MwALT
Midwest Association
of Language Testers

MwALT 2015

Conference Program

Writing Assessments and Assessing Writing: Research and Practice

October 3, 2015
Iowa City, IA
University of Iowa

TABLE OF CONTENTS

Welcome to MwALT 2015

Welcome to the 17th annual conference of the Midwest Association of Language Testers. The University of Iowa College of Education is pleased to host the conference. This gathering brings together researchers, test developers, students, teachers, and teacher educators from across the Midwest and from other regions as well. The presentations, conversations, and community-building that occur at the MwALT Conference encourage us all not only to continue but to elevate the work we do in assessing languages.

New ideas are often presented at MwALT as well as questions long grappled with in the field. The program this year does not disappoint in this regard. In our planning this year, we decided to give a central theme to the main events (plenaries and panels) to dig more deeply into one area of assessment. The theme of “Writing Assessments and Assessing Writing: Research and Practice” led to a sizable number of concurrent sessions on important issues in performance assessment such at rating, score interpretation, and performance indicators. These topics translate to the assessment of speaking, which is another area well represented in the program. Along with these areas, the conference includes sessions on new constructs, assessment literacy, and narratives of test development, which will spur innovative thinking along with practical guidance.

Our hope is to provide a day that is both scintillating and enjoyable. Thank you for joining MwALT at UIowa!

Lia Plakans, Past-MwALT President and Conference Team Leader

Conference Organizing Committee

Lia Plakans

Renka Ohta

Jui-Teng Liao

Warren Merkel

Stephanie Lynch

Volunteers

Dr. Leslie Schrier

Yu-Chi Wang

Sha Huang

Elizabeth Plummer

Jennifer Brown

Tamar Bernfeld

Steve Lynch

Yupeng Kou

Yuan Lu

Fang Wang

Gomee Park

Sponsors

The MwALT 2015 Organizing Committee sincerely appreciates the following sponsors' support:

The University of Iowa
COLLEGE OF
EDUCATION

Location: University of Iowa — Lindquist Center, College of Education

7:45-8:20	**Registration & Continental Breakfast**	**N300 Lindquist Center (Jones Commons)**
CONFERENCE SESSIONS		
8:30 – 9:35	**Welcome to MwALT** (Lia Plakans, *The University of Iowa*) **Plenary speech** **L2 Writing Placement: A (Not-So-Pretty) Behind-the-Scenes Look** **Christine Tardy** *University of Arizona* **301 Lindquist Center – South (Leonard Feldt Auditorium)**	
Concurrent Presentation Sessions		
	Presentation Session 1: 301 Lindquist Center - South	**Presentation Session 2: 204 Lindquist Center - South**
9:45 – 10:15	**A Corpus-Based Analysis on Syntactic Complexity as a Measure of Oral English Proficiency of ITAs** **Suthathip Thirakunkovit** **Rodrigo A. Rodríguez Fuentes** **Kyongson Park** *Purdue University*	**Investigating the Quality of L2 Writing: Linguistic Measures as Score Predictors** **Hyung-Jo Yoon** *Michigan State University*
10:15 – 10:45	**Designing and Implementing a Portfolio Assessment in an ITA Speaking Classroom** **Yoo-Ree Chung** *Iowa State University*	**Computer Proficiency, Language Proficiency, and Computerized Writing Assessment: A Comparability Study of Low-Level Learners** **Laura Ballard** *Michigan State University*
10:45-11:00	**BREAK – Drinks and Snacks: N300 Lindquist Center (Jones Commons)** **Sponsored by CaMLA**	

Concurrent Presentation Sessions

	Presentation Session 3: 301 Lindquist Center - South	Presentation Session 4: 204 Lindquist Center - South
11:00 – 11:30	**The Speaking Construct Realized through Theatre** **India Plough** *Michigan State University*	**The Development of an Essay Rating Scale for a Post-Entry English Proficiency Test** **Lixia Cheng** **April Ginther** **Matthew Allen** *Purdue University*
11:30 – 12:00	**Developing an In-House IEP Oral Interview Placement Test** **Peter Chiligiris** **Silvana Dushku** *Intensive English Institute at the University of Illinois*	**Relating Writing Scores to CEFR Levels: An Application of Rasch-Grouped Rating Scale Model** **Chih-Kai Lin** *Center for Applied Linguistics*
12:00 – 1:30	**Lunch (12:00-1:00) Sponsored by WIDA** **Concurrent with: Poster session (12:45-1:15), Business meeting (12:30-1:00, S 302 LC)** **N300 Lindquist Center (Jones Commons)**	
1:30 – 2:15	**Plenary Speech** **Technology and Innovation in Writing Assessment** **Carol Chapelle** *Iowa State University* **301 Lindquist Center – South (Leonard Feldt Auditorium)**	
2:15 – 2:25	**Intermission**	

