



Hawaii TESOL Virtual Conference 2022

Exploring New Dimensions in Vocabulary and Grammar

February 19, 2018



Hawai'i TESOL is pleased to welcome Eli Hinkel as our plenary speaker. Eli has taught ESL and applied linguistics, as well as trained teachers, for over thirty years and has published numerous books and articles on learning second culture, and second language grammar, writing, and pragmatics in such journals as *TESOL Quarterly*, *Applied Linguistics*, *Journal of Pragmatics*, and *Applied Language Learning*. She is also the editor of the Routledge ESL & Applied Linguistics Professional Series of books and textbooks for teachers and graduate students.

Schedule at a Glance

9:00 - 9:10 am	Welcome Remarks: Anthony Silva, Hawaii TESOL President in Room #1				
9:10 - 10:00 am	Plenary: Eli Hinkel in Room #1				
	Room #1 (Brent)	Room #2 (Perry)	Room #3 (Sam)	Room #4 (Kalehua)	Room #5
Session 1 (20 minutes) 10:10 - 10:30 am	Methodology of the English language lexico-grammatical competence formation of civil engineering students Natalia Tarasiuk	A Corpus and multimodal approach to phrasal verbs Tom Court Jeff Maloney	Communicative approach in teaching grammar online Natalia Romanchuk	Teaching the subjunctive through vocabulary knowledge Maria Levy	Teaching world Englishes vocabulary: students as researchers Joel Hung Hartse
Session 2 (20 minutes) 10:40 - 11:00 am	The educational challenges of Arab ESL students at the University of Florida Elsagheer Abdelrazak	The LMS sandbox: A new dimensional preservice teacher tool Perry Christensen	Reflection on teaching and assessing grammar in Chilean EFL Betsy Gilliland Andrea Muñoz Galleguillos,	Bad is good for second language learners: Facilitating effects of negativity bias in the domain of vocabulary learning Olena Vasylets	Pedagogical translanguaging and grammatical literacy Isidro Gonzalo
Session 3 (40 minutes) 11:10 - 11:50 am	Networking Room	QUIZLET: An effective online tool to foster vocabulary acquisition Paola Santamaria	Practical concerns with a local intensive English program's rubric revision Yu-Tzu Chang	Graduate students' grammar feedback practices: Examining writing groups interaction Rabail Qayyum	Vocabulary learning via Reddit and Instagram: Autonomy, agency, and authenticity Austin Pack Mavis Lee
12:00 - 1:00 Lunch					
Session 4 (40 minutes) 1:00 -1:40 pm	Networking Room	Using technology to practice vocabulary and grammar in context Hanh Nguyen	Developing, examining, and using a learner corpus in an EAP program Anne McCarrey Tobon Luo Xi Lin Chin Wai Gordon Lee	Instructional strategies to expand vocabulary Nicole Heinlein	
Session 5 (40 minutes) 1:50 - 2:30 pm	Networking Room	Utilizing Online Tools for Developing Academic English Word Knowledge Laura Bacchus Hikaru Ishiyama Michol Miller	Categorizing language learning strategies in EFL / ESL oral tasks William Richardo Ortiz Garcia Carol Anne Ochoa Alpala	Vocabulary and grammar across contexts: Reflections from an academic exchange Michelle Kunkel Chika Takahashi Nanami Yano Kazuho Toyota Yuna Sawada Tomoe Tanaka Ces Jocson Emma Mason	
2:30 - 3:00	Raffle				

Letter of Welcome

Dear Conference Participants,

We want to welcome you to our second Hawaii TESOL Virtual Conference. We are extremely happy that you have decided to participate with us.

Sincerely,

Brent Green & Perry Christensen
Conference Chairs

Presentation Sessions (all links to sessions are provided)

Room #1

9:00-9:10 AM

Welcome and Introduction of Plenary Speaker

Anthony Silva, Hawaii TESOL President

Brent A. Green, Hawaii TESOL Conference Co-Chair

9:10 - 10:00 AM

Plenary Speech

Eli Hinkel

Session 1 (20 minutes)

10:10 - 10:30 am

Methodology of the English language lexico-grammatical competence formation of civil engineering students

Natalia Tarasiuk

Session 2 (20 minutes)

10:40 - 11:00 am

The educational challenges of Arab ESL students at the University of Florida

Elsagheer Abdelrazak

11:10 - 2:30 pm

Networking

2:30 - 3:00 pm Raffle

Room #2

Session 1 (20 minutes)

10:10 - 10:30 am

A Corpus and multimodal approach to phrasal verbs

Tom Court

Jeff Maloney

Session 2 (20 minutes)

