfied with their academic experiences in general (Lee & Wesche, 2000; Schutz & Richards, 2003; Senyshyn et

al., 2000), we do not have much knowledge about how they think about the influence of social aspects on their academic lives (Lewthwaite, 1996). Therefore, this study aims to fill in this gap in the literature. To investigate the university adaptation, both international and domestic students at a large public university (n=180) are asked to complete a self-assessment survey and three students to participate in a follow-up interview. The findings revealed that international undergraduates have more active interaction with their instructors and advisors than with American peers. To be more specific, unlike the domestic students, the international students pointed out that the current language use and frequency of interaction in English were the main factors that cause or solve the problems in academic and social adaption. Interestingly, Chinese and Korean students showed subtle differences. My study suggests that formal academic settings such as classrooms, peer-group work, EAP courses, and language partner programs at universities could give international college students more chances to interact with other diverse peers and enhance social integration. If the university could incorporate the needs of international undergraduates, social integration could play an important role in academic integration of international students.

Sharon Pearce & Stephen O'Connell (Cambridge Michigan Language Assessments)
The CEFR in Practice: Defining the "Just Qualified" Speaker

Defining the "just qualified" test-taker is a key aspect of successful standard-setting studies (Cizek et al., 2004). Failure to take the just-qualified learner into account can result in cut scores that are too high, which can have serious consequences (Lim et al., 2013; Papageorgiou, 2010). However, despite the importance of a "just qualified" consensus understanding, it is rare to find details in the literature on how this concept is addressed in practice (Kaftandjieva, 2004; Papageorgiou, 2009).

In this paper we first explain how "just qualified" CEFR B1, B2, and C1 English learners were defined in the context of a CEFR linking study conducted for a commercially available speaking test. Then, we discuss how we qualitatively and quantitatively analyzed "just qualified" definitions that were collected both pre- and post-study from 12 judges who participated in the linking study. In two rounds of coding, we tagged our judges' responses for "limiting" language (explicitly negative language) and for "boosting" language (explicitly positive language). Our analyses show that facilitators' repeated focus on the concept of "just qualified" resulted in participants' increased usage of "limiting" language in their post-study "just qualified" definitions and decreased usage of "boosting" language in their post-study "just qualified" definitions (in comparison to pre-study definitions). The implications of an approach focusing on the concept of "just-qualified" will be discussed, as will the positive impact it had on the validity of the cut scores attained (i.e., standard error of judgment values less than a quarter of the test's standard error of measurement).

SESSION 4 (RAWLS 1086)
10:00-11:30 A.M.

YunDeok Choi & Sonca Vo (Iowa State University)
Approaches in Recent Validation Research: A Still Evolving Story

Much validation research for various tests has been done since test results entail values and social consequences on test users. This, by extension, reflects that it is critical to justify interpretations and uses of test scores for their intended purposes (Messick, 1989). As an extension of Chapelle and Voss's (2013) work, the present study examined recent trends in validation approaches: one question and three validities, evidence gathering, test usefulness, argument-based, and not explicit (implicit) in 74 empirical research articles published from 2012 to 2016 in *Language Testing* and *Language Assessment Quarterly*. In addition, different from Chapelle and Voss, we analyzed 30 empirical research reports published during the same time period by two big worldwide testing companies--Educational Testing Service (ETS) and International English Language Testing System (IELTS) regarding the validation

frameworks. It was found that, across the journal articles and research reports, the approaches were taken most frequently in a descending order: implicit and evidence gathering, argument-based, test usefulness, and one question and three validities. Very similar trends were observed in both the journal articles and research papers. These findings are in line with what Chapelle and Voss predicted in that the growing popularity of argument-based approach in current practices in validation is remarkable. This study has implications for a better understanding of the approaches for recent validation studies in the field of language testing in general, and of similarities and dissimilarities between the academic journals versus testing companies with regards to the use of validation framework in particular.

Ahmet Dursun, Catherine Baumann, James McCormick, Nicholas Swinehart, & Jason Merchant (University of Chicago)

Building a Validity Argument for a Graduate-Level L2 Reading Comprehension Exam: A Case of Domain Analysis

To replace its previously-used translation exam, the University of Chicago has recently developed a new graduate-level L2 reading comprehension exam, namely the Graduate Foreign Language Reading Comprehension Exam (GFLRCE), to measure its graduate students' ability to conduct academic research through reading in a secondary research language. Like any other language test scores that are used to make important decisions about students in higher education, this exam needs to be evaluated. The University of Chicago Language Center has utilized the Argument-based approach to validation (Kane, 2006; Chapelle, Enright, & Jamieson, 2008, 2010) as a framework to evaluate the intended uses and interpenetrations of GFLRCE scores. As a transparent research framework, this approach will guide the testing program in "…prioritizing different lines of evidence, synthesizing them to evaluate the strength of a validity argument, and gauging the progress of the validation efforts" (Xi, 2008, p. 18) and thus methodologically will offer a practical guideline to construct a validity argument (Chapelle et al, 2010) for the GFLRCE. This presentation will first present the interpretive argument, which specifies the proposed interpretations and uses of GFLRCE results by laying out a network of inferences and assumptions. Then, in an effort to critically evaluate the appropriateness and plausibility of the assumptions in the domain definition inference, results from graduate students' and faculty survey and interview responses, syllabi from Reading for Research Purposes courses, and published literature on conducting academic research through reading in a secondary language will be presented.

Elena Cotos & Yooree Chung (Iowa State University)

Investigating Functional Language Use for Domain Description

TOEFL iBT® is one of the most globally accepted, high-stakes language proficiency tests for non-native speakers of English. In the United States, there is an increasing tendency to use TOEFL iBT® Speaking scores for decisions regarding the screening or certification of international teaching assistants (ITAs). Therefore, obtaining validity evidence to demonstrate the usefulness of these scores for such purposes is necessary for both policy and practice. The argument-based approach to validation – a theoretical model consisting of a chain of inferences about the interpretations and uses of test scores, propositional warrants associated with inferences, and specific assumptions underlying respective warrants (Chapelle, Enright, & Jamieson, 2008; Kane, 1992) – has been widely employed in validation research. This study adheres to this model in order to investigate an assumption pertaining to the Domain Description inference. Specifically, our inquiry aims to determine whether the language functions elicited by TOEFL iBT® Speaking tasks are representative of language functions used in the target domain. We will report results derived from a linguistic analysis of a TOEFL iBT® speech corpus and of an ITA speech corpus. The analysis was conducted using the Knowledge Framework – a heuristic in Systemic Functional Linguistics that focuses on *doing* in the discourse related to thinking skills (Mohan, 1986). This enabled the identification and definition of language functions used by TOEFL iBT® test-takers and by ITAs in instructional settings. The results also reveal how the functions are realized linguistically and how they differ in instructional discourse and in responses elicited by the speaking tasks.

SESSION 5 (RAWLS 1062)
10:00-11:30 A.M.

Laura Ballard (Michigan State University)
Young ELLs and Computerized Reading Assessment: Are Kids Ready?