Concurrent Presentation Sessions

	Presentation Session 5: 301 Lindquist Center - South	Presentation Session 6: 204 Lindquist Center - South	Presentation Session 7: 302 Lindquist Center - South
2:25 – 2:55	**What Makes a Good Calibration Sample for Rater Norming Sessions?** **Jack Drolet** **Heekyoung Kim** **Aron Ohlrogge** **Daniel Reed**	**Developing Item Specifications for an Integrated Writing Achievement Test: Examining Prompt-Response Relationships over Time** **Ryan Lidster**	**Exploring ESL Composition Instructors' Writing Assessment Literacy** **Heon Jeon** *Ohio State University*

	Michigan State University	*Indiana University at Bloomington*	
3:00 – 3:30	**Implementing a Hybrid Assignment-Rater Norming Protocol in an ESL Writing Program** **Jin Kim** **Leyla Lambert** **Jeff Arrigo** **F. Scott Walters** *University of Illinois at Urbana-Champaign*	**Ontological Realism as a Validity Criterion for the Assessment of Strategic Competence** **Stephen O'Connell** **Steven Ross** *University of Maryland at College Park*	**Mainstream Teacher Candidates' Perspectives on ESL Writing: The Effects of Writer Identity and Rater Background** **Hyun-Sook Kang** *Illinois State University*
3:35 – 4:05	**Researching Writer Rater Processes: Are Concurrent Think Alouds the Best?** **Deirdre J. Derrick** *Northern Arizona University*	**A Corpus-Based Analysis of Lexical Richness: Can TOEFL Writing Sub-scores Predict Lexical Variation** **Kyongson Park** *Purdue University*	
4:05 – 4:20	**BREAK – Drinks and Snacks: N300 Lindquist Center (Jones Commons)**		
4:15 – 5:15	**Panel Session** **Current Issues and Dilemmas in Assessing Writing** **Organizer: Deborah Crusan,** *Wright State University* **Panel Members:** **Carol Chapelle**, *Iowa State University* **Mark Chapman**, *WIDA at The University of Wisconsin, Madison* **Lia Plakans**, *The University of Iowa* **Christine Tardy**, *The University of Arizona* **301 Lindquist Center – South (Leonard Feldt Auditorium)**		
5:15 – 5:30	**Awards Presentation & Closing**		

Plenary Speakers

Abstracts

Christine Tardy

University of Arizona

Writing Assessment for Placement from an L2 Writing Perspective

At U.S. colleges and universities, one of the most ubiquitous course requirements is that of first-year writing, often consisting of a range of courses designed to meet the needs and interests of a diverse student population. Administration of first-year writing courses therefore entails developing effective practices for placing students into the most appropriate course or determining if they have already met the institutional requirement. Simply put, writing placement is central work in a writing program, and it increasingly involves placement of students who are writing in English as an additional language. In this talk, I draw on my experience at one large state university to reveal the less visible side of L2 writing placement, including historical, social, political, and even economic influences. As I will share, decisions are often shaped by far more than sound knowledge of assessment principles. By describing some of the less-than-ideal realities that can impact placement policies and practices, I hope to offer a perspective that might contribute to an ecological understanding of second language writing assessment.

Carol Chapelle

Iowa State University

Technology and Innovation in Writing Assessment

Changes in writing practices over the past several decades have continually created the need to reexamine the way that writing is assessed. Evolving technologies at the root of change also present opportunities for test developers to design a range of new writing assessments, but acceptance of new assessments requires openness to innovation on the part of prospective test users. In language assessment, “innovation” refers to new ideas, approaches, and practices in test development and use. Innovative language tests can be particularly challenging to introduce because prospective test users may view the assessment of writing with a respect for orthodoxy that creates resistance toward new ideas, approaches, and practices. This paper explores the challenges posed by innovation in writing assessment by describing three examples of innovations in writing assessment that have been both motivated and made possible by new technologies: assessment of writing online with web-searching permitted, assessment of grammatical correctness in formative assessment of writing; and assessment of academic writing using integrated writing tasks. For each of the examples, I will describe the assessment and its context of use; I will also demonstrate the type of validation efforts that were undertaken to investigate the adequacy of the interpretations and uses of each innovation. I will suggest that the validation research should play a role in the diffusion of innovation.

Presentation Session 1 (9:45-10:45)

Abstracts

Suthathip Thirakunkovit

Rodrigo A. Rodríguez Fuentes

Kyongson Park

Purdue University

A Corpus-Based Analysis on Syntactic Complexity as a Measure of Oral English Proficiency of ITAs

This is an exploratory corpus-based study investigating the use of different syntactic features of prospective ITAs who have previously taken an oral English proficiency test (OEPT). Our research questions are: 1) What are the syntactic features that are commonly found in the examinees' responses? and 2) What are the distinguishing syntactic features that can characterize different levels of proficiency? Based on previous seminal research, ten major syntactic features were selected. Our data consist of 278 examinees from the three largest L1 groups – Mandarin, Hindi, and Korean. Biber Tagger was used to tag the syntactic features of our interests. The results show that based on the categorization proposed by Biber, Gray, and Poonpon's (2011), the OEPT examinees' responses contain both conversation and written features. Most spoken features, including the use of subordinating and adverbial clauses and finite and non-finite complement clauses, are not good predictors of the scores as these features are found quite frequently across all proficiency levels. However, some written features, which are considered more complex e.g., relative clauses, non-finite noun modifiers, and stance adverbs seem to have more potential in predicting the OEPT scores. Two possible interpretations might be that because the examinees were given opportunities to plan their speech, their responses illustrated the mixed use of both written and spoken features although spoken features occurred more than written features. Moreover, the use of either spoken or written features exclusively to measure language proficiency may miss out on some important structures of complexity in academic oral assessments.