10:40 - 11:00 am

The LMS sandbox: A new dimensional preservice teacher tool

Perry Christensen

Session 3 (40 minutes)

11:10 - 11:50 am

QUIZLET: An effective online tool to foster vocabulary acquisition

Paola Santamaria

Session 4 (40 minutes)

1:00 -1:40 pm

Using technology to practice vocabulary and grammar in

context

Hanh Nguyen

Session 5 (40 minutes)

1:50 - 2:30 pm

Utilizing Online Tools for Developing Academic English Word Knowledge

Laura Bacchus

Hikaru Ishiyama

Michol Miller

Room #3

Session 1 (20 minutes)

10:10 - 10:30 am

Communicative approach in teaching grammar online

Natalia Romanchuk

Teaching the subjunctive through vocabulary knowledge

Maria Levy

Session 2 (20 minutes)

10:40 - 11:00 am

The educational challenges of Arab ESL students at the University of Florida

**Elsagheer Abdelrazak The LMS sandbox: A new
dimensional preservice teacher tool**

**Perry Christensen Reflection on teaching and assessing
grammar in Chilean EFL**

**Betsy Gilliland Andrea Muñoz Galleguillos, Bad is good
for second language learners: Facilitating effects of
negativity bias in the domain of vocabulary learning**

Olena Vasylets

Session 3 (40 minutes)

11:10 - 11:50 am Networking

**Room QUIZLET: An effective online tool to foster
vocabulary acquisition**

**Paola Santamaria Practical concerns with a local
intensive English program's rubric revision**

**Yu-Tzu Chang Graduate students' grammar feedback
practices: Examining writing groups interaction**

Rabail Qayyum

12:00 - 1:00 Lunch

Session 4 (40 minutes)

1:00 - 1:40 pm Networking

Room Using technology to practice vocabulary and

grammar in context

Hanh Nguyen Developing, examining, and using a learner corpus in an EAP program

Anne McCarrey Tobon

Luo Xi Lin

Chin Wai Gordon Lee Instructional strategies to expand vocabulary

Nicole Heinlein

Session 5 (40 minutes)

1:50 - 2:30 pm Networking

Room Utilizing Online Tools for Developing Academic English Word Knowledge

Laura Bacchus

Hikaru Ishiyama Michol Miller Categorizing language learning strategies in EFL / ESL oral tasks

William Richardo Ortiz Garcia

Carol Anne Ochoa Alpala Vocabulary and grammar across contexts: Reflections from an academic exchange

Michelle Kunkel Chika Takahashi Nanami Yano Kazuho Toyota

Yuna Sawada Tomoe Tanaka

Ces Jocson

Emma Mason

Room #4

Room #5

Presentation Abstracts

Thursday June 1 9:15 AM

Introduction of Lyle Bachman

Lyle Bachman's career spans 47 years over a variety of domains, and trying to do justice to its breadth and importance in a short introduction is quite a challenge; however, it is one that I am privileged to accept. In doing so, I will start with Lyle's university education and early experiences that eventually led him to Thailand where his career in language testing began. Thailand is where I "officially" met Lyle and where he and I started our collaboration, one that continues to this day. I will then provide a brief "for the record" overview of the subsequent stages in Lyle's career in order to show how it developed and how he produced some of the major contributions in our field. I will then turn to the central role Lyle has played in and multi-faceted contributions he has made to the formation and sustenance of LTRC, the most important meeting for language testers, and ILTA, the most important professional association for language testers. I will conclude with my experience of some of the intellectual and personal qualities that have kept Lyle on the trajectory to become, in my opinion, the foremost language tester in the history of our field.

Adrian Palmer
University of Utah

Adrian Palmer teaches a variety of courses in L2 teaching and language assessment. One of his recent co-authored with Lyle Bachman books, *Language Assessment in Practice* (Oxford University Press), 2010, recently received the 2012 SAGE/ILTA award for the best book in the area of language testing/assessment (2009-2012).

Thursday June 1 9:50 AM

The US community college L2 language placement conundrum: A case study

Historically, community colleges in the United States have experienced a love/hate relationship with their placement procedures. On one hand, they appreciate a system that places students into specific English and math classes as a means of ensuring college success. On the other hand, students, faculty, and administrators have become frustrated with the inaccuracies of a system that doesn't seem to get it right for many students.

Because of open admissions policies instituted in the 1970s and 1980s, community colleges across the US face the challenge of ensuring that students are prepared for college level work. Unlike highly selective institutions, these challenges have high stakes implications for students and the colleges because of state and federal funding linked to student completion rates.

Additionally, US community colleges are seeing a growing number of students who speak English as an additional language and who have varying degrees of academic and English language preparation upon arrival at these institutions. In most US community colleges, a good portion of the responsibility for determining college preparedness lies solely on the institution's math and English placement exam.

