LTRC 2019

41st Language Testing Research Colloquium

Atlanta, Georgia, USA

LANGUAGE TESTING AND SOCIAL JUSTICE

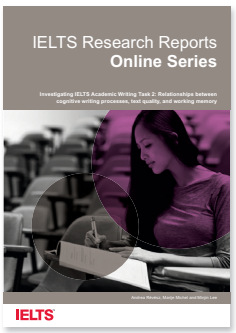
ILTA
INTERNATIONAL LANGUAGE TESTING ASSOCIATION

4-8 MARCH · 2019

IELTS™

IELTS Research Reports Online Series

IELTS Partnership Research Papers



available now at

ielts.org/research

BRITISH COUNCIL idp Cambridge Assessment English

Thank you to our sponsors!

Platinum Sponsors

Gold Sponsors

CAL CENTER FOR APPLIED LINGUISTICS

BRITISH COUNCIL
ASSESSMENT RESEARCH GROUP

Paragon
TESTING ENTERPRISES
CANADA'S *LEADER* IN ENGLISH LANGUAGE TESTING

duolingo english test

Silver Sponsors

ITTC® GEPT®
財團法人 語言訓練測驗中心

AELRC
Assessment and Evaluation Language Resource Center

OET
OCCUPATIONAL ENGLISH TEST

UNIVERSITY OF MICHIGAN
MICHIGAN LANGUAGE ASSESSMENT

TABLE OF CONTENTS

Thank you to the Assessment and Evaluation Research Center (AELRC) at Georgetown University for sponsoring the printing of the program book.

MESSAGE FROM THE ILTA PRESIDENT

On behalf of the International Language Testing Association (ILTA) Executive Board, welcome to the 41st Annual Language Testing Research Colloquium in Atlanta! It is a personal treat to welcome you to the US South, which I have called home for over 13 years now. So to welcome you Southern-style, I need to say*: Welcome all y'all!*

Given this year's conference theme, Language Testing and Social Justice, Atlanta is a particularly fitting venue for the conference. Atlanta is home of Dr. Martin Luther King, Junior, and a major center of the Civil Rights movement.

While it has been less than a year since we last convened, I trust you are as excited as I am to partake in all that our premier international conference offers. With LTRC, you are invited to engage with thought-provoking scholarship, deepen ties with the field's leading professionals, celebrate colleagues and friends, and get to know a place you may not have visited before.

Given our professional interest in languages, I thought you might appreciate the description of the term *all y'all.*

The *Dictionary of Southern Slang*, states: **All Y'all** Etymology: Intensive form of y'all

This usage states "you all" more emphatically. For example, saying "I know y'all," would mean that one knows a group of people; saying, "I know all y'all" would mean that one knows the members of the group individually.

The *all y'all* rolls off your tongue more easily if you're sipping sweet tea or mint julep!

Thank-yous are in order to all those who have worked hard to put the conference together. Special thanks to the committee co-chairs: Sara Cushing and Meg Malone. We owe thanks also to the other members of the organizing committee: Jee Wha Dakin, Barbara Dobson, Gad Lim, Elvis Wagner, Paula Winke, and Meg Montee. Your diligence and leadership are vital to the Association. Thank you to all the organizations that have provided support to this year's conference. Our appreciation goes to Cambridge Assessment English-IELTS, ETS-TOEFL, along with the British Council-Assessment Research Group, Center for Applied Linguistics, Duolingo English Test, Paragon Testing Enterprises, the Taiwan Language Training and Testing Center-GEPT, the Occupational English Test, the Assessment and Evaluation Language Resource Center, and the University of Michigan-Michigan Language Assessment. Finally, I would like to thank all of you for incurring the hassles and expense of travels to come and join the conference this year. On behalf of ILTA, thank you!!!

ILTA's pride and joy has always been its conference, LTRC. I trust that you will be able to take advantage of all the conference has to offer. The conference program includes a wide range of stimulating papers, posters and works-in-progress, and other presentations. Here, I would like to highlight the ETS-sponsored Samuel J. Messick Memorial Lecture by Joan Herman. Joan is Co-Director Emeritus of the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) at UCLA. The program also features a talk by Cathie Elder who will give the British Council-sponsored Alan Davies lecture. Also to note is the address by Dan Douglas, this year's winner of the Cambridge/ILTA Distinguished Achievement Award. Trust that you will

attend many of the presentations offered and find the research to be thought-provoking and the ensuing discussions enriching.

Among the exciting sessions scheduled at LTRC this year is the newcomer session. As per our tradition, all attending LTRC for the first time are encouraged to come to the Language Learning and Testing Foundation-sponsored session at the beginning of the conference where they get to meet one another and learn about ILTA and LTRC. It is important, in my mind, that while we look back, we also look forward, especially with the new members who are the future of the organization. This year, therefore, we are experimenting with a new format for this session. The *NEWCOMER WELCOME & INTRODUCTION TO ILTA STRATEGIC PLANNING* SESSION ON TUESDAY, MARCH 5, 4:00 - 5:15 PM allows us to welcome newcomers as well as to think strategically about our organization. The new format of structured conversations among newcomers, past and current ILTA members, as well as all LTRC attendees will help us identify strengths, growth opportunities, and aspirations for new impact areas that fulfill ILTA's mission and ensure its continued vitality. So please note that ALL Y'ALL at the conference are invited.

Please make time to attend the Annual Business Meeting of the ILTA Membership on Thursday, March 7, 12:00 - 1:45 PM. ILTA conference attendees are invited to join the meeting where they can partake in ILTA's strategic deliberations, hear updates on the Association's latest initiatives, determine how to become more involved, and make plans for future LTRCs. I take this opportunity to remind you, if you have not yet done so, to JOIN ILTA. You can easily join online at www.iltaonline.com/page/MembershipPlans. Alternatively, you can go to the Conference Reception desk where Terry Dougherty and Michele Doyle from Nardone, our management company, can help you with the process.

Be on the lookout for opportunities to unwind, socialize, and have fun with colleagues/friends—old and new! Make travel plans so you are sure to attend the opening sessions of the conference and to join me at the Opening Reception on Tuesday at 7:00 PM, held at the Dekalb History Center (Historic Decatur Courthouse). Also make sure you have tickets for the conference Banquet on Thursday at 7:00 PM at The Trolley Barn. I am told that in addition to honoring and celebrating our various awardees, the program will include *dancing*—so remember to pack your dance shoes!

Living in the US South, I have been to the Atlanta airport countless times, but have never had the opportunity to visit this outstanding city, which has many sites to explore… I encourage you to visit sites such as the Martin Luther King Junior national historical Park, the High Museum of Art, the Centennial Olympic Park, CNN, Coca-Cola, LEGOLAND, and the many wonderful neighborhoods and streets.

- Feel free to use the Facebook FRAME—created for LTRC 2019 by Erik Voss: www.facebook.com/profilepicframes/?selected_overlay_id=331546674237151
- Share ILTA and LTRC 2019 news on the Twitter account: @ILTAonline
- A WeChat account is being developed

Sincerely, Bien sincèrement, 诚挚地,
Vänliga hälsningar, بإخلاص

Micheline Chalhoub-Deville
ILTA President

WELCOME FROM THE CO-CHAIRS

Dear LTRC Participants,

Welcome to Decatur! It's wonderful to see you here after two years of planning and organizing for this event.

Our conference theme, *Language Testing and Social Justice*, explores the roots of language, language testing, and how we can meet the needs of our stakeholders within our different contexts. We look forward to discussing and engaging these crucial topics with you this week.

We want to thank our amazing organizing committee: Jee Wha Dakin, Barbara Dobson, Gad Lim, Meg Montee, Elvis Wagner and Paula Winke. Their incredible hard work shows the importance of our language testing community to our colleagues across the world who work in and for language testing organizations and higher education. They have organized the workshops that will enhance your professional development and have reviewed and finalized the stimulating program of demonstrations, works-in-progress, posters, symposia, and papers that we will see this week. We are especially glad to introduce networking dinners at this year's conference, and our hope is that these dinners will allow us to discuss issues of language testing and social justice, our research goals, and ways we can collaborate to improve language testing- and its sisters, language teaching and learning- for all speakers of all languages across a diversity of contexts.