Yoo-Ree Chung

Iowa State University

Designing and Implementing a Portfolio Assessment in an ITA Speaking Classroom

While portfolio assessment has been employed in many K-12 ESL classrooms and program evaluations, the assessment of adult ESL students' language abilities at tertiary-level institutions has by and large been approached to and researched from a product-oriented perspective of language learning and testing. Among the four major language skills, writing instruction has best benefitted from portfolio assessments, arguably because the written language mode facilitates the consolidation of students' products and the observation of their learning progress made over a course of time. On the other hand, technological advances now allow us to easily store and track down students' learning progress in L2 speaking as well. With this backdrop, this study explores a possibility of facilitating the enhancement of adult language learners' oral communication skills through a process-oriented portfolio assessment approach, situated in the context of an ESL speaking course offered for international teaching assistants (ITAs). Grounded in the frameworks of portfolio assessment suggested in the literature, the study will describe the structure and the components of the portfolio particularly designed for the instruction and evaluation of ITAs' oral communication skills at the upper-intermediate level. It will then report both the instructor's and the students' perceptions and interim evaluations about the effectiveness of the portfolio assessment on the target language ability. Finally, the study will conclude with some challenges the stakeholders encounter and room for improvement in the practice of the speaking portfolio assessment for ITAs.

Presentation Session 2 (9:45-10:45)

Abstracts

Hyung-Jo Yoon

Michigan State University

Investigating the Quality of L2 Writing: Linguistic Measures as Score Predictors

In the field of second language (L2) assessment, one of the primary objectives has been to measure L2 writing proficiency validly. Of several external factors that may influence testing results (e.g., rater, task complexity, or topic), the role of genre has received increasing attention. One recent study statistically confirmed that assessing one essay of a particular genre does not provide generalizable evidence of one's writing proficiency (Bouwer, Béguin, Sanders, & van den Bergh, 2015). Based on such findings on genre effect, our next step seems to better understand specific influences of genre on essay quality and their textual features. This study, thus, has two objectives: (1) to show how L2 learners develop their writing skills in two genres over one semester; (2) to identify combinations of linguistic measures as predictors of the essay quality of two genres. Data were collected from 51 adult ESL learners. The participants were asked to write six essays (three narrative and three argumentative essays) over the course of one semester, with the genre alternated and the topics counterbalanced. Using an analytic scoring rubric, two experienced raters evaluated the essays. Target linguistic measures (linguistic complexity and coherence) were analyzed using automatic systems. Findings indicated L2 learners obtained significantly higher scores in narrative than in argumentative essays. Moreover, the results showed that three sub-scale scores (content, organization, and vocabulary) increased over one semester. Regarding essay score prediction, different combinations of linguistic measures were identified depending on genre. This study has implications for L2 writing pedagogy and assessment.

Laura Ballard

Michigan State University

Computer Proficiency, Language Proficiency, and Computerized Writing Assessment: A Comparability Study of Low-Level Learners

Recently, many high-stakes tests have moved from paper to computer-based formats (e.g. TOEFL, WIDA ACCESS, Smarter Balanced). In response, education, technology, and language testing researchers have conducted comparability studies on these testing modes (Endres, 2012; Kingston, 2009). While this research addresses the validity and ethics behind each mode, there is a lack of research examining their effects on specific language learner populations; in particular, low-level ESL learners who potentially have limited computer exposure. To address this, I investigated: 1) whether low-level learners perform differently on paper-based and computer-based writing assessments; 2) if so, whether score differences could be attributed to computer proficiency; and 3) whether other writing features are affected by computer mode. Twenty-nine low-level English learners completed a paper-based and computer-based version of a university ESL placement test for writing. Participants completed two measures of computer proficiency: an objective measure of L1 and L2 computer-proficiency and a subjective measure, a computer skills/usage questionnaire. Results indicated no statistical difference between participants' paper-based and computer-based writing scores, suggesting computer proficiency had a negligible effect on writing scores. Further analysis, however, revealed an interaction between writing fluency and writing-mode preference, a positive correlation between (L1 and L2) computer proficiency and L2 proficiency, and an overall limited computer proficiency (17 WPM in the L1, and 15 in the L2). I use the results to discuss assessment mode effects and CBT policies and protocols, namely time limits and mode optionality.