In this presentation, I will identify the challenges of community college L2 language placement across the United States. I will then share how those challenges have impacted my institution, Salt Lake Community College and discuss ways in which we have tried to mitigate these challenges. Finally, taking a language assessment perspective, I will suggest a way forward for the L2 placement conundrum in US community colleges.

Brent Green
Salt Lake Community College

Brent A. Green is Associate Dean of ESL, College Academic Readiness, and Testing at Salt Lake Community College. His areas of expertise and interest include language assessment, TESOL methodology, and corpus-based language learning. He is co-author of *Developing Academic Literacy* (2014, Kendall Hunt). Brent also teaches TESOL endorsement courses for preservice and in-service K-12 teachers.

Thursday June 1 10:50 AM

Contextual issues that influence the construct definition for English in the workplace

In language assessment, many researchers advocate an interactionist approach to construct definition in which context is part of the construct (e.g., Chalhoub-Deville, 2003). Although proficiency in English for academic purposes has been carefully considered from an interactionist perspective (e.g., Xi, 2015), English for the workplace poses distinct and complex challenges for construct definition. Meaningful inferences about English proficiency for the workplace context could be better supported by additional research and analyses.

Given that the workplace domain of language use is amongst the most heterogeneous that can be defined, one fundamental issue is construct and domain specificity (e.g., English for the workplace vs. English for tourism). A second fundamental issue is how one identifies the reference variety of English (e.g., English as a Lingua Franca vs. more localized varieties of English such as Standard American). How one articulates and reconciles these issues has strong implications for whether and how much context is represented in the construct, as well as the goal of attainment (e.g., native-like proficiency, comprehensibility to native/local community, comprehensibility to a broader community). Over the last 15 years, the related research literature has grown substantially in the field of applied linguistics; a critical synthesis of this literature has implications for the development and validation of assessments and learning products in the English for the workplace domain (e.g., through the development of materials that are more closely aligned and relevant to the needs for English in the workplace). This presentation provides a discussion of these issues and their implications for assessment.

Jonathan Schmidgall
Educational Testing Service

Jonathan Schmidgall is an Associate Research Scientist in the English Language Learning and Assessment research center at ETS. He has conducted and coordinated validation research for local and international high-stakes language assessments, including UCLA's Test of Oral Proficiency and ETS'

TOEIC tests. His research has focused on defining constructs, test design for summative and formative purposes, the use of technology, and designing ‘real-world’ validation studies.

Thursday June 1 11:25 AM

Can the same task be used to assess both science and language proficiency?

The new standards in K-12 education in the United States focus more on the practices involved in the content areas and less on discrete pieces of knowledge. The Next Generation Science Standards (NGSS), for example, promote science learning that blends together three dimensions: disciplinary core ideas, science and engineering practices, and crosscutting concepts. Prior to NGSS, most science instruction focused primarily on the disciplinary core ideas only. The type of “three-dimensional” learning promoted by NGSS cannot be assessed using traditional multiple-choice items. Instead it will require performance-based, complex, constructed-response tasks that will engage students in language-intensive practices such as “constructing explanations” and “engaging in argument from evidence.”

During the talk, I will explore whether NGSS-aligned science assessment tasks can be used to assess English learners’ (ELs) language proficiency in addition to their ability to engage in science three-dimensional learning. Such “integrated assessments of language and content” would present several advantages over traditional English language proficiency (ELP) assessments. Language proficiency would be assessed in the context of an actual content area task. This would address the lack of correspondence between the ways in which traditional ELP tasks that aim to assess the language of the content areas engage language and the ways in which the content areas actually engage language. From a practicality perspective, integrated assessments could cut down on the amount of testing that students in school must do. Finally, these types of tasks could allow for a better understanding of how science and language proficiency develop and support each other.

During the talk I will examine how an assessment use argument could be constructed to justify the development, validation and use of integrated assessments of language and content.

Lorena Llosa
New York University

Lorena Llosa is an Associate Professor in the Department of Teaching and Learning at NYU’s Steinhardt School of Culture, Education and Human Development. Her recent studies have focused on validity issues in the assessment of academic writing and the integration of language and content in instruction and assessment.

Thursday June 1 1:50 PM

Investigating the effects of various English varieties on listening comprehension test

Introducing non-native varieties of English into a second language (L2) listening comprehension test may enhance authenticity and broaden construct representation, but at the same time, it may cause test

bias against some test takers who do not share first language (L1) backgrounds with a speaker. Unfortunately, this dilemma for ESL listening test developers and users has not been resolved yet because previous research on the effects of shared L1 on L2 listening test scores has shown only conflicting results without explicating circumstances under which a shared-L1 effect takes hold. This study thus aims to investigate the potential for a shared L1 effect on TOEFL iBT listening test scores not only by doing detailed item analyses on two groups of L2 listeners from two different L1 backgrounds: Indian and Mandarin, but also by controlling key factors including texts, item types, and the degrees of intelligibility, comprehensibility and accentedness of L2 speech.