Next, we want to thank our workshop presenters: Geoff LaFlair, Dan Isbell, Troy Cox, Darren Perrett, Brigita Séguis, Tim McNamara and Elana Shohamy. Whether you are learning about R, jMetrix, corpora or the history of civil rights, we are sure that their talents and skills will contribute to your work and our field.

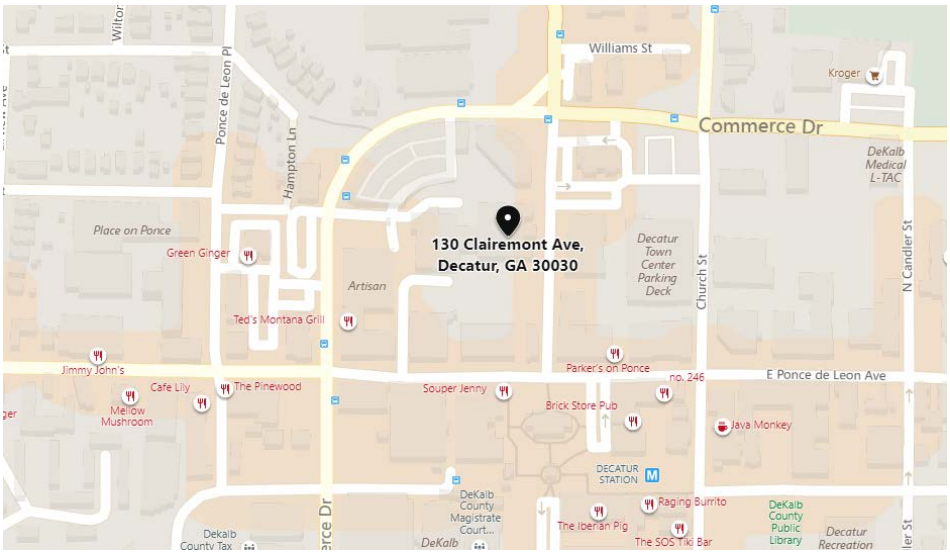
We also want to extend a warm thanks to the ILTA Board: Micheline Chaloub-Deville, Cathie Elder, Ute Knoch, Benjamin Kremmel, Yan Jin, Gerriet Janssen, and Mikyung Wolf for their commitment to the field and their hard work in leading ILTA. We are especially grateful to Jay Banerjee, who helped us manage the budget and work creatively to bring you a terrific conference at a reasonable cost. Many thanks as well to Terry Dougherty and Michele Doyle of Nardone, ILTA's association management company, for their cheerful, can-do attitude toward all things LTRC, ILTA, food, and fun.

Finally, we want to thank you! There is no conference without participants, and there are no livelier nor more engaged conference attendees than those at LTRC. Whether we are making an extrapolation argument, inquiring about methodology, or catching up about our non-testing lives, LTRC is the event of the year for many of us in language testing. We have worked not just on your behalf but because we love the energy and enthusiasm each of you brings to this meeting.

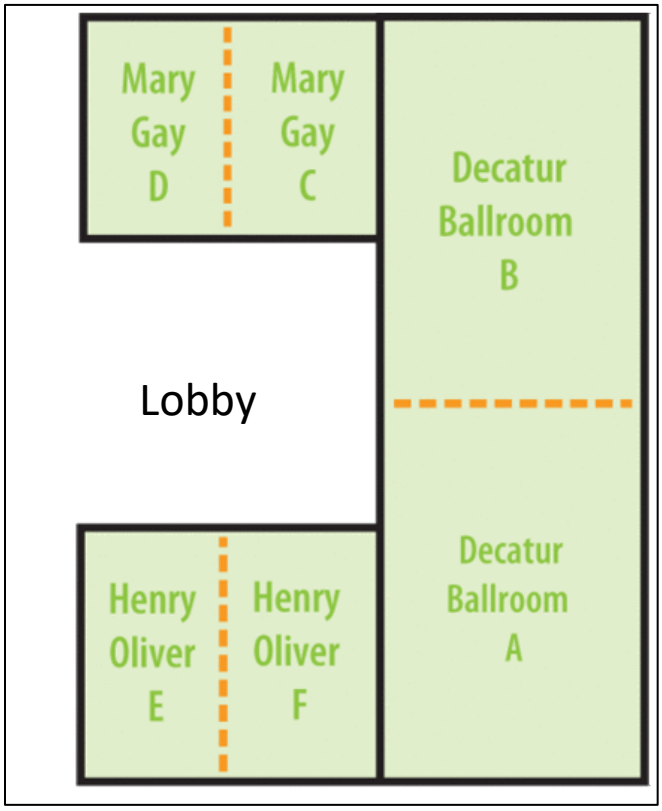
Thank you and enjoy!

Sara Cushing and Margaret (Meg) Malone
LTRC 2019 Conference Co-chairs

Courtyard Atlanta Decatur Downtown/Emory
130 Clairemont Avenue
Decatur, Georgia 30030
Direct: +1 404-371-0204



Map of Conference Rooms



CONFERENCE ORGANIZATION

Co-chairs

Sara Cushing
Georgia State University

Margaret E. Malone
Georgetown University

Organizing Committee

Jee Wha Dakin
Educational Testing Service

Barbara K. Dobson
Michigan Language Assessment

Gad S. Lim
Michigan Language Assessment

Megan Montee
Center for Applied Linguistics

Elvis Wagner
Temple University College of Education

Paula Winke
Michigan State University

Student Volunteers

Analynn Bustamante
Sanghee Kang
Xian Li
Yunjung Nam
Haoshan Ren
Yi Tan
Rurik Tywoniw
Soohye Yeom

Abstract Reviewers

Beverly Baker
Khaled Barkaoui
William Bonk
Tineke Brunfaut
Nathan T. Carr
Micheline Chalhoub-Deville
Carol Chapelle
Mark Derek Chapman
Yeonsuk Cho
Ikkyu Choi
Martyn Clark
Deborah Crusan
Alister Cumming
Sara Cushing
Jee Wha Dakin
John H.A.L. de Jong
Barbara K. Dobson
Dan Douglas
Catherine Anne Elder
Jason Fan
Timothy Farnsworth
Ardeshir Geranpayeh
April Ginther
Kirby Grabowski
Tony Green
Peter Gu
Luke Harding
Claudia Harsch
Ching-Ni Hsieh
Yo In'nami
Eunice Eunhee Jang
Gerriet Janssen
Yan Jin
Dorry Kenyon
Ute Knoch
Rie Koizumi
Benjamin Kremmel
Antony Kunnan
Yong-Won Lee
Constant Leung
Gad S. Lim
Lorena Llosa
Sari Luoma
Susy Macqueen
Margaret E. Malone
Lyn May
Tim McNamara
Fumiyo Nakatsuhara
Heike Neumann
Sally O'Hagan
Barry O'Sullivan
Gary Ockey
Spiros Papageorgiou
Lia Plakans
India Plough
David D. Qian
Daniel Reed
Stephanie Lee Rummel
Shahrzad Saif
Nick Saville
Yasuyo Sawaki
Jonathan Schmidgall
Sun-Young Shin
Charles Stansfield
Ruslan Suvorov
Masanori Suzuki
May Tan
Martha Isabel Tejada
Veronika Timpe-Laughlin
Dina Tsagari
Carolyn Turner
Alan Urmston
Alistair Van Moere
Erik Voss
Elvis Wagner
Yoshinori Watanabe
Mikyung Kim Wolf
Guoxing Yu

ILTA Executive Board 2019

President: Micheline Chalhoub-Deville (University of North Carolina, Greensboro, USA)

Immediate Past President: Catherine (Cathie) Elder (University of Melbourne, Australia)

Vice-President: Sara Cushing (Georgia State University, Atlanta, Georgia, USA)

Secretary: Ute Knoch (University of Melbourne, Australia)

Treasurer: Jayanti (Jay) Banerjee (Trinity College, London, UK)

Members at Large

Benjamin Kremmel (University of Innsbruck, Austria)

Yan Jin (Shanghai Jiao Tong University, China)

Gerriet Janssen (Universidad de los Andes, Colombia)

Mikyung Kim Wolf (Educational Testing Service, USA)

ILTA Staff

Terry Dougherty (Association Manager)
Michele Doyle (Association Assistant)