Presentation Session 3 (11:00-12:00)

Abstracts

India Plough

Michigan State University

The Speaking Construct Realized through Theatre

Davidson (LTRC, 2015) has called on the language testing community to consider the metaphor 'speaking examiners are actors.' By enlisting the performance of experts who study and create authentic characters, might we gain additional insight into the dynamics of interaction that can inform our understanding of the speaking construct? This information could then strengthen the validity argument of an existing test, be used in the development of a new test, or assist us in the training of examiners and calibration of raters. A semi-professional theatre troupe will be recruited to enact two scenes involving 5 different actors. Scene One, the "The Testing Situation," involves 3 actors—two students and one examiner; Scene Two, the "The Debate," involves two actors—friends. The same decision-making task sets the stage for both scenes. The scenes begin at the point when the actors must defend their decision, either to the examiner or to each other. Actors are asked to perform for 4-8 minutes. The recorded performances, portions of which will be shown, will be examined for interactional features that current research indicates are significant to our understanding of the speaking construct (e.g., turn-taking, topic management, listener support strategies, body language). Additionally, after the performances, the actors will be interviewed to explore the composition and development of their characters as well as the rhetorical script that they created for the performance. Comparisons between the two scenes and between the findings of Scene One and those of the research on performance-based paired speaking tests are made.

Peter Chiligiris

Silvana Dushku

Intensive English Institute at the University of Illinois

Developing an In-House IEP Oral Interview Placement Test

An oral interview that effectively assesses a student's speaking skills is an integral part of an in-house placement test that can help ensure effective placement for a specific IEP's leveled-curriculum. The presenters document step-by-step a data-driven project to design a reliable oral interview placement test at a large Midwestern IEP. The weaknesses identified in the IEP's previous oral interview and the changing needs of the IEP's student population informed the decision to develop a new oral interview as part of the IEP's placement test. The presenters document the steps taken to research and design this oral proficiency test on the basis of these considerations. These steps include designing and revising the test specifications, writing the items and protocol, piloting the questions, and analyzing the data over several semesters. In addition, an oral interview-scoring rubric was developed, piloted, and revised several times. The statistical analysis of the data collected over two semesters, together with student and rater feedback, informed the oral interview rubric and test item revisions. A series of rater training workshops aimed at ensuring better rater calibration and enhancing inter-rater reliability was organized each semester. Finally, the presenters report on the project outcomes and highlight the lessons learned from this project. They emphasize the need for ongoing data collection and analysis, as well as rater training and calibration. They also demonstrate the value of teacher involvement in the design and implementation of such a test for capacity building and professional development.

Presentation Session 4 (11:00-12:00)

Abstracts

Lixia Cheng
April Ginther
Matthew Allen

Purdue University

The Development of an Essay Rating Scale for a Post-Entry English Proficiency Test

The dramatic increase in enrollment of international undergraduate students at U.S. universities not only reflects a national trend of shifting undergraduate demographics but also highlights the need for effective evaluation of newly admitted international students' English language proficiency. To better inform language instruction in an English for Academic Purposes program at a large public university, an internet-based post-entry English proficiency test, the Assessment of College English - International (ACE-In) was developed. This presentation focuses on the development of an empirically derived rating scale for the writing assessment included in the ACE-In. Drawing on the literature of L2 rating scale development (e.g., Fulcher, Davidson, & Kemp, 2011; Upshur & Turner, 1995), we began by analyzing a sample (n=42) of first-semester international students' ACE-In essays to identify the categories and elements (i.e., constructs and variables) present and emerging levels of performance. A series of rating and discussion sessions were iteratively conducted with 33 additional essay samples until agreed-upon descriptors were established and an acceptable level of inter-rater reliability reached. These rater norming sessions not only served the purpose of developing and refining an essay rating scale, but also helped to build a community of practice by providing a venue for raters to share what they value as writing instructors (Kauper, 2013). This presentation provides a practical example of developing an empirically derived rating scale for a timed writing assessment. With the emphasis on instructor values, we provide a model for creating effective communities of practice through rating scale development.

Chih-Kai Lin

Center for Applied Linguistics

Relating Writing Scores to CEFR Levels: An Application of Rasch-Grouped Rating Scale Model

The purpose of a cut-score study is to establish cut scores on a test in relation to some performance standards. The cut scores serve to aid the interpretation of test scores with respect to the skills and proficiency levels represented in the standards (Cizek & Bunch, 2007; Tannenbaum & Cho, 2014). The current paper presents a cut-score study of an English writing proficiency test developed for teachers of English in an EFL context. In this context, it is important to gather evidence that the scores are meaningful and can be interpreted in terms of the Common European Framework of Reference (CEFR). The paper reports an application of the Rasch-Grouped Rating Scale Model to determine cut scores in relation to CEFR levels, and it touches on how the use of the psychometric model was informed by the design of writing rubrics. In particular, a pool of writing tasks was developed for piloting, and each task was aimed at a CEFR level from A1 to B2. Each level-specific task was scored on a rating rubric specifically developed for the target CEFR level. One advantage of this model-based application is that it establishes CEFR cut scores directly from the pilot data (by considering task difficulty, rater severity and rubric step measures), and as such it does not require a separate study, independent of the pilot, to determine the cut scores. The assessment practice and technical aspect in the presentation will be of interest to those who are involved in other performance-based assessment contexts.