In this study, three native English, three Hindi and three Mandarin speakers were selected to represent different L1s and levels of intelligibility, comprehensibility and accentedness. Using these speakers and accents that are not currently included in the listening sections of the TOEFL iBT, 18 new listening passages were created on the basis of disclosed iBT listening test tasks. One-hundred and fifty undergraduate/graduate students enrolled in a large Mid-Western university first took the retired TOEFL iBT listening comprehension test comprising only native English speakers' recorded texts. This was used as an external measure of control for their English listening proficiency. They took another listening comprehension test consisting of American English-, Hindi-, and Mandarin- accented lectures, and then completed a posttest survey after they finished the test. Preliminary results suggest that both Indian and Mandarin listeners performed better on the items in which they listened to lectures recorded by native speakers or speakers who shared the same L1. In contrast, they performed worse when they listened to lectures recorded by speakers with relatively low- or mid-level comprehensibility and more accentedness. However, the relationship between accented lectures and test performance was not straightforward. It seems that item types mediated the impact of accented lectures on test takers' performance, which is in line with the findings of previous research (Harding, 2012).

Sunyoung Shin, Ryan Lidster, and Senyung Lee
Indiana University

Sun-Young Shin is Associate Professor in the Department of Second Language Studies at Indiana University. His research interests include test bias, standard setting, and rater behaviors. His work has been published in *Language Testing*, *Language Teaching Research*, *Assessing Writing*, *Language learning and Technology*, and *ReCALL*.

Thursday June 1 2:25 PM

Uneven profiles of test-takers' oral proficiency as a threat to validity in an EFL speaking test

Uneven development of subskills in language learners has been perceived as a threat to validity in language performance assessment when holistic scoring is practiced, but not so in the case of analytic scoring (Hughes, 2003). However, the prevalence of halo effects can eliminate this advantage of analytic scoring, because salient weakness on one aspect of test-taker performance frequently results in deflated scores on other aspects (Cai, 2015).

To test the hypothesis that uneven profiles pose a threat to validity even in the case of analytic scoring, the researcher selected the audio recordings of 12 test-takers' oral composition from the Test for English Majors, Band 4, Oral Test (TEM4-Oral), a major test for EFL learners in Chinese colleges.

Three rater-trainers were asked to rate these recordings analytically on three aspects—content, language use, and pronunciation—according to which different types of oral proficiency profiles were obtained. A sample of 51 raters then rated the same recordings analytically on the same three aspects. Mixed ANOVA was conducted to examine the interaction between aspect and type of profile. To explore the rating processes associated with uneven profiles, five raters participated in verbal protocols in which they justified their ratings of three test-takers.

The findings revealed the dual nature of the potential of analytic scoring to accommodate uneven profiles—the scores given by the raters generally retained the even-uneven distinction between the types of profiles, but the differences between the two form-related scores and the content score were reduced for most test-takers. The mixed ANOVA discovered significant interaction between aspect and type of profile. Post hoc comparison of interaction effects using estimated marginal means found that the form-deficient uneven profiles obtained a significant gain on the form-related scores with reference to the content score, whereas the content-deficient uneven profiles achieved a greater gain on the content score with reference to the form-related scores.

In line with the quantitative findings, the verbal protocols unveiled a tendency among the raters to expect even profiles rather than uneven profiles. While raters still distinguished between the three scales and tended to give different scores on different scales, the expectation of even profiles undermined the effort to differentiate the scales in the rating process.

These findings support the hypothesis that analytic scoring is not immune to uneven profiles. This has important implications for rater training and monitoring.

Hogwen Cai

Guangdong University of Foreign Studies

Hongwen Cai, an associate professor of Applied Linguistics in the Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies in China, obtained his PhD from UCLA in 2012. His current research interests include validity issues in speaking assessment and diagnostic and formative assessment of EFL listening.

Thursday June 1 3:25 PM

Defining constructs for summative vs. formative assessment using the interactionalist approach

The interactionalist approach to construct definition (Chapelle, 1998; Chalhoub-Deville, 2004; Bachman, 2007), which highlights the prominent role of language use context, has been explored with a primary focus on applications in a summative assessment context. Attention has been drawn to design challenges and considerations and generalizability and fairness issues associated with highly contextualized tasks. Implications of the interactionalist approach for formative assessment, however, have not been examined in a systematic way. This presentation will focus on how to apply the interactionalist approach in the design and development of speaking assessment for summative and formative purposes.