LTRC 2019 Co-Chairs

Sara Cushing (Georgia State University, Atlanta, Georgia, USA)

Margaret E. Malone (Georgetown University, USA; American Council on the Teaching of Foreign Languages)

ILTA Nominating Committee 2019

Chair: Ikkyu Choi (Educational Testing Service, USA)

Kellie Frost (University of Melbourne, Australia)

Yo In'nami (Chuo University, Japan)

Ruslan Suvorov (University of Hawai'i at Mānoa, USA)

AWARD COMMITTEES

Cambridge/ILTA Distinguished Achievement Award

Chair: Carolyn Turner (McGill University, Canada)

Micheline Chalhoub-Deville (University of North Carolina, Greensboro, USA)

Yong-Won Lee (Seoul National University, South Korea)

Nick Saville (Cambridge Assessment English, UK)

ILTA Student Travel Awards

Chair: Cathie Elder (University of Melbourne, Australia)

Claudia Harsch (University of Bremen, Germany)

Benjamin Kremmel (University of Innsbruck, Austria)

Lado Award (to be awarded at LTRC)

Chair: Beverly Baker (University of Ottawa, Canada)

William Bonk (Pearson, USA)

Jason Fan (University of Melbourne, Australia)

Peter Lenz (Universite de Fribourg, Switzerland)

Lorena Llosa (New York University, USA)

Megan Montee (Center for Applied Linguistics, USA)

Lia Plakans (University of Iowa, USA)

Yasuyo Sawaki (Waseda University, Japan)

ILTA Best Article Award 2016

Chair: Yan Jin (Shanghai Jiao Tong University, China)

Tim McNamara (University of Melbourne, Australia)

Troy Cox (Brigham Young University, USA)

Atta Gebril (American University in Cairo, Egypt)

Slobodanka Dimova (University of Copenhagen, Denmark)

AWARDS

ILTA Student Travel Awards, 2019

Hyunah Kim, University of Toronto, Canada

Natalia de Andrade Raymundo, University of Campinas, Brazil

Jacqueline Ross TOEFL Dissertation Award

2019 Winner: Dr Saerhim Oh. *Investigating test-takers' use of linguistic tools in second language academic writing assessment.* Teachers College, Columbia University, USA
Supervisor: James E. Purpura

Caroline Clapham IELTS Masters Award, 2018

Chi Lai Tsang *Examining Washback on Learning from a Sociocultural Perspective: The Case of a Graded Approach to English Language Testing in Hong Kong*. University College London (UCL), UK

Supervisor: Talia Isaacs

ILTA Best article award 2017

Winner not yet announced at time of printing

The Davies Lecture Award

Cathie Elder (University of Melbourne, Australia)

Bringing Tests to Justice: Can We Make a Difference?

Samuel J. Messick Memorial Lecture Award

Joan Herman (University of California, Los Angeles, USA)

Testing and the Public Interest

Cambridge/ILTA Distinguished Achievement Award

Dan Douglas (Iowa State University, USA)

Context, Language Knowledge, and Language Use: Current Understandings

Robert Lado Memorial Award

To be announced at the LTRC banquet

TIRF 2017 Doctoral Dissertation Grant Awardees in Language Assessment

Jenna Altherr Flores
Doctoral University: University of Arizona
Dissertation Title: *Multimodality, Social Semiotics, and Literacy: How LESLLA Learners from Refugee Backgrounds Make Meaning in Assessment Texts*
Research Supervisor: Dr. Chantelle Warner, University of Arizona

Mingxia Zhi
Doctoral University: The University of Texas at San Antonio
Dissertation Title: *Investigating the Authenticity of Paper- and Computer-Based ESL Writing Tests*
Research Supervisor: Dr. Becky Huang, The University of Texas at San Antonio

Minkyung Kim
Doctoral University: Georgia State University
Dissertation Title: *Assessing Second Language Writing in Higher Education: A Longitudinal Study*
Research Supervisor: Dr. Scott Crossley, Georgia State University

Panjanit Chaipuapae
Doctoral University: Northern Arizona University
Dissertation Title: *Accents and Workplace Listening Comprehension of Thai Undergraduates in Asian ELF Contexts*
Research Supervisor: Dr. Joan Jamieson, Northern Arizona University

Parisa Safaei
Doctoral University: Laval University
Dissertation Title: *Investigating the Interactiveness of IELTS Academic Writing Tasks and Their Washback on EFL Teachers' Test Preparation Practices*
Research Supervisor: Dr. Shahrzad Saif, Laval University

Salomé Villa Larenas
Doctoral University: Lancaster University
Dissertation Title: *Exploring the Language Assessment Literacy of Teacher Educators in Chile*
Research Supervisor: Prof. Tineke Brunfaut, Lancaster University

Sonca Vo
Doctoral University: Iowa State University
Dissertation Title: *Effects of Task Types on Interactional Competence in Oral Communication Assessment*
Research Supervisor: Dr. Gary Ockey, Iowa State University

CONFERENCE SCHEDULE

Monday, March 4

Pre-Conference Workshops

Time	Event	
8:00am to 3:00pm	**Conference registration open**	
9:00am to 4:00pm	**Workshop 1 (Day 1): *R & RStudio for Reproducible Language Test Analysis, Research, and Reporting*** Leaders: Geoffrey T. LaFlair, Duolingo and Daniel R. Isbell, Michigan State University Location: Mary Gay C	**Workshop 2: *Constructing Measures: An Introduction to Analyzing Proficiency Test Data Using IRT with jMetrik*** Leader: Troy L. Cox, Center for Language Studies Location: Henry Oliver

Tuesday, March 5

Pre-Conference Workshops & Events

Time	Event	
8:00am to 7:00pm	**Conference registration open**	
9:00am to 4:00pm	**Workshop 1 (Day 2): *R & RStudio for Reproducible Language Test Analysis, Research and Reporting*** Leaders: Geoffrey T. LaFlair, Duolingo and Daniel R. Isbell, Michigan State University Location: Mary Gay C	**Workshop 3: *Corpus-based development and validation of language tests: Using corpora of and for language testing*** Leaders: Darren Perrett, Cambridge Assessment English and Brigita Séguis, Cambridge Assessment English Location: Henry Oliver
Time	**Event**	
9:00am to 1:00pm	**Workshop 4: *The Civil and Human Rights of Language: Guided Tour of the Atlanta Civil & Human Rights Museum*** Leaders: Elana Shohamy, Tel Aviv University and Tim McNamara, University of Melbourne Meeting Location TBD	
Time	**Event**	
11:00am to 3:30pm	**ILTA Pre-Conference Executive Advisory Board Meeting** Location: TBD	
4:00 to 5:15pm	**Newcomer & Strategic Planning Session** Location: Decatur B	

Tuesday, March 5

Time	Event
5:15 to 6:45pm	**Alan Davies Lecture** ***Bringing tests to justice: can we make a difference?*** Dr. Catherine Elder, University of Melbourne *Sponsored by the British Council* Location: Decatur B
7:00 to 8:30pm	**Opening Reception** Historic DeKalb Courthouse (101 East Court Square, Decatur)

Wednesday, March 6

Time	Event		
7:15 to 5:00pm	**Conference registration open**		
7:30 to 8:30am	**Latin American Association of Language Testing & Assessment Meeting** Location: Decatur B		
	Concurrent Sessions: Demonstrations		
Time	**Mary Gay C**	**Henry Oliver**	**Decatur A**
8:00 to 8:30am	***Automated essay scoring: An objective support to human raters*** Matthew Kyle Martin, Matthew Wilcox	***Talk2Me Jr: A Digital Language and Literacy Assessment Tool*** Samantha Dawn McCormick, Hyunah Kim, Jeanne Sinclair, Clarissa Lau, Megan Vincett, Christopher Douglas Barron, Eunice Eunhee Jang	***Personalized language learning as language assessment: A case study of two large learner corpora*** Burr Settles, Masato Hagiwara, Erin Gustafson, Chris Brust
8:30 to 10:35am	**Opening Symposium:** ***Local needs and global priorities in ensuring fair test use: synergies and tension in balancing the two perspectives*** Jamie Mark Dunlea, Jessica Wu, Yan Jin, Wei Wang, Haoran Yang, Quynh Nguyen, Barry O'Sullivan Location: Decatur B		