Many assessments have moved from paper to computer-based formats, including second-language tests for young learners (e.g. ACCESS 2.0). In response, language testing researchers have conducted comparability studies on these testing modes (Endres, 2012; Kingston, 2009), but there is a lack of research examining mode effects on specific populations: young English language learners (ELLs). The potential effects that emerge from technologically-based assessments are a concern for those who have little experience with technology, and these effects could be aggregated with the cognitive overload that ELLs experience when tested in a second language (Genesse, Linholm-Leary, Saunders, & Christian, 2005). In response, in a larger study to which these data belong, I investigated: 1) whether ELL and native-English-speaking children were able to complete a computer-mediated reading test; and 2) how these children's reading behaviors differed. I used eye-tracking methodology to uncover how children interacted with the computerized assessment. In the current study, I triangulated data by using an open-ended picture-drawing task (Brown & Wang, 2011) and interviews with 31 children in order to also investigate their emotional responses to a computerized version of the TOEFL® Primary™ reading test. I investigated instances where computers and emotions emerged in the children's drawings or post-assessment interviews to uncover their emotional stance related to the computerized assessment. Results showed that, despite low familiarity with computerized testing, 94% of children had positive or neutral feelings about taking the computerized reading assessment. Considering the quantitative and qualitative data, I discuss policy issues related to computerized assessment of young ELLs.

Wayne E. Wright, Trish Morita-Mullaney, Rudy Rico, Marquetta Straight, Chen Li (Purdue University)
Introducing the Purdue English Language Learners Language Portraits (Purdue ELLLPs)

One of the biggest challenges facing teachers of English language learners (ELLs) is planning and delivering instruction that is appropriate to each students' level of English proficiency (Wright, 2015). In order to do so, teachers need to become adept at using a variety of formative language assessments to determine where students are at and to track their progress over time in developing their listening, speaking, reading, and writing abilities in English (Gottlieb, 2015). To help teachers learn how to effectively use and interpret assessment results, faculty and graduate students in the College of Education's Department of Curriculum & Instruction are developing a free online resource—the *Purdue English Language Learners Language Portraits (Purdue ELLLPs)*. The *Purdue ELLLPs* features linguistic portraits of diverse ELL students from different grade levels, ethnic backgrounds, and levels of English proficiency. Each portrait includes (1) sociolinguistic background of the student, (2) video clips of the student engaged in oral language tasks, (3) video clips of the student reading texts aloud, (4) student's responses to reading comprehension questions, and (5) samples of the student's writing. Along with this rich language data, users are provided with tools to assess each student's oral language proficiency, reading, and writing in English. In this presentation we will discuss the development of the *Purdue ELLLPs*, provide an overview of the website, demonstrate the use of the site to practice assessing the English language proficiency of ELL students.

Jui-Teng Liao (University of Iowa)
Effects of Testing Format on Second Language Reading Performance

Currently, multiple-choice questions (MCQs) and short-answer questions (SAQs) are the most widely-adopted testing formats for assessing L2 English reading proficiency. Compared with SAQs, MCQs are more popular because of their rating efficiency, interrater reliability, and content validity (Plakans &

Gebril, 2013); however, SAQs demand written response without prescribed options, generating a high degree of authenticity (Qian & Pan, 2013).

The purpose of this study is to compare the influences of MCQs and SAQs on reading performance, metacognitive awareness, test completion processes, and task perceptions. In order to gain deeper insight into the impacts of testing formats, this study adopts a mixed methods approach. The quantitative strand compares test performance between MCQ and SAQ reading tasks and investigates the correlation between test performance and metacognitive awareness; the qualitative strand explores the test completion processes and participants' task perceptions.

Thirty-two low-intermediate to intermediate level English learners completed a reading task in MCQ and SAQ formats. Immediately after completing each type of reading task, participants filled out a 5-point Likert-scale survey to specify and rate the frequency of metacognitive strategies they used. Eleven participants took part in an interview describing their test-completion processes and task-perceptions. The results of a paired sample t-test and correlation analysis are discussed along with the qualitative results. Implications for the use of both test formats in classroom-based assessments will be discussed.

SESSION 6 (RAWLS 1071)
10:00-11:30 A.M.

Lixia Cheng (Purdue University)
Pragmatic Appropriateness in an ESL/EFL Oral Discourse Production Test

This paper describes an experimental study on how pragmatic appropriateness in ESL/English as a Foreign Language (EFL) request production can be impacted by English proficiency (high or low) and learning setting (ESL or EFL).

Chinese ESL/EFL participants of low and high oral English proficiency, as identified by an independent speaking test, recorded their responses to two oral discourse completion tasks sharing the same pragmatic feature: P+D+R+ (a.k.a. PDR-high). In these situations, the request recipient has higher socio-institutional power; a large social distance exists between the interlocutors; and the request involves great imposition on the recipient.

Six adult L1 English listeners rated each of 80 Chinese ESL/EFL participants' oral English requests using an analytic rubric adapted from Hudson, Detmer, and Brown's (1995) rating scale. A 2x2 factorial ANOVA test revealed significant main effects of English proficiency and learning setting on the composite score of pragmatic appropriateness (based on five rating criteria: use of typical expressions, amount of information given, levels of formality, directness, and politeness). Furthermore, there was a significant interaction between proficiency and learning setting. These findings imply that ESL/EFL proficiency and learning setting are influential factors that affect pragmatic appropriateness in oral English production of requests in PDR-high tasks; the ESL learning environment most substantially benefits low proficiency learners as far as pragmatic appropriateness is concerned.

This study has provided empirical evidence to support theories about the factors influencing ESL pragmatic appropriateness. It also has practical implications for the rating of pragmatic appropriateness in ESL/EFL speech act production.

F. Scott Walters (University of Illinois at Urbana-Champaign)
Widening the Scope of CA-Informed L2 Pragmatics Testing

Communicative language ability models assume two general sub-abilities, grammatical and pragmatic competence (Bachman 1990; Celce-Murcia et al. 1995). While testing the former has a

long history, pragmatic competence is partly invisible to test methods in the new field of second-language pragmatics testing (SLPT) as they limit researchers interested in theoretical models of SLA and constrain user decision-making. Indeed, the theoretical basis for SLPT, speech act theory (Searle 1969), and the method usually employed, the discourse completion task (Hudson et al. 1995) are called into question by findings from conversational analysis (CA) (Golato 2003). CA has been useful in *post hoc* L2-test validation (Ross 2007; Youn 2014). However, CA as a resource for *a priori* test-task design is still in its infancy although research (Author 2013; 2016) suggests it holds promise.

The present study builds on earlier studies by widening both the sample of learner-proficiencies and the range of pragmatic targets, the latter derived from the CA literature (Pomerantz 1978, 1984; Schegloff 2007; Wong and Waring 2010). The protocol involves ESL adults of low-intermediate to advanced proficiency engaging in a role-play, the responses conveyed to two raters differentially trained in CA. Response data were transcribed and analyzed according to determine (1) whether the wider pool of test tasks can generate results that can validly infer L2 intermediate pragmatic ability; (2) whether CA-informed prompts can elicit targeted skills; and (3) the impact of "CA proficiency" on rater behavior. Qualitative and quantitative results will be presented and suggestions for further research offered.