Presentation Session 5 (2:30-4:00)

Abstracts

Jack Drolet

Heekyoung Kim

Aron Ohlrogge

Daniel Reed

Michigan State University

What Makes a Good Calibration Sample for Rater Norming Sessions?

The importance of selecting appropriate essay samples for rater norming cannot be overestimated. Typically, a main criterion for the selection of essays to include in calibration sets is consensus among experienced raters. Other criteria often used are wide score ranges (in order to represent the breadth of the rating scale) and subscore patterns in the case of analytic rubrics. Several interesting questions surround this process. Why do raters exhibit strong agreement on some calibration essays but not on others? Is it best to include only the noncontroversial samples, or could there be value in trying to explain discrepant cases? Put another way, what makes a calibration sample good or bad? To investigate these questions, a qualitative analysis of 50 essays was done based on essays representing varying levels of agreement. The essays, which were previously rated for a large-scale test, were recirculated among 5 of the most experienced raters in the rating pool with an accompanying survey. The survey aimed to make explicit each rater’s rationale and to determine whether disagreement resulted from failure to follow the rubric, or whether the rubric itself was problematic (underspecified, ambiguous, etc.). The presenters will summarize the main findings and then engage the audience in a discussion of the implications for rater training practice both in terms of method and policy.

Jin Kim

Leyla Lambert

Jeff Arrigo

F. Scott Walters

University of Illinois at Urbana-Champaign

Implementing a Hybrid Assignment-Rater Norming Protocol in an ESL Writing Program

Achieving high inter-rater reliability across multiple sections of an ESL writing course is essential for reasons of validity (Messick, 1989; Kane, 2012) and fairness (Xi, 2010), and adequate rater training is necessary to achieve these (McIntyre, 1993; Weigle, 1994, 1998; Elder et al., 2007). However, implementation of face-to-face norming sessions is problematic in large institutions due to large numbers of graduate-student instructors and mutual time constraints (Hamilton et al., 2001). To meet these practical challenges, a group of writing instructors in an ESL writing program at a large university developed a hybrid, assignment-rater norming protocol that includes an asynchronous online assignment-rater norming and a synchronous, face-to-face session. Two different versions of this protocol were implemented with a one-year gap between the two. The first version, a fully online protocol, was piloted in fall 2014 and produced mixed results. A second version, the focus of the present study, is a hybrid online/f2f protocol intended to build on the results of the pilot by examining (1) instructors' performances and perceptions of the two different versions of the online assignment-rater norming site created with Moodle course management software, (2) the effects on training of using different rubrics and integrating a synchronous session, and (3) the benefits and challenges of developing and implementing the hybrid protocol. Quantitative and qualitative data analyses include examination of survey results from participants, analysis of essay grading results, and observation notes from face-to-face sessions. Implications for improving the assignment-rater norming practices in large institutions will be discussed.

Deirdre J. Derrick

Northern Arizona University

Researching Writer Rater Processes: Are Concurrent Think Alouds the Best?

Research on writing rater cognition has typically used concurrent think aloud procedures (e.g., Barkaoui, 2007; Lumley, 2002). Many researchers who use concurrent think alouds assume that the procedure is relatively straight-forward and results in reliable data, despite the fact that rating is a cognitively-demanding task (Barkaoui, 2011; Lumley, 2005). Although think alouds have yielded insights into rater decision-making processes and writing scale interpretation, they have been found to affect veridicity (accurate representation of the process) and reactivity (alteration of the process) to varying degrees (Barkaoui, 2011). Retrospective stimulated recalls are an alternative method used in second language and assessment research (e.g., Cohen, 1984), typically with speaking raters. Stimulated recalls can result in highly accurate memories, particularly if the object(s) used to stimulate and support the recall are appropriate (Gass & Mackey, 2000). This methodology, however, has not previously been used to explore writing rater processes. The present study compares the two procedures to explore possible differences in the information provided by each. Think alouds and stimulated recalls from four experienced raters were collected as part of a larger study to provide validity support for the revised ECPE writing scale. Reports were coded and compared to assess trends in the quality and quantity of information obtained from the two procedures. Even with a small sample size, certain trends emerged. Qualitative and quantitative analysis of the data suggest that the two procedures might provide different information that could be used to answer different research questions.