Wednesday, March 6

Time	Event			
10:35 to 10:55am	**Coffee Break** Location: Lobby			
Concurrent Sessions: Papers				
Time	**Mary Gay C**	**Henry Oliver**	**Decatur A**	**Decatur B**
10:55 to 11:25am	***Understanding teacher educators' language assessment literacy practices while training pre-service English teachers in Chile*** Salomé Villa Larenas	***Reading Self-Concept and Reading Achievement in Monolingual and Multilingual Students: A Cross-Panel Multiple-Group SEM Analysis*** Christopher Douglas Barron	***Rater cognition and the role of individual attributes in rating speaking performances*** Kathrin Eberharter	***Test Consequence on Rural Chinese Students: Investigating Learner Washback of National College Entrance English Exam*** Yangting Wang, Mingxia Zhi
11:30am to 12:00pm	***Using the Language Assessment Literacy Survey with an Under Researched Population: The Case of Uzbekistan EFL Teachers*** David Lawrence Chiesa	***What did the Reading Assessment Miss?*** Elizabeth Lee	***The Effect of Raters' Perception of Task Complexity on Rater Severity in a Second Language Performance-based Oral Communication Test*** Yongkook Won	***Examining Washback on Learning from a Sociocultural Perspective: The Case of a Graded Approach to English Language Testing in Hong Kong*** Chi Lai Tsang
12:00 to 1:30pm	**Lunch Break on your own** *Language Testing* Editorial Board Meeting, Location: Henry Oliver			

Wednesday, March 6

Concurrent Sessions: Works in Progress **1:30 to 3:00pm Location: Decatur A**
1. ***"That's a waste of time, going back and reading the text again!" – Cognitive processes in an integrated summary writing task*** Sonja Zimmermann
2. ***Studying item difficulty. Insights from a multilingual foreign language assessment*** Katharina Karges
3. ***Development of Scales for the Assessment of Young Learners Functional Writing Proficiency*** Gustaf Bernhard Uno Skar, Lennart Joelle
4. ***Investigating the Interactiveness of IELTS Academic Writing Tasks and Their Washback on EFL Teachers' Test Preparation Practices*** Parisa Safaei, Shahrzad Saif
5. ***The methods dealing with dependent effect sizes in a meta-analysis: a review in reading research area*** Jingxuan Liu, Xiaoyun Zhang, Hongli Li, Xinyuan Yang
6. ***Towards the Democratisation of the Assessment of English as a Lingua Franca*** Sheryl Cooke
7. ***Understanding Young Learners' Spoken Academic Language Development through Analyzing Oral Proficiency Test Responses*** Megan Montee, Mark Chapman
8. ***Validating the use of a web-based rating system for oral proficiency interviews*** Jing Xu, Anne Clarke, Andrew Mullooly, Claire McCauley
9. ***Test preparation materials for the Test of Workplace Essential Skills (TOWES): Validating materials for adult literacy and numeracy*** Claire Elizabeth Reynolds
10. ***Students' and Teachers' Perception and Use of Diagnostic Feedback*** Hang Sun
11. ***Effects of Reader and Task Variables in L2 Reading Comprehension and Speed*** Toshihiko Shiotsu
12. ***The Effect of Genre on Linguistic Features of Source-Based Essays by Tertiary Learners: Implications for Construct Validity*** Sukran Saygi, Zeynep Aksit
13. ***Investigation of Social Justice Violation in an English Proficiency Test for PhD Candidates in Iran*** Masood Siyyari, Negar Siyari

Wednesday, March 6

Concurrent Sessions: Works in Progress **1:30 to 3:00pm Location: Mary Gay C**
14. ***Developing a Digital Simulation to Measure L2 Intercultural, Pragmatic and Interactional Competence: Initial Pilot Results*** Linda Forrest, Ayşenur Sağdıç, Julie Sykes, Margaret E. Malone
15. ***Using unscripted spoken texts in college-level L2 Mandarin assessment*** Xian Li
16. ***LAL: what is it to teachers in their classrooms?*** Sonia Patricia Hernandez-Ocampo
17. ***A case study of evidence-centered exam design and listening: Washback from placement test development to program tests, KSAs, and pedagogy*** Gerriet Janssen, Olga Inés Gómez
18. ***Development of an ITA Assessment Instrument based on English as a Lingua Franca*** Heesun Chang
19. ***Applying a summative assessment speaking test for formative assessment gains: The case of a computerized speaking test in Israel*** Tziona Levi, Ofra Inbar-Lourie
20. ***Creating a Socially Responsible Language Diagnostic Tool to Support At-Risk Students at a Canadian Technical College*** Nathan J Devos
21. ***Examining values and potential consequences in argument-based approaches to TOPIK-Speaking validation process*** Soohyeon Park, Gwan-Hyeok Im, Dongil Shin
22. ***Aviation English Proficiency Test Design for a Group of Brazilian Military Pilots: A Case Study*** Ana Lígia Barbosa de Carvalho e Silva

Time	Event
3:00 to 3:20pm	**Coffee Break** Location: Lobby
3:20 to 5:20pm	**Symposium:** ***Language Proficiency Assessment and Social Justice in the US K-12 Educational Context*** Mark Chapman, Margo Gottlieb, Keira Ballantyne, H. Gary Cook, Paula Winke, Todd Ruecker, Micheline Chalhoub-Deville Location: Decatur B

Wednesday, March 6

Concurrent Sessions: Papers			
Time	**Mary Gay C**	**Henry Oliver**	**Decatur A**
3:20 to 3:50pm	***Investigating the validity of a writing scale and rubric using a corpus-based analysis of grammatical features*** Susie Kim	***Assessing Workplace Listening Comprehension of Thai Undergraduates in English as an Asian Lingua Franca Contexts*** Panjanit Chaipuapae	***Justifying the Use of Scenario-Based Assessment to Measure Complex Constructs of Communicative Language Competence*** Heidi Liu Banerjee
3:55 to 4:25pm	***French Learners' Use of Sources in an Integrated Writing Assessment Task*** Anna Mikhaylova	***Item-level analyses of a listening for implicature test: Evidence against an implicature subskill construct?*** Stephen O'Connell	***Differential Item Functioning in GEPT-Kids Listening*** Linyu Liao
4:30 to 5:00pm	***Formative Assessment through Automated Corrective Feedback in Second Language Writing: A Case Study of Criterion*** Giang Thi Linh Hoang	***Content-rich videos in academic L2 listening tests: A validity study*** Roman Olegovich Lesnov	***Raters' Perceptions and Operationalization of (In)Authenticity in Oral Proficiency Tests*** John Dylan Burton
5:05 to 5:35pm	***Individualized Feedback to Raters: Effects on Rating Severity, Inconsistency, and Bias in the Context of Chinese as a Second Language Writing Assessment*** Jing Huang, Gaowei Chen	***Examining the effects of foreign-accented lectures on an academic listening test at the item level using differential item functioning analysis*** Sun-Young Shin, Ryan Lidster, Senyung Lee	***Investigating variance sources and score dependability of an ITA speaking test for construct-related validity and fairness: A mixed method G-theory study*** Ji-young Shin
6:45pm	**Networking Dinners (advance registration required)** Meet in Lobby		

Thursday, March 7

Time	Event		
7:15am to 5:00pm	**Conference registration open**		
Concurrent Sessions: Demonstrations			
Time	**Mary Gay C**	**Henry Oliver**	**Decatur A**
8:00 to 8:30am	***Integrating Social Justice and Student Performance Online*** Noah McLaughlin	***Development and Use of an Automatic Scoring Model for Spoken Response Test*** Judson Hart, Troy Cox, Matthew Wilcox	***BEST Plus 3.0: Assessing Speaking Using a Multi-Stage Adaptive Test*** Megan Montee Daniel Lee
Concurrent Sessions: Symposia			
Time	**Decatur A**	**Decatur B**	
8:35 to 10:35am	***Toward Social Justice in L2 Classroom Assessment Theory and Practice: The Potential of Praxis*** Matthew E. Poehner, Ofra Inbar-Lourie, Constant Leung, Tziona Levi, Luke Harding, Tineke Brunfaut, Remi van Compernolle, Angela Scarino	***Aligning language tests to external proficiency scales: validity issues*** Sha Wu, Lianzhen He, Jianda Liu, Han Yu, Richard J. Tannenbaum, Spiros Papageorgiou, Ching-Ni Hsieh, Shangchao Min, Hongwen Cai, Jie Zhang, Jamie Dunlea, Richard Spiby	
10:35 to 10:55am	**Coffee Break** Lobby		
11:00am to 12:00pm	**Messick Lecture** ***Testing and the Public Interest*** Dr. Joan Herman, UCLA/CRESST *Sponsored by Educational Testing Service* Location: Decatur B		
12:00 to 1:45pm	**Lunch Provided (first 100 takers)** **ILTA Annual Business Meeting** Location: Decatur B		