I will provide a brief orientation to the multi-faceted approach to construct definition that is aligned with the interactionalist view (Xi, 2015). When the interactionalist approach is used to define the

constructs of assessments for formative purposes, I argue that the constructs require a more dynamic and fluid characterization. The primary focus has shifted to defining a learning progression model that puts learners on a learning trajectory, which needs to reflect both expansion of language use contexts and growth in higher-order and foundational skills in relation to particular contexts.

In a summative context, major test design challenges include ensuring accessibility of contexts to test takers and figuring out the trade-off between rich representation of contextual factors in tasks and generalizability of tasks. In a formative environment, a central focus is on whether the feedback information provided by formative assessment helps us to understand how learners develop higher-order and foundation skills and expand their capability to apply these skills in increasingly more complex contexts. Given multiple assessment opportunities afforded by formative assessment and relaxed operational constraints, it is also possible to represent much richer contexts in designing formative assessment tasks. Accessibility of contexts, a major design issue associated with assessment for summative purposes, becomes less of a concern in a formative context as feedback about how test takers interact with different language use contexts is highly useful information that can be fed back into teaching and learning.

I will conclude the presentation with thoughts about key research areas that will help expand applications of the interactionalist approach in different assessment contexts.

Xiaoming Xi **Educational Testing Service**

Xiaoming Xi works at Educational Testing Service. Her research spans broad areas of theory and practice, including validity, fairness, test constructs, test design, scoring and technology. She is Co-editor of the Routledge book Series on “Innovations in Language Learning and Assessment”, and is Associate Editor of Language Assessment Quarterly.

Thursday June 1 4:00 PM

Fundamental Validity Issues in K-12 English Language Proficiency Assessments: Construct and Consequences

English learners (ELs) in U.S. public schools constitute the fastest growing segment of the nation’s K-12 population. These EL students deal with the dual challenge of learning rigorous material in all subjects while also developing the English language proficiency needed to meaningfully participate in academic contexts. To support EL students’ needs, the federal educational law (Every Student Succeeds Act, 2015 formerly No Child Left Behind Act) required that all states be held accountable for EL students’ progress in English language proficiency attainment (National Research Council, 2011). The federal law also stipulated that states must develop or adopt English language proficiency (ELP) standards and ELP assessments. As a result, large-scale, standards-based ELP assessments have been used with EL students in kindergarten (K) through grade 12 across the states.

In addition to accountability purposes, states use ELP assessments to determine EL students’ readiness to exit from EL status (Abedi, 2007; Wolf, Farnsworth, & Herman, 2008). This is a relatively high-stakes decision, as students will follow different academic paths depending on whether they are designated as ELs. Appropriate assessments of students’ ELP will provide useful information to guide

the types of services and support that individual students may need (e.g., dual language support, immersion program, ESL courses, and accommodations). On the other hand, faulty EL exit decisions resulting from inadequate ELP assessment results could lead to unequal opportunities to learn. To ensure that ELP assessments are used appropriately as they are intended for students and schools, empirical investigation of validity is of paramount importance.

In light of these considerations, this session focuses on two key, interrelated areas of validity concerns: the construct and the consequences of K–12 ELP assessments. Recently, both academic content-area standards and ELP standards have increased their rigor to prepare students for college and career. Accordingly, newly-developed ELP assessments contain more challenging tasks that measure higher-order language skills compared to assessments developed based on previous ELP standards. I will discuss how the ELP construct has been redefined and operationalized in recent K–12 ELP standards and assessments through examination of sample tasks. Then, I will describe what ramifications these new ELP assessments have for making high-stakes decisions about EL students. Drawing from a validation framework proposed by Bachman and Palmer (1996, 2010), I will also discuss how such fundamental construct validity issues are closely tied to consequential validity issues. I will conclude this session with a set of critical research areas pertaining to construct and consequential validity, which have important implications for practice and policies to support EL students' needs.

Mikyung Wolfe
Educational Testing Service

Mikyung Kim Wolf is a Managing Senior Research Scientist in the Center for English Language Learning and Assessment Research at ETS. Wolf's research areas include technology-enhanced language assessments, alignment between assessments and standards, formative assessment, and validity issues in assessing K-12 English language learners in the United States as well as worldwide.