Aysenur Sagdic (Indiana University)
The Effects of Task Mode on Assessing Pragmatic Inferential Skills

Several studies contributed to the ongoing discussion of L2 pragmatics assessment (Bouton, 1994; Brown & Ahn, 2011; Grabowski, 2009; Hudson, Detmer, & Brown, 1995; Liu, 2006; Roever, 2006; Taguchi, 2005; Youn, 2013, 2014; Walters, 2007). However, further research is needed to examine the features of assessment instruments and their effects on assessing L2 pragmatic ability reliably. This study aims to investigate the extent of task mode (i.e., paper-and-pencil based and audio-based) on the assessment of native speakers' pragmatic ability in comprehending implicature. Thirty-eight native English speakers (NESs) at a US university took a multiple-choice pragmatic test assessing the ability to understand conversational implicature and indirect speech acts (i.e., refusals, requests) and a background questionnaire. The participants were divided into two groups: one group completed the paper-and-pencil test and another completed the audio-formatted test. The findings revealed that (a) both instruments were sufficiently reliable in measuring the ability to understand implied meaning; and (b) task mode had no significant impact on native speakers' ability to comprehend implied meaning. The significance of the results is twofold. First, it is another step to explore the nature of native speaker pragmatic competence. Second, it reveals the effect of the task mode on native speaker pragmatic competence. Implications for assessing L2 pragmatics are discussed and suggestions are provided for improving instrument reliability and the correspondence between characteristics of target language use and pragmatic assessment tasks.

SESSION 7 (RAWLS 1086)
2:30-4:00 P.M.

Saerhim Oh (Teachers College, Columbia University)
Investigating L2 Learners' Use of Linguistic Tools in an Online Writing Test

Advances in technology in recent years have greatly reformed the way we write (Lunsford, 2006; Purpura, 2016). When second language (L2) learners write with a computer, they easily check their spelling and grammar, and find the meanings of uncertain words. However, most writing tests prohibit these tools, based on the assumption that using them in tests provides an inaccurate measure of writing ability. However, if L2 writing ability were re-conceptualized to include these tools, the assessment

would more closely simulate writing behaviors in the real world, and we would be better placed to generalize examinees' performance to writing ability in authentic contexts (East, 2008).
The current study investigated 39 adult L2 learners' use of linguistic tools in an online English writing test. The purpose of the study was to 1) examine the difference between L2 test-takers' writing scores with and without linguistic tools; 2) identify the nature of test-takers' use of linguistic tools in writing tests; and 3) understand test-takers' perceptions of using linguistic tools in assessment. It was found that generally, test-takers received a higher score when linguistic tools were permitted; however, the interpretation of their ability in terms of the different components of the construct of writing ability were the same regardless of whether they had access to linguistic tools or not. Additionally, test-takers mostly used linguistic tools to check their vocabulary and spelling, and a conflict between the extent to which test-takers want to use linguistic tools in tests and their perception of test fairness was identified.

Lia Plakans & Atta Gebril (University of Iowa & American University in Cairo)
Integration in Academic Assessment: Lexical Diversity, Textual Borrowing and Proficiency

Integrated reading-writing tasks are increasingly appearing in assessments used in high-stakes decisions such as college admissions, creating a need for careful study of these newer assessment types. Questions have been raised about the role of source vocabulary in test takers' writing on integrated tasks and, consequently, how scores from these tasks should be interpreted. The presentation investigates issues related to the influence of textual borrowing on lexical diversity and the difference in lexical diversity across test scores on integrated tasks. To this end, 130 students in a Middle Eastern university completed a reading-based integrated task. The essays were analyzed for lexical diversity using CLAN software, a computer program developed to compute lexical diversity. Then to illuminate the impact of the source texts, vocabulary originating from the reading was removed from the essays, and the D index was recomputed for a lexical diversity score with borrowed vocabulary omitted. A paired samples t-test and Analysis of Variance (ANOVA) were used to answer the research questions. The results showed that borrowing from source texts significantly affects the lexical diversity values in integrated writing. Further, the results demonstrated that lexical diversity plays a substantial role in integrated writing scores.

Phuong Nguyen (Iowa State University)
Linguistic Analysis of Essays from Integrated Writing Tasks

This study investigates the differences in linguistic use in responses to integrated writing tasks among test takers of three proficiency levels (level B, D, and Pass) in an English placement test. The corpus, consisting of 429 essays (121,392 words) was a sub-corpus of an English placement test at a large mid-west university. The essays were analyzed for vocabulary distributions, phraseological patterns, and various lexico-grammatical features such as type/token ratio, normalization, noun phrases, *that* clauses and *to* clauses (controlled by verbs, adjectives, and nouns), and adverbial clauses. Results indicate that the groups significantly differed in their use of words from the General Service List 1,000 words (GSL1K; see Nation, 1990) and the Academic Word List (Coxhead, 2000), their employment of nominalizations, and their type/token ratio. Additionally, the Pass group also tended to used more lexical bundles and borrowed less from the writing prompts compared to the other two groups. Regarding nominal features, more proficient writers employed significantly more phrasal noun modifiers such as attributive adjectives, nouns as premodifiers and prepositional phrases. However, writers of three proficiency levels did not differ in their use of finite relative clauses, a clausal feature modifying nouns. Other clausal features functioning as postmodifiers were used very rarely in all proficiency levels although descriptive statistics showed that the Pass group employed these features more frequently. Findings from the study provide implications for writing assessment using integrated tasks and academic writing instruction.

SESSION 8 (RAWLS 1062)
2:30-4:00 P.M.

Aleksandra Swatek & Aleksandra Kasztalska (Purdue University & Southern Arkansas University)
"Write an Email to Your Friend in England": Assessment of the Written English Matura Exam in Poland

At the end of 12 years of Foreign Language learning, 87% Polish students graduating from high school choose to take the mandatory foreign language *matura* exam in English (Centralna Komisja Edukacyjna, n.d.). A high stakes exam administered by the National Examination Board (NEB), matura's scores are used as the sole criterion for admission into Polish universities. However, despite its importance, the matura has not received much attention from assessment researchers.

In the proposed presentation, we examine the theoretical and ideological underpinnings of the written portion of the English matura exam (EME). First, we analyze theoretically its purpose, task design, and rubrics. Specifically, using Bachman and Palmer's (1996) framework for testing, we discuss the purpose and task of the examination, inferences about language ability and language use of the test takers. Furthermore, we consider task dimensions: subject matter, prompt, rhetorical task, audience, genre, etc., for the basic and advanced EME (Purves et al. 1984). We also critically analyze the detailed analytic rubric and scoring examples provided by the NEB.

The results of this initial analysis are juxtaposed against the discourse used by the Polish government to describe the linguistic and cultural goals of EME. We consider these choices in the wider sociopolitical context of Poland and explain how EME documentation reflects a broader goal of the European Union to foster tolerance through the learning of foreign languages. At the same time, we argue that such openness is not always reflected in the linguistic standards valued by the test makers.

Rongchan Lin (Teachers College, Columbia University)
A Multi-Contextual Approach to Investigating Chinese Language Assessment in Singapore

This paper examines Chinese language assessment in Singapore with reference to Kunnan's (2005, 2008) Test Context Framework (TCF). The discussion revolves around the contexts specified in the TCF, namely 1) political and economic; 2) educational, social, and cultural; 3) technology and infrastructure; and 4) legal and ethical. The paper concludes with future directions in research. Through analyzing Chinese language assessment in Singapore via various lenses, this paper aims to further emphasize that language assessment does not only entail language-related issues. It is hoped that this paper would inspire further research regarding language assessment in other multilingual and multicultural societies.