Thursday, March 7

Concurrent Sessions: Posters **1:30 to 3:00pm Location: Lobby**
1. ***Do students' motivation and locus of control impact writing performance through their perceived writing competency?*** Clarissa Lau, Chris Barron
2. ***Deconstructing writing and a writing scale: How a decision tree guides raters through a holistic, profile-based rating scale*** Hyunji Park, Xun Yan
3. ***Assessing EFL college students' speaking performance through Google Hangouts*** Yu-Ting Kao
4. ***Is it fair to use scores from a test of grammar and vocabulary to refine grade boundary decisions in other skill areas?*** Karen Dunn, Gareth McCray
5. ***Test taker characteristics as predictors of holistic score on independent and integrated-skills writing tasks*** Analynn Bustamante, Scott Crossley
6. ***An Investigation of the Validity of a New Speaking Assessment for Adolescent EFL Learners*** Becky Huang, Alison Bailey, Shawn Chang, Yangting Wang
7. ***Instructors as Agents of Change: A Systematic Approach to Developing Proficiency-Oriented Assessments in Less Commonly Taught Languages*** Shinhye Lee, Ahmet Dursun, Nicholas Swinehart
8. ***Multimodality, Social Semiotics, and Literacy: How LESLLA Learners from Refugee Backgrounds Make Meaning in Official U.S. Naturalization Test Study Materials*** Jenna Ann Altherr Flores
9. ***Raters' decision-making processes in an integrated writing test: An eye-tracking study*** Phuong Nguyen
10. ***Story of an education system accountable for exam-success but not for learning: A washback study*** Nasreen Sultana
11. ***Using multimodal tasks to promote more equitable assessment of English learners in the content areas*** Scott Grapin, Lorena Llosa
12. ***Familiarizing standard-setting panelists with the CEFR: A three-step approach to attaining a shared understanding of just-qualified candidates*** Sharon Pearce, Patrick McLain, Tony Clark
13. ***Using Machine Learning Techniques in Building and Evaluating Automated Scoring Models for ITAs' Speaking Performances*** Ziwei Zhou
14. ***A systematic review: Ensuring high quality ELP assessments for all*** Jo-Kate Collier
15. ***Bridge to Seven (Language Testing and Social Justice)*** Johanna Motteram

Thursday, March 7

Concurrent Sessions: Posters **1:30 to 3:00pm Location: Lobby**
16. ***Beyond the Test Score: Developing Listening Test Feedback & Activities to Empower Young Learners and Teachers of English*** Brent Miller, Luke Slisz, Patrick McLain, Rachele Stucker, Renee Saulter
17. ***Cyberpragmatics: Assessing Interlanguage Pragmatics through Interactive Email Communication*** Iftikhar Haider
18. ***Reverse-engineering L2 reading and listening assessments for sub-score-reporting purposes*** Yeonsuk Cho, Chris Hamill
19. ***Scenario-based tasks for a large-scale foreign language assessment: a mixed-methods exploratory study*** Malgorzata Barras, Katharina Karges, Peter Lenz
20. ***Developing a local-made English test for Thai EFL grade 6 students: Concurrent validity and fairness issues*** Jirada Wudthayagorn, Chatraporn Piamsai, Pan-gnam Chairaksak
21. ***Holistics and analytic scales of a paired oral test for Japanese learners of English*** Rie Koizumi, Yo In'nami, Makoto Fukazawa
22. ***Accessibility in testing: generating research from good practice*** Richard David Spiby, Judith Fairbairn
23. ***Listening to test-takers' perspective in the validation process: the case of the Aviation English Proficiency Exam for Brazilian Air Traffic Controllers*** Natalia de Andrade Raymundo
24. ***Pre-service and in-service language teachers' conceptions of LA: towards the construction of LAL knowledge base*** Sonia Patricia Hernandez-Ocampo
25. ***Language assessment literacy in Brazil: analyses of undergraduate and graduate courses at federal universities*** Gladys Quevedo-Camargo, Matilde V. R. Scaramucci
26. ***Impact of Language Background on Response Similarity Analysis*** James Robert Davis
27. ***Japanese EFL Learners' Speech-in-Noise Listening Comprehension Process: Use of Context Information*** Ryoko Fujita
28. ***Certifying language ability for immigration purposes in Switzerland*** Peter Lenz
29. ***Comparing rater and score reliability under holistic and analytic rating scales in assessing speech acts in L2 Chinese*** Shuai Li
30. ***Exploring raters' perceptions of Oral Proficiency Interview Tasks as "promotion" or "demotion"*** Jeremy Ray Gevara, Troy Cox, Larissa Grahl, Logan Blackwell
31. ***Mapping the Path to Advanced Second Language Literacy in Adults Using Eye-Tracking: A Look at Portuguese*** Troy Cox, Larissa Grahl, Logan Blackwell

Thursday, March 7

Time	Event			
3:00 to 3:20pm	**Coffee Break** Location: Lobby			
Concurrent Sessions: Papers				
Time	**Mary Gay C**	**Henry Oliver**	**Decatur A**	**Decatur B**
3:20 to 3:50pm	***Analyzing stakeholders' voices in the aviation context: a glocal perspective*** Natalia de Andrade Raymundo	***Positioning students as active learners: An examination of student-generated question quality in literacy assessment*** Hyunah Kim, Megan Vincett, Samantha Dawn McCormick, Melissa Hunte, Xue Lin	***Enhancing the Interpretability and Usefulness of TEPS Section Scores Through Alignment with CEFR*** Heesung Jun, Euijin Lim, Yong-Won Lee	***Assessing textual sophistication and linguistic complexity in L2 writing*** Jianling Liao
3:55 to 4:25pm	***The domain expert perspective on workplace readiness: Investigating the standards set on the writing component of an English language proficiency test for health professionals*** Simon Davidson	***Do test accessibility features have the intended effect for K-12 English learners?*** Ahyoung Alicia Kim, Meltem Yumsek, Mark Chapman, H. Gary Cook	***High-stakes tests can improve learning - Reality or wishful thinking?*** Jessica Wu, Judy Lo, Anita Chun-Wen Lin	***Linguistic Tools in Writing Assessment: Their Impact on Test-takers' Writing Process and Performance*** Saerhim Oh

Thursday, March 7

Concurrent Sessions: Papers				
Time	**Mary Gay C**	**Henry Oliver**	**Decatur A**	**Decatur B**
4:30 to 5:00pm	***How valid are language tests used in the overseas-trained nurse registration processes?*** Ute Knoch, Sally O'Hagan	***Empowering K-12 Teachers to Make Better Use of High-Stakes Summative ELP Assessments*** Alexis Lopez	***Source Use Behavior and Raters' Judgement in L2 Academic Writing*** Pakize Uludag, Heike Neumann, Kim McDonough	***Unpacking the textual features, vocabulary use, and source integration in integrated listening-to-write assessments for adolescent English language learners*** Renka Ohta, Jui-Teng Liao
5:05 to 5:35pm	***Assessing clinical communication on the Occupational English Test: The intersection of cognitive and consequential validity*** Brigita Séguis, Barbara Ying Zhang, Gad Lim	***Strategies Used by Young English Learners in an Assessment Context*** Lin Gu, Youngsoon So	***Writing Assessment Training Impact and Mexican EFL University Teachers: A Proposed Categorization*** Elsa Fernanda Gonzalez	***Japanese university students' paraphrasing strategies in L2 summary writing*** Yasuyo Sawaki, Yutaka Ishii, Hiroaki Yamada
6:30 to 9:30pm	**Banquet (Ticket required)** The Trolley Barn 963 Edgewood Ave NE, Atlanta, Georgia			