Thursday June 1 4:35 PM

The development of a curriculum-based adaptive diagnostic testing system for young EFL learners in Taiwan

Diagnosis and Certification of English Competence (DCEC) is a computerized testing system that aims to adaptively diagnose the English competences of elementary and middle school students in Taiwan for the purpose of fulfilling the local need for remedial instruction. Developed and administered by the Research Center for Psychological and Education Testing at National Taiwan Normal University, DCEC consists of subtests in four skills (vocabulary, grammar, reading, and listening comprehension) at five levels, from grade four to nine, for a total of twenty tests. The test suites were developed based on the common core standards of the Grade 1-9 curriculum and the classroom performance standards outlined in the Standard-based Assessment of Student Achievements for Junior High School Students (SBASA). DCEC enables the identification of student achievement by grade level and outlines multi-dimensional diagnostic feedback that aligns with the curriculum content and instruction. The system incorporates IRT and standard-setting methods to design its adaptive scheme for different subtests for the purpose of curriculum alignment. In the system, test takers receive different combinations of screening and diagnostic testlets to determine their grade level and performance on different content dimensions. Moreover, which combinations of testlets to receive depends on whether their performance meets the standard-setting criteria established by panels of judges for each testlet. The presentation will

introduce the system, its design principles, and development processes. It will demonstrate the online system and its comprehensive diagnostic feedback report, outline the validation evidence collected, and end with challenges encountered during the development and administration processes.

Hsinmin Liu
National Taiwan Normal University

Hsinmin Liu received her PhD in Applied Linguistics at UCLA in 2014. She currently works as a test developer and researcher at National Taiwan Normal University, where she is assisting in the development of computerized language tests in English and Chinese. She is most interested in issues and methods related to narrowing the gap between real-life language use and the traditional language testing systems in Taiwan.

Friday June 2 9:15 AM

Do speaking assessments which require listening differentially impact the scores of more and less proficient test takers?

Increasingly, tasks which aim to assess speaking ability require test takers to listen as well as speak. For instance, both integrated listen-speak tasks and interactive tasks, such as paired and group tests require listening as well as oral production. It seems apparent that listening should be considered part of the construct when such tasks are used, but research suggests that it may be that only oral production matters (e.g., Lee, 2006). This research aimed to shed light on this issue by determining the extent to which test takers with different oral abilities might be differentially impacted by speaking tasks that require listening and ones that do not. Scores from a paired oral presentation, group oral, individual picture and graph description, and TOEFL iBT integrated and independent speaking tasks were analyzed for this purpose. Data were from 222 Japanese university students, who were studying English as a foreign language. A mixed effects ANOVA, with listening (tests which require listening versus ones that do not) as a within subjects variable and oral ability (high and low) as a between subjects variable, was used for the analysis. Results showed an ordinal interaction between listening and oral ability, indicating that the tasks which were expected to require listening were differentially more difficult for the low level learners. In addition, a main effect for listening indicated that the test tasks which were expected to require listening were more difficult than the ones that did not. The construct of speaking ability will be discussed in light of the changes in the task types that are being used to assess it.

Gary Ockey
Iowa State University

Gary Ockey, Associate Professor of Applied Linguistics and Technology, investigates second language assessments, with a focus on the use of technology and quantitative methods to better measure oral communication. He has served as the Editor of the *TOEFL Research Report Series*, and is currently an associate editor of *Language Assessment Quarterly*.

Friday June 2 9:50 AM

How Generalizable Are Different Components of the Speaking Construct across Tasks?

The generalizability of test performance on speaking tasks has been a great concern in second language (L2) speaking assessment. The current study attempts to examine the generalizability of the different ability components of L2 speaking construct as well as the effects of speaking tasks on each ability component (e.g., grammar, vocabulary, pronunciation). Nine hundred and ninety-nine students at a Hong Kong university were asked to take a computer-based English as a Second Language (ESL) speaking test with five different tasks (e.g., summarization, role play, description). Their spoken responses were rated with an analytic rating scale which included five aspects of the L2 speaking construct: 1) Task Fulfilment and Relevance, 2) Grammar and Vocabulary, 3) Pronunciation, 4) Confidence and Fluency, and 5) Clarity of Presentation. A mixed-methods approach was adopted to look into the issue of score generalizability and task effects on each ability component. Confirmatory factor analysis results showed that the superior fit of a Higher-order trait-Unrelated method model validated the test design and provided the most parsimonious explanation of the relationships among the five dimensions and overall speaking proficiency. The analytic scores were found to have much larger factor loadings on the trait factors than on the method factors. However, Task 4 which asked students to leave a telephone message, showed relatively large effects on the ability components measured by this task. IRT analyses indicated that this task was the easiest one among all the five tasks and did not discriminate among examinees as well as the other tasks. Generalizability (G) theory analyses revealed that Grammar and Vocabulary and Pronunciation proved to be the most dependable dimensions while Task Fulfillment and Relevance had the largest task (t) and person-by-task (pt) variance components. The large task or person-by-task effects were considered as a significant threat to the generalizability of test scores and the validity of score interpretations and uses. The possible reasons behind the large errors due to tasks are discussed. Implications are also provided for the task design and development of the speaking tests (e.g., appropriateness of difficulty levels of the speaking tasks).