Presentation Session 6 (2:30-4:00)

Abstracts

Ryan Lidster

Indiana University at Bloomington

Developing Item Specifications for an Integrated Writing Achievement Test: Examining Prompt-Response Relationships over Time

The use of integrated read-to-write tasks in language assessment continues to expand due to numerous factors, but above all else because there is evidence that read-to-write tasks correspond better with the construct of Academic English (e.g. Plakans, 2008). A growing body of research has (favorably) compared read-to-write tasks to independent writing tasks; however, not all read-to-write tasks are equally successful in eliciting desired responses, and yet there are very few studies comparing various designs of read-to-write tasks to each other. In addition, while previous literature on task design offers much in the way of general guidelines (e.g. Davidson & Lynch, 2002), there are few studies empirically demonstrating how changes to item specifications impact response behavior. This study investigates how response patterns on a read-to-write assessment differed according to the prompt specifications. Specifically, this study examines an achievement test students in high-intermediate level classes in an Intensive English Program at a large Midwestern university. Based on feedback from students and instructors, as well as empirical analyses of the written responses, the item specifications evolved several times. The construct, however, remained largely constant: namely, the ability to write an argumentative essay using multiple sources for support while responding to a counterargument. We present evidence based on content analysis of student responses that prompts based on current specifications not only elicit the target construct, but are more effective in doing so than previous iterations. We discuss the challenges of eliciting argumentative writing and how they have been addressed in our program.

Stephen O’Connell

Steven Ross

University of Maryland at College Park

Ontological Realism as a Validity Criterion for the Assessment of Strategic Competence

Strategic competence has proven to be a construct difficult to assess in language testing. In some assessments, such as the oral proficiency interview, strategic competence is assessed through the use of a role play with a complication that is devised to force the candidate to resolve a transactional problem. Assessment of candidate performance on the role play is highly subjective, and is contingent on the raters’ accurate identification of interactional evidence of strategic competence. Validation of the strategic competence exhibited in role plays has to a large extent been interpretive. As a backing for an interpretive argument, the criterion of ontological realism is used in this study. Eleven samples of English-as-a-foreign-language OPI role-plays with a complication (administered to Japanese L1 speakers) were judged by 52 untrained English native speakers asked to assess candidates’ successful completion of the role-play task. Untrained raters’ judgments regarding successful completion of the role-plays provide evidence that strategic competence is an identifiable construct to naïve native speakers of English, and by doing so provides evidence of ontological realism, a key criterion for a validity claim (Borsboom et al., 2004). Evidence in support of the ontological validity of the strategic competence role play is presented through conversation analysis augmented by quantitative analyses describing sources of variation among the naïve raters.

Kyongson Park

Purdue University

A Corpus-Based Analysis of Lexical Richness: Can TOEFL Writing Sub-scores Predict Lexical Variation

In this study, I investigate the lexical variety in the L2 written corpus as a predictor of L2 writers' English writing levels. This study aims to discover whether the score of students' performance on timed exams (TOEFL) measures alongside or against their lexical variation as demonstrated on an untimed first assignment. Specifically, I examine three research questions: 1) Can TOEFL writing sub-scores predict lexical variation in international first year composition (FYC) papers? 2) Is there a correlation between nationality (Chinese, Korean, and Indian) and lexical variation in students' autobiographical writing? 3) Can the researcher determine change in the lexical variation between the first draft and the last draft in autobiographical narrative writing? Based on Webb and Nation's (2010) research, I assess the lexical variation among lexical richness measuring type-token ratio (TTR), standardized type-token ratio (STTR), mean of word length, and frequency of use of equal-character length words (one to 11 character length words). The new local second language writing (SLW) corpus data developed and established by a large Midwestern University are used to provide primary texts and background information of all participants, and sub-corpus of specific international students' data (n=132) are selected for this research. The results show that TOEFL scores might not be a reliable variable to indicate the lexical variation in FYC writing, although it distinguishes the students with high TOFEL writing scores from the students with low and intermediate TOEFL writing scores. On the other hand, the nationality of the students correlates with lexical variation by presenting a strong effect on lexical richness in FYC writing. My findings reveal that although there is a significant change in text quantity between the first draft and the last draft, the change in text quality do not occur with regard to TTR or STTR. This indicates that students who have low or intermediate TOEFL scores might need to take extra English courses before taking a FYC class. This study calls attention to the ways students' TOEFL scores and other demographic information, such as nationality or L1 background, hold larger pedagogical implications for FYC instructors.

Presentation Session 7 (2:30-4:00)

Abstracts

Heon Jeon

Ohio State University

Exploring ESL Composition Instructors' Writing Assessment Literacy

Assessment literacy refers to teachers' knowledge about what, when, and how to assess students' progress of learning by using various assessment tools. Despite the importance of assessment literacy, ESL composition instructors are not well trained with what, when and how to assess writing effectively. In addition, few studies focused on ESL composition instructors' writing assessment literacy in university setting. This study aims at answering the following two questions: (a) What are ESL composition instructors' perceptions about assessing writing and (b) How do ESL composition instructors assess writing? In order to answer the first research question, a survey was developed and distributed to ESL composition instructors and the results of survey were analyzed quantitatively. Among them, three survey respondents participated in the subsequent study investigating the second research question: how they assess students' writing. Three data gathering instruments – (a) interviews, (b) field notes, and (c) documents related to writing assessment practices – were used for answering the question. The triangulated data were analyzed according to the two stages of data analysis: (a) open coding (b) analytical coding. The primary findings of this study illustrate that survey respondents were not highly aware of the importance of effective writing assessment. In addition, the three participants' writing assessment practices varied according to the years of teaching experiences. This study would contribute not only to revealing the gaps between perceptions and practices among ESL composition instructors but also to raising the importance of systematic writing assessment.