Friday, March 8

Time	Event		
7:15am to 1:00pm	**Conference registration open**		
7:30 to 8:30am	**Language Assessment Literacy Special Interest Group** Location: Henry Oliver		
	Concurrent Sessions: Papers		
Time	**Mary Gay C**	**Henry Oliver**	**Decatur A**
8:30 to 9:00am	***Investigating raters' scoring processes and strategies in paired speaking assessment*** Soo Jung Youn, Shi Chen	***Multilingual Assessment Reflecting Multilingual Educational Policy: Toward Assessment for Justice*** Elana Goldeberg Shohamy, Michal Tannenbaum, Anna Gani	***Developing lists of empirical English word difficulties specific to each L1*** Steve Lattanzio, Alistair Van Moere, Jeff Elmore
9:05 to 9:35am	***Rater behavior in a high-stakes L2 examination: Does test takers' perceived first language matter?*** Ari Huhta, Sari Ohranen, Mia Halonen, Tuija Hirvelä, Reeta Neittaanmäki, Sari Ahola, Riikka Ullakonoja	***Social justice and washback in language testing in Norway*** Marte Monsen	***Exploring the Impact of Bilingual Education Types on DIF: Implications for Vocabulary Test Development*** Suchada Sanonguthai
9:40 to 10:10am	***Not Unwarranted Concordances But Warranted Convergences: Approaches to Standard Setting and Maintenance Using Subject Experts*** Gad Lim, Barbara Ying Zhang, Brigita Seguis	***Intended and unintended consequences of reforming a national school-leaving exam and their role for validation*** Benjamin Kremmel, Carol Spoettl, Veronika Schwarz	***A Knowledge-based Vocabulary List (KVL): German, Spanish, and Chinese Results*** Norbert Schmitt, Barry O'Sullivan, Laurence Anthony, Karen Dunn, Benjamin Kremmel

Friday, March 8

Time	Event			
8:30 to 10:30am	**Symposium:** ***Transformative teacher-researcher partnerships in language assessment*** Beverly Baker, José Manuel Martínez, Ni-La Lê, Erika B. Kraus, Azad Hassan, India C. Plough, Xun Yan, Ha Ram (Hannah) Kim, John Kotnarowski, Hyunji (Hayley) Park, Jamie L. Schissel, Mario López-Gopar, Constant Leung, Julio Morales, James R. Davis Location: Decatur B			
10:35 to 11:00am	**Coffee Break and Group Photo** Lobby			
Concurrent Sessions: Papers				
Time	**Mary Gay C**	**Henry Oliver**	**Decatur A**	**Decatur B**
11:00 to 11:30am	***Understanding Writing Process of Adult EFL Learners in a Writing Assessment Context*** Ikkyu Choi	***How do raters learn to rate? Many-facet Rasch modeling of rater performance over the course of a rater certification program*** Xun Yan, Hyunji Park	***Language assessment and student performance in South African higher education: The case of Stellenbosch University*** Kabelo Wilson Sebolai	***The Impact of an External Standardized Test on Teaching and Learning for Young Learners: A Year 1 Baseline Study in Turkey*** Mikyung Kim Wolf, Alexis Lopez, Jeremy Lee
11:35am to 12:05pm	***What aspects of speech contribute to the perceived intelligibility of L2 speakers?*** Willam Bonk, Saerhim Oh	***Establishing a Validity Argument for a Rating Scale Developed for Ongoing Diagnostic Assessment in an EFL University Writing Classroom: A Mixed Methods Study*** Apichat Khamboonruang	***Exploring teacher understandings and beliefs as a basis for benchmarking assessments for university foreign language programs*** Noriko Iwashita	***Investigating the consequential validity of the Hanyu Shuiping Kaoshi (Chinese proficiency test) by using an Argument-based framework*** Shujiao Wang

Friday, March 8

Time	Event			
12:05 to 1:35pm	**Lunch Break on your own** *Language Assessment Quarterly* Editorial Board Meeting, Location: Henry Oliver			
	Concurrent Sessions: Papers			
Time	**Mary Gay C**	**Henry Oliver**	**Decatur A**	**Decatur B**
1:35 to 2:05pm	***Examination of test-taking strategies used for two item types during L2 listening assessment*** Ruslan Suvorov	***Academic language or disciplinary practices? Reconciling perspectives of language and content educators when assessing English learners' language proficiency in the content classroom*** Lorena Llosa, Scott Grapin	***Placement Testing: One test, two tests, three tests? How many tests are sufficient?*** Kathryn Hille, Yeonsuk Cho	***Developmental frameworks for writing in Denmark, Norway, and the US: A Cross-national comparison*** Jill V. Jeffery, Nikolaj Elf, Gustaf Bernhard Uno Skar, Kristen Campbell Wilcox
2:10 to 2:40pm	***Exploring the relationships between test value, motivation, anxiety and test performance: The case of a high-stakes English proficiency test*** Jason Fan, Yan Jin	***The role of feedback in the design of a testing model for social justice*** Slobodanka Dimova	***Mitigating rater bias in L2 English speaking assessment through controlled pairwise comparisons*** Masato Hagiwara, Burr Settles, Angela DiCostanzo, Cynthia M. Berger	***Examining the Structure, Scale, and Instructor Perceptions of the ACTFL Can-Do Statements for Spoken Proficiency*** Sonia Magdalena Tigchelaar

Friday, March 8

Concurrent Sessions: Papers				
Time	**Mary Gay C**	**Henry Oliver**	**Decatur A**	**Decatur B**
2:45 to 3:15pm	***Establishing appropriate cut-scores of standardized tests for a local placement context***	***Towards social justice for item writers: Empowering item writers through language assessment literacy training***	***Use of automated scoring technology to predict difficult-to-score speaking responses***	***Building a Partial Validity Argument for the Global Test of English Communication***
	Gary J. Ockey, Sonca Vo, Shireen Baghestani	Olena Rossi, Tineke Brunfaut	Larry Davis, Edward Wolfe	Payman Vafaee, Yuko Kashimada

Time	Event
3:15 to 3:35pm	**Coffee Break** Lobby
3:35 to 5:00pm	**Distinguished Achievement Award Lecture** ***Context, Language Knowledge, and Language Use: Current Understandings*** Dan Douglas, Iowa State University *Sponsored by Cambridge ESOL/ILTA* Location: Decatur B
5:00 to 5:30pm	**Wrap up & Thanks** Location: Decatur B

Saturday, March 9

Time	Event
1:30 to 3:30pm	**Joint AAAL/ILTA Invited Colloquium** ***Assessing lingua franca competence*** Location: Atlanta Sheraton, Capitol North Ballroom

THE CAMBRIDGE / ILTA DISTINGUISHED ACHIEVEMENT AWARD

Sponsored by Cambridge ESOL/ILTA

Context, Language Knowledge, and Language Use: Current Understandings

Dan Douglas, Iowa State University

Friday, March 8
3:35 p.m. to 5:00 p.m. Location: Decatur B

In this presentation I will discuss specific purpose test performance from the point of view of the interplay between language and context by considering some current studies, focusing on academic, aviation, and medical discourse. The argument is that until we better understand and define the nature of language knowledge in specific contexts of use, we risk missing the target of our assessments: the measurement of specific purpose language ability. Recalling Weir's (2005) notion of context validity, I contend that we do in fact have a better understanding of these matters now than we did back in 2000 when I proposed a broader definition of specific purpose language ability than was common at the time. However, we still have far to go, as I will try to demonstrate in my review of some current studies.

Weir, C. 2005. Limitations of the Common European Framework for developing comparable examinations and tests. Language Testing 22.3: 281-300.

Award Announcement

Dan Douglas completed his PhD in Applied Linguistics at Edinburgh University in Scotland in 1977. His dissertation was entitled: A Study of Reading Among Secondary School Pupils in a Developing Nation. He went on to work at several universities including the English Language Institute at the University of Michigan until 1985 when he began as Assistant Professor in the English Department at Iowa State University. He remained there and in 1990 was promoted to Associate Professor with tenure, in 1997 promoted to Professor, and is presently Professor Emeritus after retiring in 2009. Besides his national and international work in language testing, he served his own institution in several ways: teaching a variety of courses from composition for foreign graduate students to testing language for specific purposes; supervising not only local graduate students, but also generously working with visiting scholars; serving as coordinator of the TESL/Applied Linguistics area, and volunteering for a variety of other committees.