Yujie Jia

Defense Language Institute, Foreign Language Center

Yujie Jia, Ph.D., works at the Testing Division of Defense Language Institute Foreign Language Center as the project lead to build a validity argument for the ILR OPI. She earned her degree in Applied Linguistics at UCLA with a specialization in language assessment. Prior to joining DLIFLC, she worked on the development and validation of the automated speaking proficiency tests at Pearson.

Friday June 2 10:50 AM

Generalizability of content analytic rating scales for assessing university-level Japanese EFL learners' summarization performance

The last decade witnessed a rapid increase of studies that examined the process and product of L2 learners' performance on academic writing tasks based on source texts from a variety of perspectives (Plakans, 2015). Despite the rapid advancement of research in this area, recent studies have focused mainly on complex task types requiring a synthesis of multiple source texts rather than straight summarization of a single source text. Moreover, partly reflecting the conceptualization of source-based writing scores as measures of writing in the field, only a relatively small number of previous studies (e.g., Plakans & Gebril, 2013; Sawaki, 2003; Sawaki, Quinlan, & Lee, 2013) have examined the

content of learner-produced source-based writing task responses. However, the ability to accurately and yet succinctly represent source-text content in its summary is an indispensable part of basic academic literacy for L2 writers. Accordingly, it is imperative to design and develop assessment criteria for summary content that are easy to implement and yet exhibit sufficient generalizability, the use of which could enhance the representation of the source-based writing construct in summarization performance feedback provided to learners.

This study aimed at examining L2 summarization performance profiles of Japanese students majoring in English at a private university in Tokyo. A total of 162 students enrolled in academic English courses wrote summaries in English on two out of four source texts at two difficulty levels randomly assigned to them. Based on qualitative analyses of expert summaries and discussions with academic writing course instructors, analytic rating scales were developed for language use as well as for content coverage/accuracy and integration based on van Dijk and Kintsch's (1983) macrorules for synthesizing source text content. Two independent raters scored each learner-produced summary based on the above-mentioned rating scales.

This presentation summarizes results of multivariate generalizability theory analyses conducted for examining the effects of source-text difficulty levels, source texts, and raters on score generalizability and the relationships among the above-mentioned analytic rating scales. Implications of the study findings for designing diagnostic performance feedback for university-level Japanese EFL learners on summary content and other aspects of summarization performance (e.g., source use, cohesion) will be discussed.

Yasuyo Sawaki
Waseda University

Yasuyo Sawaki, Professor, the Faculty of Education and Integrated Arts and Sciences, Waseda University, Tokyo, earned her Ph.D. from UCLA in 2003. After working as a research scientist at Educational Testing Service for five years, Sawaki has been teaching various undergraduate- and graduate-level courses in language assessment, teacher training, and academic English.

Friday June 2 11:25 AM

Test-ifyze: Software for Assessment Design and Instruction

Bachman and Palmer (2010) described a justification-based approach to assessment development employing four documents: (a) justification (the Assessment Use Argument [AUA] and backing), (b) initial planning, (c) design statement, and (d) blueprint. Five years of experience in teaching the material in the book and using it to design language tests have shown the traditional book format to be challenging. Primarily, the challenges stem from the fact that the process of using the four documents in test design and development is somewhat non-linear and that there is a large amount of information that needs to be kept track of across the four documents.

Early on in their work with the book, Bachman and Palmer realized that a well- designed software program might be useful in helping address the challenges associated with tracking information across the four documents. Using software would allow designers to focus more on the development of the assessment rather than the organization of the information. However, their initial attempts to find

competent software designers proved both very expensive and, ultimately, futile.

In 2012, Palmer and one of his students, Daniel Dixon, started work to develop more advanced software to manage the document creation process and automate the organization of information. The current software, *Test-ifyze* (which uses the Microsoft's InfoPath® application) provides an interface that is intuitive, easy to learn, and easy to use. Language testers who are familiar with MSW documents and typical web-based form completion tasks can easily navigate and use the software. The purpose of this presentation is to show potential users how the software works and what it accomplishes that a paper-based approach cannot.

In this presentation we will first show how the *Test-ifyze* interface looks and how basic operations are used to create documents and manage interactions among them. Then we will show an example of how *Test-ifyze* is used to create a type of assessment task by modifying the characteristics of a target language use task template. Participants will be able to access the software on line and try it out during the conference.