Ananda Muhammad (Iowa State University of Science and Technology)
Alignment of Target Language Use and the High School English National Examination in Indonesia

Despite having adopted the Communicative Language Teaching approach many years ago, Indonesia's English national examination has never changed from the multiple-choice format. This resulted in the non-alignment between the target language use (TLU), which requires students to produce their own language, and the exam itself—thus the authenticity of the exam is questioned. Within the context of this study, the TLU is characterized by the Indonesian high school English subject syllabus as well as the English textbooks recommended by the syllabus. In this study, I described the task characteristics of the English national examination and compared them to the task characteristics of the TLU based on Bachman and Palmer's (1996) framework. In addition to the task characteristics, I also identified and examined some of the exam's qualities with the aim of gauging test usefulness, namely reliability,

construct validity, and authenticity. Based on the results of these examinations, I came to the conclusion that besides the issue of authenticity, the exam also has problems with its reliability and construct validity. This study concludes with the implication that stem from the disconnect between the TLU and the high school English national examination in Indonesia. Suggestions are also made for the improvement of the administration and assessment methods of the English national examination in the Indonesian context.

SESSION 9 (RAWLS 1071)
2:30-4:00 P.M.

Ha Ram Kim (University of Illinois at Urbana-Champaign)
Towards a Profile-Based Rating Scale for Post-Admission Writing Placement Tests

When developing post-admission writing placement tests, it is common for test developers to design a rating scale that only reflects placement options rather than performance levels or profiles. The placement-based scales, while efficient and practical, tend to be limited in its usefulness in informing writing instruction. This study reports on the revision of a placement-based rating scale for a post-admission writing placement test in a large Midwestern US university, to reflect placement as well as writing performance profiles in the new scale.

The scale revision was conducted in three stages. First, key writing criteria were identified by analyzing the composition course syllabi; Next, interviews were conducted with writing instructors to gauge the estimated range of writing performance profiles and levels in ESL composition courses. Then, following a data-driven scale development approach (Fulcher, Davidson & Kemp, 2011), four experienced instructors underwent a series of iterative essay rating and discussions, which resulted in a profile-based rating scale, with five typical writing performance profiles. This new rating scale was piloted on 9 raters with 60 essays, which helped to refine the scale descriptors for better operationalization. Finally, the five performance profiles were linked to different placement recommendations.

Compared with the placement-based rating scale, the revised scale prompts raters to pay more attention to the description of writing performances, which, in turn, provides useful diagnostic information for the writing instructors. In addition, the new profilebased rating scale strengthens the alignment across the ESL curriculum, instruction, and assessment, ultimately enhancing the quality of the test.

Senyung Lee (Indiana University)
A Data-Driven Rating Scale for Assessing Coherence in L2 Writing

It is crucial to inform stakeholders of assessment criteria in clear language to justify the intended use of an assessment. Coherence is an important property of writing quality as it explains a reasonable amount of variance in second language (L2) writing (Bae & Bachman, 2010). Although the importance of coherence in writing is well-acknowledged, the definitions of coherence to date are rather vague, causing subjective understanding of the construct (Todd, Khongput & Darasawang, 2007). Coherence is often operationalized in combination with cohesion (Bae & Lee, 2012), and the bare term coherence is included without explanation in the rubrics of some high-stakes writing tests (iBT TOEFL writing, GRE analytic writing).

This study developed a four-point rating scale for assessing coherence in L2 argumentative writing. Empirically-derived, binary-choice, boundary-definition scale development method was used because it generates rating descriptors which fit a specific testing context (Upshur & Turner, 1995). Argumentative essays written by ESL learners at an American university were examined. Six experienced ESL writing teachers formulated yes/no questions that distinguish coherence of higher- and lower-level essays, through think-aloud protocol and retrospective interviews. Three questions were formulated: Did the

ideas logically flow in a sophisticated fashion? 2) Did you have to make effort to understand the text? 3) Was there any paragraph that does not contribute to the main idea of the essay? When presented to another six writing teachers, these questions were effective in rating the coherence of the same essays. Implications for how to operationalize coherence in L2 writing are discussed.

Reuben Vyn (The University of Iowa)
A Validity Argument for a Foreign Language Writing Performance Assessment Rubric

As student achievement on high-stakes tests are increasingly used to evaluate teacher effectiveness, demonstrating validity of all aspects of assessments including design, administration, and scoring becomes critical to ensuring that decisions made as a result of their use are as fair and ethical as possible. Responding to calls for extending the use of validity arguments to scoring rubrics (Chapelle, 2012; Crusan, 2010; Janssen, 2015), the purpose of this study was to examine the validity of one district's foreign language writing performance assessment rubric and the process by which it was implemented. With a portion of teachers' evaluation and pay dependent on their students' performance on this assessment, the importance of establishing a validity argument for both the assessment and its rubric is critical. In this paper, qualitative analysis of the rubric and its implementation was framed using Toulmin's (1958) model of argumentation with the aim of identifying specific issues and proposing recommendations for improving the validity and related scoring inferences (Bachman & Palmer, 2010). In this analysis, several instances of construct irrelevance were found. Elements recommended for removal or revision included the entire Task Completion domain, the minimum number of clauses in the Vocabulary domain, and rubric criteria and task directions addressing writing style, among others. These findings highlight the importance of regularly analyzing a rubric's alignment with a task's underlying constructs, as well as the need for establishing such validity arguments.

SESSION 10 (RAWLS 1086)
4:15-5:15 P.M.

Ziwei Zhou (Iowa State University)
Exploring the Relationship Between Fluency and Speaking Performance of ITAs

Previous studies suggest that L2 fluency measures can influence the evaluation of speaking performances, but with different degrees of contribution (Bosker et al., 2013; Ginther et al., 2010). Such relationships are still under-researched for international teaching assistants (ITA), who play important roles in undergraduate education in American or Canadian universities. This study, built on a small-scale pilot study, focuses on 100 prospective ITAs at a large Midwestern university in US and aims to investigate the relationships between fluency measures and their speaking performances in an in-house speaking test for ITAs. Three categories of fluency measures, namely, speed, breakdown, and repair, were calculated in the form of 7 variables, based on the automated results from a modified version of de Jong, Persoon, and Quené's (2010) Praat script as well as manual annotation of the speech samples. The raw holistic scores of the speaking performance were analyzed using FACETS (Linacre, 2015) to calculate inter-rater reliability and to obtain corrected or fair scores. A multiple regression was conducted with the fluency measures being independent variables and the fair scores of speaking performance as dependent variable. Results indicated that mean syllable duration was the most significant predictor of the corrected score, while both breakdown and repair fluency measures made varied contributions to the raters' decisions. These findings are informative for ITA programs to better understand the fluency characteristics of ITAs and their contribution to speaking performances. This study also sheds light on a further development of the pronunciation-related rating rubric of the speaking test.

Daniel Isbell (Michigan State University)
Assessing Comprehensibility and Accentedness in L2 Korean Speech

In L2 pronunciation research, the constructs of *comprehensibility* and *accentedness* have been extensively researched (Derwing & Munro, 2015). These constructs are typically assessed via listener judgments of L2 speech on 9-point Likert scales, and these judgments have generally been found to be reliable. Linguistic factors influencing these judgments have also been uncovered (e.g., Saitou, Trofimovich, & Isaacs, 2016). However, the scales have been shown to be difficult for raters to use, particularly in differentiating middle scale points (Isaacs & Thomson, 2013). Furthermore, the majority of research using these scales has focused on L2 English, with little work done on L2 Korean.