Since his retirement Dan Douglas continues to be active and among other work, conducts research on the assessment of language ability in specific academic and professional contexts, the effect of context on second language acquisition and use, language for specific purposes and assessing language through computer technology.

Dan has had diverse research interests in the field of language testing, but one of his most valuable academic contributions is his extensive research on the assessment of language ability in academic and professional contexts. This is reflected in several of his numerous peer-reviewed publications including books, book chapters and journal articles. His book entitled *Assessing Languages for Specific Purposes*

(Douglas, 2000) has had a major influence. This important work stimulated language testers' thinking on the interplay between language constructs and contexts of language use and provided detailed guidelines for constructing contextualized language tests for specific purposes. Some examples of the impact of this work can be seen in Weir's (2005) notion of "context validity" in his socio-cognitive validation framework; and in the theoretical basis for tests such as Cambridge Assessment's revision of the Business English Certificates (BEC), the Occupational English Test (OET) for health professionals (Hoekje, 2015), and in aviation English tests for pilots and air traffic controllers (Kim & Elder, 2015).

The influence of Dan's expertise can also be seen in his extensive professional service to the language testing community. He served as President of ILTA at two different times (2005 and 2013-2014), with connecting roles of Vice-President and Past President. He also was Member and Chair of the ILTA Awards Committee (1994-1997). Other service can be seen in his positions as Chair and Member of the TOEFL Award Committee for Outstanding Doctoral Dissertation, Secretary of MwALT, Co-Editor of Language Testing with John Read, Co-Chair with Micheline Chalhoub-Deville, Craig Deville and Felicity Douglas of LTRC St. Louis 2001, Chair of the 2005 MwALT conference and Member of the TOEFL Grants and Awards Committee. In addition, Dan has served on the Editorial Board of several academic journals for over a decade, including Language Testing, English for Specific Purposes, Journal of English for Academic Purposes, and ESP Malaysia. In 2015, Dan, together with his wife Felicity, conducted a two-day ILTA workshop on Classroom Language Assessment in Ghana and paved the way for creating the Ghana Association of Language Testing, the first ILTA-affiliated organization in Africa.

Another aspect of Dan's service was his active engagement in the development of operational language tests. He was a consultant for the TOEFL 2000 Development Project, Chair of the Test of Spoken English Committee, and Member of the TOEFL Committee of Examiners at the Educational Testing Service. He also played a leading role in revising the English Language Placement Test at Iowa State University (2006) and was Principal Investigator for the Pearson Education Online Business Test Development Project (2001).

Finally, in the words of those who put forth Dan's nomination, he has been "a kind and inspiring mentor who introduced many young scholars into the field of language testing." Some of these young scholars are now working as language testing professionals, while others have become academics.

Given the extent and impact of Dan Douglas's contributions to the field of language testing, and his scholarship, mentorship and diverse service, the Award Committee has reached a unanimous decision in granting him the 2019 Distinguished Achievement Award.

Carolyn Turner (Chair), Micheline Chalhoub-Deville, Yong-Won Lee and Nick Saville

DAA Award Committee

ALAN DAVIES LECTURE

Sponsored by the British Council

Bringing tests to justice: can we make a difference?

Cathie Elder, University of Melbourne

Tuesday, March 5
5: 15 p.m. to 6:45 p.m. Location: Decatur B

The conference theme of social justice, when applied to language assessment, invites consideration of issues of validity, fairness, ethics, social consequences, accountability, responsibility and professionalism – all addressed in the writings of Alan Davies (1931-2015). The paper begins by briefly sketching Alan's contribution in these areas and considering his influence and the relevance of his ideas to current discussions about how, or to what extent, language testers can move towards more socially responsible practice. In what ways (if at all) has our thinking evolved since Alan's time? How useful are existing ethical codes and validation frameworks in guiding and regulating our practices? Is a focus on validity and validation and the creation of a 'professional milieu' within our sphere of expertise the best we can do to promote socially just outcomes? What strategies might individual language testers or professional organizations like ILTA pursue in the interests of achieving better communication with stakeholders and more equitable uses of tests in their policy contexts?

Such questions are explored with reference to particular language assessment scenarios I and my colleagues have encountered, where language tests have been misunderstood, misrepresented or misused with potential or actual adverse consequences. I argue, with the benefit of hindsight, for greater attention by language testers to stakeholder understandings of language, language test scores and the reasonings underlying test use and better training for our profession in the tools for effective and persuasive engagement in the policy domain.

Discussants will respond to the issues raised and group discussion will follow.

Dr. Cathie Elder is Principal Fellow in the School of Languages and Linguistics at the University of Melbourne and a former Director (2007-12) of the Language Testing Research Centre (of which Alan Davies was the founding Director). In the latter role she led a number of major research and development projects on LSP testing and on school- and industry-based language standards/assessment frameworks. She also held senior academic positions at Monash University (2004-6) and at the University of Auckland (2000-2004), where she set up the university-wide Diagnostic English Language Needs Assessment (DELA) program.

Dr. Elder served as Chair of the TOEFL Committee of Examiners at the Educational Testing Service in Princeton from 2006 to 2008 and as co-editor of Language Testing from 2007 to 2011. She was a founding member and co-president (2013) of the Australian Association for Language Testing and Assessment of Australia and New Zealand (ALTAANZ) and President of the International Language Testing Association (ILTA) (2017-2018).

Dr. Elder has been teaching and researching in the areas of language testing assessment and program evaluation for nearly 30 years, publishing widely in peer-reviewed journals and edited volumes and contributing to major reference works in the field. She collaborated with Alan Davies on the *Dictionary of Language Testing* (CUP 1999) and on the *Handbook of Applied Linguistics* (Blackwell 2004) and with her colleagues at the Language Testing Research Centre produced a festschrift in Alan's honour entitled *Experimenting with Uncertainty* (CUP 2001). Her PhD research, completed in 1997, considered questions of fairness in the use of common school-based examinations for language learners from heritage and non-heritage language backgrounds. She has an abiding interest in the thorny issue of construct definition in language assessment and how language proficiency is construed by various stakeholders in the assessment enterprise, including teachers, university admissions officers, academics from different disciplines, health professionals, aviation personnel, lawyers and politicians.

THE SAMUEL J. MESSICK MEMORIAL LECTURE

Sponsored by Educational Testing Service

Testing and the Public Interest

Dr. Joan Herman, UCLA Center for Research on Evaluation, Standards, and Student Testing (CRESST)

Thursday, March 7
11:00 a.m. to 12:00 p.m. Location: Decatur B

How well is testing serving the public interest? How might its productive impact be strengthened? These are central questions for the presentation. It starts by considering definitions of public interest, a complex concept that embeds social values and whose interpretation varies depending on where one is in the body politic. With a focus on high stakes testing and the extent to which it benefits (or not) teaching and learning, I consider available evidence from the United States and internationally on impacts on curriculum and instruction, student outcomes, and fairness. A very recent study using PISA data from 58 countries (Bergbauer, Hanushek & Woessmann, 2018) is emblematic in finding that accountability testing positively effects student performance, with larger effects for school-based rather than individual student-based accountability and larger effects for school systems that are low performing.

Admittedly, however these effects tend to be relatively small. I hypothesize that one reason for the limited effect is the testing community's inattention to use and to the changes in practice that are necessary for serious improvement in learning. Logic-based theories of action too often assume test information will magically promote effective action, yet the use of assessment to promote meaningful learning depends on knowledge and capacity, repertoires of alternative action, and motivation to change, any and all of which may be in short supply. Synthesizing evidence from the literatures of knowledge utilization, assessment use, data-based decision-making, and socio-cultural theories of organizational and individual learning, I make the case for greater attention to the different types of uses that may support improvement and to the social context of use and draw implications for improving practice.

Joan Herman is Co-Director Emeritus of the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) at UCLA, where she currently serves as senior research scientist. Her research has explored the effects of testing on schools and the design of assessment systems to support school planning and instructional improvement. Her recent work focuses on the validity and utility of teachers' formative assessment practices and the assessment of deeper learning. She also has wide experience as an evaluator of school reform.