Adrian Palmer, Daniel Dixon
University of Utah

Adrian Palmer teaches a variety of courses in L2 teaching and language assessment. One of his recent books, *Language Assessment in Practice* (Oxford University Press), 2010, recently received the 2012 SAGE/ILTA award for the best book in the area of language testing/assessment (2009-2012).

Friday June 2 1:50 PM

Internationally-developed or locally-developed language assessments?

The blueprints for the design, development, administration, scoring and reporting of well-known international English language proficiency assessments are developed and operationalized into products in hothouse laboratories. These products are then prepared for export and consumption to nations, regions, or city schools hundreds or thousands of miles away. The main assumption made in the development of these products is that the end-user is a consumer, local expertise is most likely not available, and that the products do not need to fit the local circumstances perfectly. This line of thinking no doubt will help many local and regional schools who may find their capacity to develop language assessments difficult mainly due to the lack of expertise and resources.

On the other hand, local agencies are developing language assessments or adapting international assessments. These assessments are likely to have the active involvement of local experts, acceptance from local language teachers and local government, and the greater likelihood that these products would meet local priorities, resources, legislation and policies. But will such locally-developed language assessments be constrained by local circumstances to the extent that they will have to modify the construct, constrain the content, restrict the task format, re-conceptualize concepts such as validity, fairness, and relativize ethical matters such as justice? This issue also brings to the fore the question of whether standards for assessments (such as the APA, AERA, NCME Standards, 2014) written in a universalist fashion ought to be the standard for global language assessments or whether contextually-bound local standards ought to be the standard for local language assessments. This talk will explore this topic with illustrations from the field of language assessment.

Antony Kunnan
University of Macao

Antony John Kunnan, Professor of English at the University of Macao, was Lyle's first Ph.D. student at UCLA. He worked with Lyle on the TOEFL-Cambridge Comparability study that became his dissertation topic. Antony and Lyle have worked on many projects together, including the *Statistics for Language Assessment Workbook*. Antony's new book *Evaluating Language Assessments* (Routledge, 2017) expands on Lyle's Assessment Use Argument.

Friday June 2 2:25 AM

The Curse of Explanation: Model Selection in Language Testing

Statistical models in language testing aim to explain. Language testing researchers have used models, such as linear regression and factor analysis, to provide explanations for, and evaluate hypotheses about relationships between theoretical constructs. The translation of a substantive theory into an estimable and testable model is seldom straightforward, and therefore, the same theory can yield different models. Often these models provide alternate explanations. The pursuit of explanation dictates that the researcher select one best model and interpret it without considering the uncertainty in the model selection process. This can be a curse. Theories behind model selection are complex (e.g., Leeb & Potscher, 2005). In practice, model selection can be dangerous.

The focus on explanation used to be universal (Breiman, 2001), and still appears the norm in social science disciplines (Shmueli, 2010). However, over the last two decades, prediction of future or unseen events has become the default goal of modeling in certain fields (e.g., computer science) with impressive results. It is now well understood how modeling for explanation and prediction is different (e.g., Berk, 2016; Shmueli, 2010). One of the key differences is model selection. Prominent modeling-for-prediction approaches are specifically designed to avoid model selection. The rationales for these approaches and their success illustrate why model selection can be dangerous in practice. If model selection is essential for modeling in language testing, it is important to be aware of its danger.

In this presentation, I will introduce the challenges of model selection within the context of language testing research, and provide several suggestions to address them. I will illustrate how popular tools for model selection, such as model diagnostics and fit indices, are agnostic about the “true” model and the process of model comparison and selection. I will also show how models with substantive differences can yield almost identical data, and therefore, cannot be reliably distinguished based on data. Both simulated and real-world data will be used to clearly demonstrate the challenges while contextualizing the demonstration. I will conclude with suggestions to address the model selection challenges within the modeling-for-explanation approach and to leverage the success of the modeling-for-prediction approach.

Ikkyu Choi
Educational Testing Service

Ikkyu Choi is an associate research scientist in the Research and Development division at Educational Testing Service. He earned his PhD from the University of California, Los Angeles, with a specialty in

language assessment. His research interests include second language development profiles, test taking processes, and quantitative research methods for language testing data.

Friday June 2 3:25 AM

Some reflections on mentoring students

Lyle Bachman

University of California, Los Angeles

Lyle Bachman has received numerous honors from universities and professional organizations and has published widely in the areas of language testing, educational measurement, and Applied Linguistics. His current activities include consulting with test developers on practical test development issues and training classroom language teachers in classroom-based language assessment.

A Special Thanks to our Sponsors!

Institutions:



Listening. Learning. Leading.

Individuals:

Youngsoon So