This presentation examines the use of translated comprehensibility and accentedness scales in judgments of L2 Korean speech samples. To investigate the functioning of the scales, understand how naïve listeners judged the constructs, and discover challenges in the rating process, 10 native speakers of Korean were recruited to listen to speech samples (including read-aloud and picture description tasks) from learners enrolled in first and second year university Korean courses. Comprehensibility and accentedness scores were analyzed for internal consistency, interrater reliability, and fit to the multi-faceted Rasch model. Additionally, listeners answered post-rating debriefing questions regarding factors influencing construct judgments and difficulties encountered in the rating task. Results indicate that the comprehensibility and accentedness scores were reliable, but listeners had difficulty using mid-range scale points. Listeners also varied somewhat in their views of the constructs. These findings highlight areas for improvement in the assessment of L2 speech for research purposes.

SESSION 11 (RAWLS 1062)
4:15-5:15 P.M.

Elnaz Kia & Valeriia Bogorevich (Northern Arizona University)
Linking Rating Behavior to Criteria: Evaluation of Paired Speaking Task

This paper reports on a mixed-methods approach to evaluate the rating criteria used for a local paired-speaking task and to examine potential biases between the raters and the examinees, examinees' L1, and the criteria. Five raters' examined 56 paired speaking speech samples from examinees of Arabic, Chinese, and Portuguese L1. Prior to rating, individual training sessions were held for each rater. Moreover, raters' rating behavior was recorded via think-alouds. Many-faceted Rasch measurement (MFRM, Linacre, 2015) was used to analyze the scores awarded by the raters and to measure the bias toward individual categories and test takers. Facets modeled in the analysis were examinee, rater, L1, and criteria. Insights were added to the quantitative results based on the raters' comments during the think-aloud sessions. Results indicated that there was no bias between the raters and the examinees or between the raters and the examinee's L1. However, significant biases were located between the raters and the criteria. Suggestions are made regarding improving the rating criteria and improving rater consistency.

Anna Mikhaylova (University of Iowa)
The Effects of French Teaching Assistants' Backgrounds on Writing Assessment

Raters vary in their characteristics that may affect their scoring behavior: their cultural and linguistic backgrounds (Kim & di Gennaro, 2012; Johnson & Lim, 2009), work experience (Cumming et al., 2002), and rater experience (Barkaoui, 2010). The mixed findings of the existing research (Guo et al., 2013) prevents second language researchers from making strong conclusions on the topic. To address the previous mixed findings, the current project is concerned with an in-depth investigation of the relationship between raters' cultural, educational, and professional backgrounds and their assessment of writing.

The purpose of the study is to elicit similarities and differences in the views on assessment of writing among three French teaching assistants at a U.S. university (two French native speakers and one non-native speaker) and to see how their different cultural, educational, and professional backgrounds could affect their views on assessment of writing at the elementary level of French.

The data collected consisted of individual interviews with each participant, a group discussion observation, and copies of three random students' compositions graded by each participant. The qualitative data analysis revealed that the two French native speakers provided a larger amount of more explicit improvement as well as achievement feedback assuming students read their comments. However, the French non-native speaker provided a smaller amount of more implicit improvement feedback assuming students do not read his comments.
The significance of the topic under investigation lies in the necessity to ensure raters' consistency in their scoring behavior and, thus, accuracy and reliability of students' scores.

SESSION 12 (RAWLS 1071)
4:15-5:15 P.M.

Jie Gao (Purdue University)
A Comparison of Two Automatic Writing Evaluation Tools in China: Friendly to Use and Easy to Accept?

In the era of digital writing and composing, automatic writing evaluation (AWE) tools are gradually marking their places in standardized testsand second language writing classrooms. Equipped withtheaim of assisting instructors withgrading work, helping students pin down grammar errors and providing multiple forms of feedback, they are also faced with quite a few challenges. By comparing two of the most widely used automatic writing evaluation tools in China, Teaching Resources Platform(TRP)and pigai.org., the study examined their differences in grammar error categorization, detection accuracy, as well as student users' various perceptionsin a setting of English argumentative writing course.Besides receiving revision suggestions from the two tools, participants in the studyalso obtained a combined form of feedbackfrom the course instructor and TRP. Though TRPhad a better performance in errordetection and correction, the study shows Ad significant difference between students' recognition of thetools' assistance in indexes such as grammar accuracy and argument articulation.Correction and guidance from the human instructor remain to bethe priority for students when it comes for feedback credibility and efficiency.Meanwhile, the practicality and necessity of using AWE tools in second language teaching classrooms have also been discussed, with the hope of directing towards more investigation about human-machine relationship as well as forms of feedback provision for second language writing and writing assessment

Virginia David (Western Michigan University)
Timed-Writing and Process-Based Writing Exams: Comparing Learners' Performance and Perceptions

The purpose of this study is to compare learners' performances on and perceptions of two writing exams: a timed-writing (TW) exam and a process-based timed-writing (PBTW) exam. Eighty-one ESL students had 45 minutes to write an essay for the TW exam. The same participants read an article and watched videos about a topic, discussed the topic, and planned their essays before they wrote the PBTW exam. The order and topics of the exams were counter balanced to diminish test or topic effects. The participants answered a questionnaire about their perceptions of the exams and eighteen participated in interviews for more information about their perceptions. The results of the study revealed that, although the learners' scores did not differ, they expressed a preference for the PBTW exam because it provided them with background information about the topic, they had time to plan, and they could use ideas from

the source materials. The learners scored significantly higher for content and punctuation in the PBTW exam, while they scored significantly higher in spelling in the TW exam. In addition, the participants wrote significantly longer essays and more words per minute in the PBTW exam. Furthermore, they used more sophisticated vocabulary and more nouns in the PBTW exam. The scores that students received in the exams correlated moderately (.391), suggesting that the exams measure different constructs. The results of this study suggest that the PBTW exam may be more appropriate to evaluate the construct of academic writing within the context of higher education.

POSTER SESSION
1:30 - 2:30 P.M. (RAWLS LOBBY)

Mariam Alamyar, (Purdue University)
An Interpretative and Comparative Review on Writing Assessment in First and Second Language

The internet has revolutionized how people find information and compose that information. Starting in elementary school and continuing on through the university level, everyone is integrating digital writing into their writing classes. This development of emerging technologies brought an array of opportunities and challenges to writing/composition teachers in assessing digital writing projects, and they have been struggling and finding it a daunting task to assess digital writing. Thus, this paper is an interpretive and comparative assessment review of two contemporary books, with more than 18 collections of articles that are fundamental to the theme of assessing digital writing. In essence, the presenter explore issues relevant to tools, heuristics, methods, techniques, criteria, pedagogical practices, and principles people were concerned about in 2005 in assessing digital writing projects up to now. Additionally, it addresses challenges and opportunities that digital writing assessment presents to teachers, learners, program administrators, and institutions as a whole. The researches indicated there has always been a struggle between the valorization of traditional writing assessment and the new digital writing assessment. The long history of writing assessment, and the debates that have surrounded it, tells us that digital writing assessment of students composing is a complex and difficult topic. Indeed, digital writing scholars have definitely been trying to consider the ethical issues involved in researching and teaching with digital media. Nevertheless, to understand these issues better, they need to be translated into research on digital writing assessment, particularly validity studies of large-scale digital writing assessment.