Hyun-Sook Kang

Illinois State University

Mainstream Teacher Candidates' Perspectives on ESL Writing: The Effects of Writer Identity and Rater Background

This study explored the extent to which the ethnic identity of a writer and the background (gender and area of teaching) of a rater can influence mainstream teacher candidates' evaluation of ESL writing as a source of variability, using a matched-guise method. Teacher candidates were led to believe that a one-page essay presented to them was produced by an ESL learner whose first language was either Chinese or Spanish, respectively. The essay on public transportation in a foreign city was elicited from an ESL learner enrolled in an intensive English program, and was manipulated to incorporate error patterns often observed among Chinese- and Spanish-speaking learners. One-hundred-sixty-three undergraduate students enrolled in a teacher education program at a U.S. university were asked to score the ESL essay holistically, provide qualitative comments, identify the three most troublesome errors, and offer suggestions on how to improve ESL writing. No significant effects of the writer's identity on holistic scoring were detected. However, a different picture emerged in the results of qualitative comments, rank-order of error gravity, and advice on how to improve ESL writing. The teacher candidates revealed different categories of rater responses, presumably influenced by the ethnolinguistic identity of the writer. While rater background in terms of gender and area of teaching had significant effects on the global scoring of the ESL writing, they did not show any significant impact on the nature of the responses in terms of general comments, order of error gravity, or advice on L2 writing skill development.

Poster Session (12:00-1:30)

Abstracts

Deirdre J Derrick

Northern Arizona University

Assessing High-stakes Writing: A Validation Project for the New ECPE Writing Scale

The development of writing scales should be research-based, with validity a consideration at each stage of the process, since a priori development, even when followed by post hoc empirical studies, can result in problems with validity and reliability (Fulcher, 1996). Kane's (1992, 2013) argument-based approach to validation offers a framework to guide validity work even during the development process. Last year, the Examination for the Certificate of Proficiency in English (ECPE) writing scale was redesigned, resulting in an analytic scale with five criteria: topic development, organization and connection of ideas, grammar and syntax, vocabulary, and authorial voice. The authorial voice criterion, based on the work of Zhao (2013) is new, and has not been previously incorporated into high-stakes, large-scale standardized tests of English proficiency. This presentation describes a project to provide support for the evaluation inference in Kane's framework. The study looked at how experienced ECPE raters interpreted and applied the new scale, with particular focus on the authorial voice criterion. Four experienced ECPE raters participated in a calibration session and provided concurrent think alouds and stimulated recalls. Qualitative and quantitative analysis of the data indicated that raters were able to apply the scale consistently overall, although there remained some confusion as to how to interpret and apply the authorial voice criterion. Further analysis revealed ways in which the authorial voice criterion could be strengthened.

Senyung Lee

Indiana University at Bloomington

ESL Learners' Strategies for Marking Footing in Prompt-based Argumentative Writing

Prompt-based argumentative writing is a task often used in high-stakes writing tests (Jeffery, 2009). Previous research reported that ineffective argumentative essays are characterized by making claims without supporting details and listing supporting claims as if they are universally accepted knowledge (e.g., Cumming, Kantor, Baba, Eouanzoui, Erdosy, & Jamesse, 2005). This study investigates how English as a second language (ESL) learners indicate orientations of ideas in their supporting claims in timed, prompt-based argumentative writing. What Goffman (1981) calls *principal* (i.e., the person who came up with the idea) in his framework of production format was adopted in order to identify the orientations of ideas in second language (L2) writing. Understanding how L2 writers specify the principal in timed writing tests can provide insight into the relationship between the way the principal is indicated and the overall writing scores.

Eight essays written by intermediate and high-intermediate ESL learners were analyzed. The principal of each supporting claim was categorized based on whether the principal was specified (e.g., *Psychologist say that TV shows affect children's behavior*.) or not (e.g., *Children tend to imitate what they see on TV*.) and the way the principal was specified. It was found that principals of supporting claims were unspecified in the majority (i.e., 82%) of the supporting claims made in the essays, regardless of learners' English proficiency level. Potential implications for L2 writing instruction and scoring objectives are discussed.