Matthew Allen, (Purdue University)
Measuring ESL Reading Fluency Development with Assisted Repeated Reading

Much remains unknown about how to define, measure, and develop reading fluency for ESL students at different proficiency levels (Anderson,1999; Grabe, 2009; 2014; Lems, 2012; Taguchi & Gorsuch, 2012). These practical and theoretical issues are addressed in this presentation of findings from a single-case design (ABAB) study with a small number of participants (n=12 university students). Two research questions are addressed: (1) What are participants' silent and oral reading rates? (2) Does use of an audio-assisted repeated reading program contribute to increased reading rates? In this study, reading fluency is operationalized as the number of words per minute read by participants, and is supported by several other measures to indicate participants' accuracy and comprehension. Empirical data for selected participants will be displayed in graphs showing their (1) recognition vocabulary; (2) baseline measurements of silent and oral reading rates at multiple points in time; and (3) progress in the repeated reading intervention. The graphs will show how silent and oral reading rates compare within and across participants, and the extent to which the reading intervention (IV) increased participants' reading fluency rates (i.e., led to a stable change in the DVs). The use of graphic displays to visualize quantitative data is a hallmark of single-case designs, and the method of careful visual analysis of data to identify trends translates well to a poster session. This study advances our understanding of L2 reading fluency, with implications for assessment, curriculum and instruction, and student motivation.

Guangyan Chen, (Texas Christian University)
Is it Fair for English-Speaking Learners in the U.S to Take the New HSKs?

This study assesses the factor structure of the New HSK and its invariance across subgroups of test takers who differ in their native languages and cultures. The New HSK is a national standardized test used in China to assess the Chinese proficiency of non-native speakers. It is by far the most authoritative Chinese proficiency test. The subgroups refer to the two groups: the group of test takers who are from English-speaking countries and the one who are from Asian countries. There are six levels of the New HSK. The author chooses the New HSK Level 3 (HSK3) to assess the invariance in the factor structure because it is the lowest level that consists of four sections—listening, reading, writing, and speaking. This choice is based on the assumption that the lower the proficiency level is, the higher the language and cultural influence is on test takers' performance. The study results indicate that a correlated four-factor model corresponding to the four language abilities of listening, reading, writing, and speaking best accounts for the factor structure of the test. In addition, the underlying construct has the same structure across the two test-taker subgroups. However, the latent construct differ in their means across the subgroups. This study provides empirical support for the current score reporting practice for the New HSK3. It suggests that the test scores have the same meaning across the two test-taker subgroups. It also identifies possible test-taker background characteristics that affect Chinese language abilities as measured by the New HSK3.

Jean Young Chun & Claire Fields, (Indiana University)
Use of Hedges and Intensifiers by NNS in Oral Discussion

A majority of oral speaking test rubrics lacks the pragmatic aspect of speech, which fails to present information on how strongly or weakly international students impose their opinions on their interlocutors in discussion. The misuse or lack of pragmatic features such as hedges and intensifiers might lead to conflict or misunderstanding in discussion by unintentionally giving an impression of forcefulness. The current study investigated how high-level non-native speakers of English use hedges and intensifiers in discussion to mitigate their claims in comparison to native speakers of English. Sixteen native speakers of English and 32 non-native speakers of English (Korean and Mandarin speakers) were paired with another participant of the same gender and native language and participate in brief English discussions. The numbers and types of hedges and intensifiers that occurred during discussions were examined and compared between native speakers and non-native speakers. The preliminary results show that male and female native speakers used significantly more hedges than intensifiers in the oral discussions. However, the ratios of hedges and intensifiers used by non-native speakers of English were less consistent across gender, native language, and length of residence. The findings encourage discussion on whether oral speaking sections of placement tests for matriculated international students should measure pragmatic features of speech given that highly proficient non-native speakers tend to vary considerably in pragmatic competence

Jung Han & Hyun Jin Cho, (Purdue University)
Challenges of Elementary School ELLs in Mainstream Classes and Effective Use of Classroom-based Assessments of ELLs

Every K-12 ELL faces double the academic challenge in that they must acquire English proficiency while also developing subject knowledge for them to be able to compete with students who are well-established in the English language. However, teachers may feel that feedbacks from state-level large-scale assessments are too general and vague to identify students' specific language proficiencies. This study investigates the academic challenges of ELLs and how mainstream teachers can use classroom-based assessments for teaching ELLs. In this study, the researchers carried out a one-year field study that included several mainstream classroom observations, ESL class observations, various assessments of ELLs' English language proficiencies, and new mini-lessons for them. The participants were three

Chinese ELL students in elementary school who had just arrived in the US. The assessments and mini lessons were developed based on PreK-12 ELP standards (TESOL, 2006), WIDA ELP standards (2007 edition), and WIDA ELD standards. The results have shown that ELLs needed additional linguistic, graphic, and visual support in content area learning to achieve academic success as shown in Gottlieb's research (2005). The findings of this study also indicate that classroom-based assessments provide more specific and accurate information of ELL's proficiency than standardized large-scale assessments in both identifying their academic and linguistic needs and implementing proper instructions according to each of their special needs. In order to examine the teachers' perception of the role of the classroom-based assessment, surveys and in-depth interviews are needed for the further study.

Hyun-Ju Kim, (Danook University, S. Korea)
Implementation of Criterion-Referenced College Scholastic Ability Test in Korea

The English section of the College Scholastic Ability Test (CSAT) in Korea will be implemented as the criterion-referenced tests from 2018 and the fixed cut-off scores will be used to classify students' performance levels. Over the 22 years of the CSAT history in Korea, the testing policies regarding the structure of the test, test administration, test materials, score reporting, score interpretation, etc. have gone through numerous changes. The present policy regarding score reporting is still in debate although it has already been decided. Since the Angoff method to set cut-off scores can influence students' performance level depending on varying test difficulties, it is necessary to examine both methods (Angoff and standard setting methods) in order to provide valid and credible scores to students. The purpose of this study is to investigate the feasibility of standard setting method for the CSAT although it has not been selected to be used in CSAT in Korea. The current study has examined the processes of bookmark method with empirical examples, and suggests the bookmark method could be a promising approach for the criterion-referenced test for Korean CSAT.