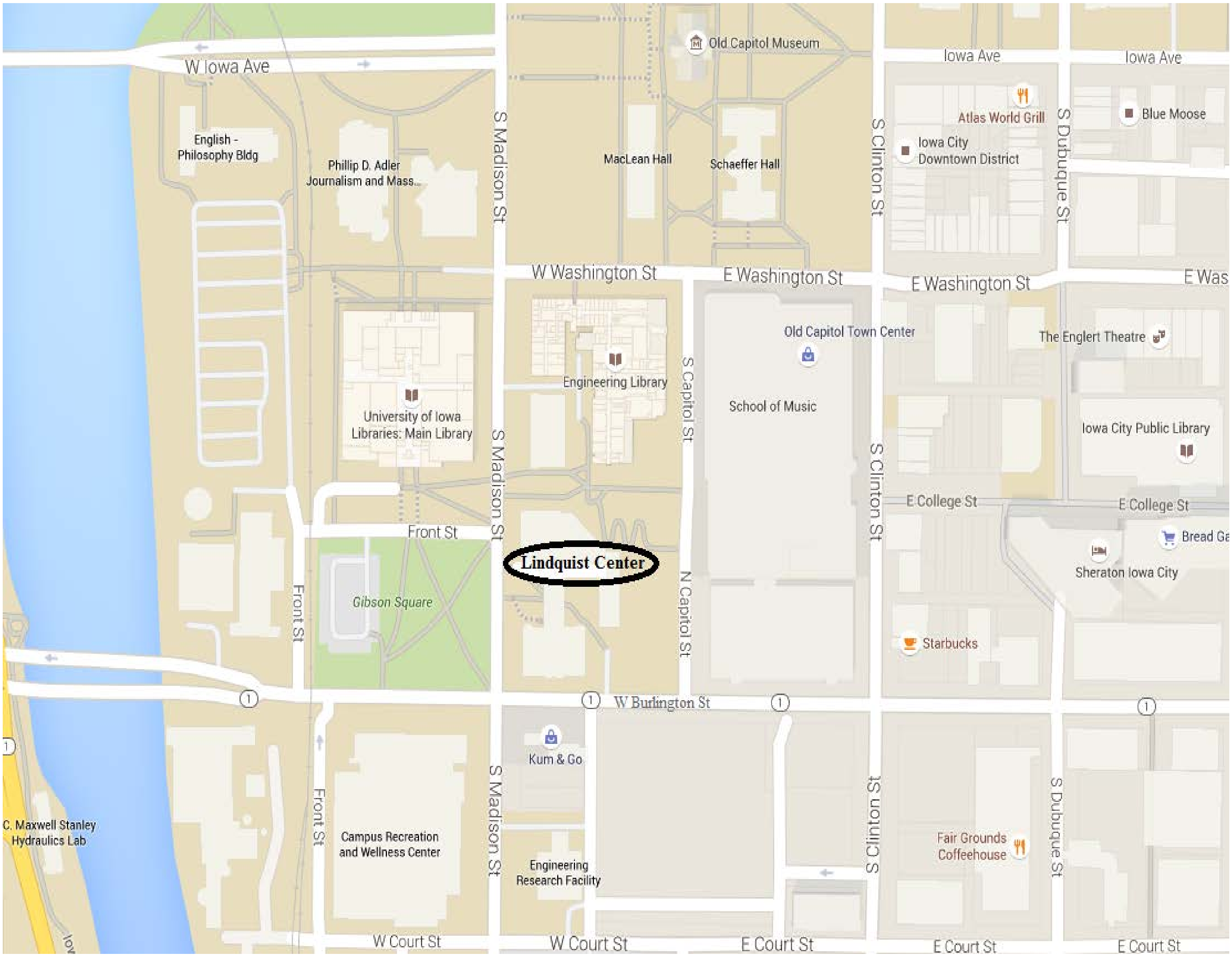
Renka Ohta
University of Iowa

Reliability in Holistic vs. Analytic Scoring of a Reading-to-write Task: Generalizability Theory Approach

The purpose of this study is to investigate whether different rating scales affect score reliability, and if so the magnitude of difference between the scores. Integrated writing assessments, which require L2 learners to combine multiple language skills to write a composition, have been ubiquitous in EFL/ESL writing courses as they reflect real-life writing activities commonly seen in English-medium universities. When evaluating integrated writing, both holistic and analytic scales have been used. Some researchers have pointed out that the particular scale has been chosen without legitimate justification. Therefore, this study intends to help EFL/ESL educators establish legitimacy in selecting a rating scale for integrated writing assessments from the perspective of score reliability. Five experienced ESL/EFL teachers at the secondary or post-secondary levels participated in this study. The raters evaluated 60 argumentative essays written by EFL learners using holistic and analytic scales with one month in between ratings. The writing scores were analyzed using generalizability theory, which enables researchers to compute the variance components related to examinees, raters, and examinee × rater interactions (Shavelson & Webb, 1991). I employed a fully crossed, univariate one-facet $p \times r$ design for the holistic ratings and a fully-crossed multivariate one-facet $p^\bullet \times r^\bullet$ design for the analytic ratings. For both designs, p*ersons* is the object of the measurement and *raters* is a facet of the measurement. For analytic ratings, *scales* is a fixed facet. The poster details the results of the study and concludes with the study's limitations and implications for future research.

Index of Presenters

University of Iowa Campus Map



My Notes