Susie Kim & Ok-Sook Park, (Michigan State University)
Validation of a Korean Speaking Proficiency Assessment

In recent years, Korean has become one of the most commonly taught languages in U.S. higher education (Goldberg, Looney, & Lusin, 2015). Accordingly, significant changes have taken place for Korean language programs with the very first publication of the *National Standards for Korean Language Learning* by the American Council on the Teaching of Foreign Languages (ACTFL) in 2015. To support these changes, the American Association of Teachers of Korean has launched a project to develop a *National Standards*-based curriculum for college students, in an effort to better articulate and align language programs with these standards. As such, the assessment of learner performance following these ACTFL proficiency guidelines has become essential in moving towards standardization. Though ACTFL provides Oral Proficiency Interviews (OPI) for Korean, a readily available alternative tool for measuring global oral proficiency would be beneficial for evaluating the students, courses, and the language program itself in accordance with *National Standards* and proficiency guidelines. The main focus of this project is to validate a web-based Korean Speaking Proficiency Assessment (KSPA) developed at a large Midwestern university. Ten students enrolled in 300- and 400-level Korean courses self-assessed their proficiency based on ACTFL's Can-Do Statements (ACTFL, 2013), took KSPA and OPIc, and completed feedback questionnaires for both tests. Three Korean instructors at different U.S. institutions participated as raters for KSPA and provided evidence for and comments on their ratings. Together, this work-in-progress assessment tool aims to investigate multiple aspects of the validity of the test from different perspectives of the stakeholders

SungAe Kim, (Purdue University)
Variables that Affect English Reading Difficulty for English Language Learners

One of the most important considerations in developing test items for standardized tests is to control the level of test difficulty in order to appropriately discriminate the scores within test takers. If test items are too easy or too difficult for test takers, it may not be possible to appropriately discriminate high scores. Since the impact of the test scores is huge on test takers, developing appropriate test items is very important. In relation to controlling the level of test difficulty, identifying factors affecting test difficulty should be prioritized before administrating the test. In standardized tests, reading plays an important role, and reading achievement for English language learners is very important for their academic success. There have been studies regarding the variables affecting second language reading such as Freedle and Kostin's (1993), Jin and Park (2004), Chang (2004), etc. In this paper, I will review previous studies about variables that affect reading test difficulty in standardized tests with English language learners and analyze tests which have been implemented in K-12 education, using Bachman's (1990) framework of Test Method Facets (TMF). According to TMF, there are five categories including testing environment, test rubrics, input, expected response, and the relationship between input and response. This paper focuses on the nature of the language of test items which are characterized by length, propositional content, organizational characteristics, and sociolinguistic characteristics within standardized English tests.

Xiaorui Li, (Purdue University)
What does it Mean? Predicting Oral Proficiency by Using Referential it

The referential coherence is one of the important aspects that influences listener's information mapping during oral communication. This study examined the relationship between the referential *it* usage in item responses and the holistic scores on a semi-directed measure of oral English proficiency among Mandarin speakers. The spoken responses of 60 respondents to two items (Compare & Contrast and Newspaper Headline) on the Oral English Proficiency Test (OEPT) administrated at Purdue University have been analyzed. All respondents in this study represent three levels (Level 3, 4, and 5) of the OEPT1 scale that ranges from the lowest Level 2 to the highest Level 6. All tokens of referential *it* were identified as either recoverable or non-recoverable, and all observations of recoverable *it* were categorized as either anaphoric or cataphoric. The study results indicated a significant increase of the referential *it* ratio and the recoverable *it* ratio in responses from Level 3 to Level 4, but the differences were less noticeable between Level 4 and Level 5. Additionally, a higher ratio of anaphoric *it* was observed in responses to Compare & Contrast and a higher ratio of cataphoric *it* was observed in responses to Newspaper Headline. This study suggested that the referential *it* ratio and the recoverable *it* ratio might be two effective predictors of oral English proficiency especially in distinguishing low proficiency responses from intermediate proficiency responses. Compare & Contrast appeared to be a more desirable elicitation tool for anaphoric *it*, whereas Newspaper Headline was more efficient in eliciting cataphoric *it*.

Yi Li, (Southwest University, China)
Automated Essay Feedback Generation and Its Impact on EFL Writing Revision

Writing essays is a very important skill for college students, but a difficult task too. It is particularly true for English as a Foreign Language (EFL) students in China. It would be useful for students if they can receive timely and effective feedback about their writing. Due to the high student-teacher ratio in China, teachers often cannot provide individualized feedback to their students in a timely manner. The new technology--automatic essay feedback generation--provides the possibility to help Chinese students write in English and receive feedback from instructors. The study includes two parts: (1) build up and (2) evaluate the automatic essay feedback generation. We first analyzed 1290 teacher comments on their 327 English-major students, and annotated the feedback on seven aspects of writing, including *the*

grammar, spelling, sentence diversity, structure, organization, supporting ideas, coherence and conclusion, for each paper. Then, machine learning approach with an automatic feedback classification experiment was conducted. Finally, we compared the impact of the system generated-indirect corrective feedback (ICF) with human teachers' direct corrective feedback (DCF) in two English writing classes (N=56 in ICF class; N=54 in DCF class) at a key Chinese university through a web-based assignment management system. The study results indicated that the feasibility of this approach that system generated in ICF can be as useful as direct comments made by the teachers in terms of improving the quality of the content regarding to the *structure, organization, supporting ideas, coherence and conclusion,* and encouraging students to spend more time on self-correction.

Mayu Miyamoto, (Purdue University)
A Scale Development Project for Performance-Based Test (PBT)

Although performance assessment of oral proficiency is increasing in work settings and high-stake testing (McNamara, 1990), many of the foreign language course classroom assessments are still relying on the traditional written formant. These paper tests often focus on grammatical accuracy or lexical knowledge rather than oral proficiency. There is a huge discrepancy between what they learned to do in class and what they are tested on. In order to fill in this gap, a new testing system called "Performance-based Test (PBT)" was developed and put into practice at a university in Midwest. It was developed as an achievement test for languages courses to evaluate learners' speaking ability. It is conducted on computers and it consists of five tasks: a monologue, a read aloud, a read & answer, Q&A, and a role-play. However, when the test was developed, there was no corresponding scale. There was a strong need of a rating scale that's specifically developed for this test as Fulcher (2013) claims that "the new testing purpose requires a new rating scale". A new rating scale was developed from the collected data of student responses, and this presentation will focus on the development process of a new scale, as well as descriptors, benchmarks and a grading rubric tree.

GoMee Park, (University of Iowa)
Use of TOEIC: de facto Language Policy in South Korea

In South Korea, TOEIC scores have been used and accepted *naturally* as almost the sole gauge of evaluating one's English proficiency in all sectors of Korean society. Hence the TOEIC has been a de facto language policy throughout the nation. Acknowledging the TOEIC as the prevalent and apparent de facto language policy in Korea, this paper focuses on investigating what ideology is reflected in the TOEIC that affects language policy in Korea, the role of TOEIC as a mechanism, and the effects of TOEIC as a de facto language policy in Korean society. Various documents have been reviewed and investigated including the government publications related to English language policy, and a series of the ETS Test Taker Reports of TOEIC, as well as newspaper articles and advertisements. The findings revealed that predominance of English language was enhanced by the introduction of neoliberalism and globalization in Korean society that led to the use of the TOEIC test as an easy way to measure one's competitiveness. The phenomenon is not a part of natural and inevitable process of globalization but an output of *ideology* spread through various mechanisms. Such a practice has been entrenched through various devices such as government policies, major corporations that require TOEIC scores as one of the employment qualifications, and universities and mass media in Korea that acknowledge TOEIC scores as *the* assessment tool for quantifying an individual's capability

Sockwun Phng, (Iowa State University)
Man/Machine: Human and Computer Raters in L2 Oral Assessments

In English-medium classrooms, oral proficiency is important for non-native speakers of English to possess because learning is a group process requiring student-teacher talk and student-student talk.

English placement tests in universities are important for that reason. Since speech events in the target language use domain is largely interactive, measuring the speaking construct means measuring interactional competence on top of comprehensibility and intelligibility, which comprise appropriate use of phonology, appropriate and accurate use of vocabulary and grammar, and appropriate fluency. This paper presents a comparison of human raters and computer raters in terms of measurable versus non-measurable speech features as well as in terms of the qualities of reliability, authenticity, and practicality. Human raters excel at rating for interactional competence, appropriate and accurate use of vocabulary and grammar, and appropriate fluency, and they meet the quality of authenticity. Computer raters, on the other hand, excel at rating for appropriate use of phonology and appropriate fluency, and they meet the qualities of reliability and practicality. This paper concludes with a suggestion for cooperative rating between human and computer raters whereby both groups of raters contribute their expertise to second language oral assessments to create a reliable, authentic, and practical rating process that measures the entire construct for speaking.

Ji-young Shin, (Purdue University)
The Use of "Actually" in Korean College Aptitude Scholastic Test

This presentation investigates the use of a discourse maker, "actually," in the English listening test sections of College Scholastic Aptitude Test (CSAT) in Korea from 1994 to 2016. Based on the comparison of the use of "actually" in Korean CSAT with its authentic counterpart in spoken corpus (Michigan Corpus of Academic Spoken English, Santa Barbara Corpus of Spoken American English) and previous studies (Aijmer, 2002; Karlsson, 2014), the presentation aims to address the different use of "actually" in the national oral proficiency test. Considering the significant function of discourse makers and their frequent use in spoken communication (Aijmer, 2002), it is important that the input from Korean CSAT should expose test takers to authentic use of "actually" for their oral communication skill improvement. However, this issue has been neglected in previous research.
Thus, this study analyzed the use of "actually" in both dialogues and monologues in the listening test scripts of CSAT in terms of its frequency, sentential role (an adverb or discourse marker), syntactic distribution (initial or final of a clause or utterance), and functional use (textual/interpersonal, contrastive/emphatic). According to the preliminary results from the comparison with authentic spoken discourses, Korean CSAT predominantly used "actually" as an initially-positioned adverb with more contrastive function. The pedagogical implication of this misrepresentation in Korean CSAT will be discussed especially considering its exceptional washback effect on nationwide English education (Choi, 2013).

Evan Simpson, (Iowa State University)
Shifting Baselines: How Changing from NRT to CRT Affects Placement

The question of whether to use an NRT or a CRT framework for interpreting placement test results is widely discussed within language programs. To that end, while Brown (1989) found that their reading placement test accurately reflected both students' reading proficiency relative to each other as well as their reading ability in relation to the program's stated criteria, a similar analysis of Iowa State University's English Placement Test (EPT) has yet to be conducted. Therefore, this paper presents the results of a study that replicated and expanded on Brown's original study by analyzing the EPT reading and listening scores. Furthermore, this paper addresses the possible ramifications of changing interpretive frameworks by identifying the differing outcomes for stakeholders. In conclusion, this project, by closely analysing and interpreting the data, highlights the importance of clearly prioritizing the goals of placement tests.

Magda Tigchelaar, (Michigan State University)
Using Self-Assessments to Predict Spoken French Proficiency

Research on self-assessment has revealed that language learners are generally poor judges of their spoken performance, but that the use of can-do statements may help to sharpen their judgments (VanPatten, Trego, & Hopkins, 2015). The purpose of this study was to analyze the self-assessed spoken French abilities that students said they "can do" in relation to ACTFL (2012) proficiency scores they received on an oral proficiency interview (OPIc). My secondary aim was to assess the scales that have been used to convert OPI ratings to numeric scores. Prior to taking the ACTFL OPIc test, 216 learners of French rated can-do statements related to their speaking skills. I conducted a series of regression analyses to determine how well self-assessment scores predicted the rating of participants' spoken proficiency by certified ACTFL raters. I found that the strength of the relationship between self-assessment and OPIc rating was strongly influenced by the type of numeric scale that was used: Using ordinal regression, a majority (65%) of the variance in OPIc scores was explained by self-assessment scores. Using linear regression, when the scores were converted to equal-interval scales, self-assessment scores explained approximately 30% of the variance. On a graduated scale that reflected the increasing distances between ACTFL (2012) proficiency levels, only 20% of the variance was accounted for. I discuss the results in terms of using self-assessments as a tool for assessing spoken proficiency for placement and instructional purposes and advise on the implications of converting OPI ratings to numeric scores for research purposes.

Taichi Yamashita, (Iowa State University)
The Significance of Triangulation in Written Corrective Feedback Research

The validity of measurements of second language writing has been discussed by itself (Polio, 2001), and other fields (e.g. task-based language teaching) have also partly but immensely contributed to this discussion (Ellis & Yuan, 2004; Larsen-Freeman, 2006; Ong & Zhang, 2010; Vyatkina, 2012) as their subsidiary purpose. However, such measurements have not been utilized in written corrective feedback (WCF) research practice (Polio, 2012), and they usually rely on a single measurement with excessive attention to accuracy. To point out the risk of such convention in WCF research, the present WCF study triangulated students' product and each writing construct (i.e. accuracy, complexity, fluency). Seventeen students taking an intermediate Japanese course at a U.S. university participated in the study. They underwent three writing tasks and three revision tasks with either focused (N=9) or unfocused (N=8) WCF for three weeks in total. Accuracy of their products (i.e. pretest, immediate posttest, delayed posttest) was measured by the ratio of error-free clauses and total clauses and by errors divided by total characters. Complexity was measured by characters per t-unit and by dependent clauses divided by total clauses. Fluency was measured by characters per minute and characters per non-dysfluency. Correlation studies revealed that (1) the accuracy measurements were correlated very well, (2) the measurements for complexity were not correlated well enough, and (3) neither were those for fluency. The results indicated the necessity of triangulation of students' written product and further triangulation of each construct especially for complexity and fluency in future WCF research.

NOTES

NOTES

Opportunities for Language Assessment Professionals & Scholars

Research Grants

The **Spaan Research Grant Program** was established to recognize Mary Spaan's contributions to the field of language assessment through her work at the University of Michigan. Spaan grants provide financial support for those wishing to carry out research projects related to second or foreign language assessment and which also investigate an aspect of CaMLA's tests.

- Proposals are invited from graduate students, faculty, and other language assessment professionals
- Applications can be submitted online at **CambridgeMichigan.org/spaan**

Applications must include:

- Name(s) of applicant(s) and contact information
- Research project details
- Curriculum vitae for each researcher named on the application form
- Student applications must include contact information for two academic references

Deadline for applications: November 21, 2016

Find out more at **CambridgeMichigan.org** (search "Spaan").

Internships

The **CaMLA Internship Program** provides professional training and research opportunities in English language assessment. Successful applicants will work closely with CaMLA staff on projects related to their interests and skills.

Who we're looking for . . .

CaMLA interns are often English language teaching professionals with an interest in assessment, or graduate students studying linguistics, psychology, foreign language assessment, education, psychometrics, or related fields.

- Interns have the opportunity to work on a wide range of assessment projects
- Internships typically take place from May through August

Deadline for applications: January 27, 2017

Find out more at **CambridgeMichigan.org** (search "Internships").

About CaMLA

CaMLA—Cambridge Michigan Language Assessments—is a not-for-profit collaboration between the **University of Michigan** and **Cambridge English Language Assessment**, two institutions with long and distinguished histories in the field of language assessment, teaching, and research.

CAMBRIDGE ENGLISH Language Assessment
Part of the University of Cambridge

UNIVERSITY OF MICHIGAN

CambridgeMichigan.org


PURDUE
UNIVERSITY®