Dr. Herman is noted in bridging research and practice. Among her books are *Turnaround Toolkit*; and *A Practical Guide to Alternative Assessment*, both of which have been popular resources for schools across the country. A former teacher and school board member, Dr. Herman also has published extensively in research journals and is a frequent speaker to policy audiences on evaluation and assessment topics. She is past president of the California Educational Research Association; has held a variety of leadership positions in the American Educational Research Association, National Organization of Research Centers, and Knowledge Alliance; and is a frequent contributor at the National Academies' National Research Council (NRC). An elected member of the National Academy of Education and a fellow of the American Educational Research Association, Dr. Herman is current editor of Educational Assessment, served on the Joint Committee for the Revision of Standards for Educational and Psychological Testing, and is a member NRC's Board on Testing and Assessment. She received her BA in Sociology from the University of California, Berkeley, was awarded an MA and Ed.D in Learning and Instruction from the University of California, Los Angeles and is a member of Phi Beta Kappa.

PRE-CONFERENCE WORKSHOPS

Workshop 1: R & RStudio for Reproducible Language Test Analysis, Research, and Reporting

Monday, March 4, 9:00 a.m. - 4:00 p.m.
Tuesday, March 5, 9:00 a.m. - 4:00 p.m.

Location: Mary Gay C

Geoffrey T. LaFlair, Duolingo**, Daniel R. Isbell**, Michigan State University

When R and RStudio are used together (with other tools internal and external to the RStudio environment) they provide an excellent toolkit for carrying out analyses relevant to research in the field of language assessment as well as the practice of evaluating language assessments. Additionally, this toolkit allows for data processing and analysis that is reproducible, helping to ensure quality control in practical settings and transparency in research. Combining basic data and statistical functions with packages specifically for working with test data, we will provide attendees with a foundation in using R for language testing.

Workshop Goals:

- Participants will attain basic working familiarity with R and RStudio, including how to read in,
- subset/filter, summarize, and produce basic visualizations of data.
- Participants will be able to use R to carry out common test data processing and reporting tasks,
- including reliability analyses, item analyses, and score reporting.
- Participants will be able to use R for analyzing test data for validation/research purposes, including
- test equating, Rasch/IRT and CFA analyses.

Geoff LaFlair is an Assessment Researcher at Duolingo where he works on the research and development of the Duolingo English Test. Prior to joining Duolingo, he was an assistant professor in the Department of Second Language Studies at the University of Hawai'i Mānoa. He is the author and maintainer of the rcrtan R package, an R package for criterion-referenced test analysis. He is also a lesson maintainer and a curriculum advisory committee member for Data Carpentry's social sciences lessons. His research has appeared in *Language Testing*, *Applied Linguistics*, and *The Modern Language Journal*.

Daniel R. Isbell is a PhD candidate in Second Language Studies at Michigan State University, where he has successfully introduced students and faculty to R through workshops and tutorials. Dan's operational testing experience includes the Program in Intensive English at Northern Arizona University, the Testing Office at MSU's English Language Center, and an internship at Michigan Language Assessment. His assessment interests involve rater effects, test fairness, and diagnostic language assessment. Dan's other research interests lie in L2 pronunciation and instructed SLA. Dan's work has been published in *Language Testing*, *Studies in Second Language Acquisition*, and *Language Learning & Technology*.

Workshop 2: Constructing Measures: An Introduction to Analyzing Proficiency Test Data Using IRT with jMetrik

Monday, March 4, 9:00 a.m. - 4:00 p.m. Location: Henry Oliver

Troy L. Cox, Center for Language Studies

jMetrik was developed as an open source computer program for conducting psychometric analysis including classical item analysis, differential item functioning, and item response theory. Since it is free, it is an ideal tool for language assessment professionals to learn as they can take it with them as they progress throughout their career.

Goals:
Use jMetrik to:

- describe a test-taking population,
- score tests with dichotomous and polytomous data,
- calculate descriptive statistics and interpret the results,
- graph test results and visually inspect test data,
- calculate item statistics (both classical and IRT),
- calculate the reliability of a test instrument, and
- conduct a Rasch IRT analysis.

Troy L. Cox, PhD, is a faculty member in the Linguistics Department at Brigham Young University and serves as the Associate Director of the Center for Language Studies. He is also a certified ACTFL oral proficiency tester/trainer and has used his testing expertise as a forensic linguist and in test development projects in a number of different languages. His research interests include language proficiency testing, the integration of technology with assessment, objective measurement and self-assessment.

Workshop 3: Corpus-based development and validation of language tests: using corpora of and for language testing

Tuesday, March 5, 9:00 a.m. - 4:00 p.m. Location: Henry Oliver

Darren Perrett, Cambridge Assessment English, **Brigita Séguis**, Cambridge Assessment English

Corpora are being increasingly used in language testing for a wide range of purposes and in a variety of ways. While there is a general consensus that a corpus can be defined as a collection of written or spoken materials, there is less understanding about the specific properties that are required for a collection of texts to be classified as a corpus, the various functions that corpora can fulfil specifically in the context of language testing and the advantages and limitations of the various tools that are being developed for corpus analysis.

In this workshop, we will explore two major connections between corpus linguistics and language testing, namely using corpora for language testing, as well as creating corpora of language testing. Prerequisites: attendees are requested to bring their own laptops. As part of the workshop, participants will be given access to L1 corpora, as well as temporary access to the full Cambridge Learner Corpus (CLC) via a corpus analysis tool called 'Sketch Engine'.

Goals:

- We will provide participants with the necessary skills for accessing and navigating the existing corpora, and demonstrate how they can be used to perform a range of activities relevant to development and validation of language tests, such as item writing and identification of criterial features at different proficiency levels;
- We will also equip participants with the practical tools to create their own corpora by providing an overview of the available corpus interfaces, tagging, as well as understanding the importance of high-quality metadata.

Darren Perrett holds a Master's degree (MA) from Lancaster University and is currently studying for his PhD in Education from Leeds University. He is an Assessment Manager at Cambridge Assessment English, a non-teaching department of the University of Cambridge. He is responsible for content production and process improvements using automation and machine learning techniques. His main research interests include using corpus tools to inform item production and identifying criterial features of the Cambridge Assessment English item bank.

Brigita Séguis is a Senior Research Manager at the Research and Thought Leadership Department of Cambridge Assessment English. Her current areas of research/interest include the Occupational English Test (OET), corpus linguistics, multilingualism and impact studies. She holds a DPhil in General Linguistics and Comparative Philology from the University of Oxford.

Workshop 4: The Civil and Human Rights of Language: Guided Tour of the Atlanta Civil & Human Rights Museum

Tuesday, March 5, 9:00 a.m. - 1:00 p.m.

Elana Shohamy, Tel Aviv University, **Tim McNamara**, University of Melbourne

Come with workshop leaders to the Center for Civil and Human Rights in Atlanta, Georgia for a guided tour of the building, exhibits and displays, with a focus on how language is an integral part of civil and human rights.

Workshop Agenda:

9:00 AM: Meet in the conference hotel lobby at the Registration Desk

9:10 AM: Travel as a group on public transportation to the Center for Civil and Human Rights in downtown Atlanta. (Each person must pay for his or her own public transportation.)

10:00 AM: Tour of the Center, guided by an Educational Specialist on Language at the Center

11:30/Noon: Discussion of the tour and issues related to civil and human rights and language with Drs. Shohamy and McNamara

Lunch: On your own in downtown Atlanta

Elana Shohamy is a professor of Language Education at the School of Education, Tel Aviv University where she researches topics of multilingualism within the contexts of critical framework, language rights and co-existence.

Her work in language testing focuses on the power and misuses of tests in education and society and more recently on multilingual testing. The work in language policy addresses mechanisms used for creating de facto policies that penalize immigrants and minority groups.

In the past decade Elana researches ample dimensions of Linguistic Landscape (LL), referring to languages displayed in public spaces with a focus on LL as arenas of conflicts and contestation. Her books include: *The Languages of Israel: Policy*, *ideology and practice* (w/ B. Spolsky, 1999); *The power of tests: A critical perspective of the uses of language tests* (2001); *Language policy: Hidden agendas and new approaches* (2006); *Linguistic landscape: expanding the scenery*, (ed. w/ Durk Gurter, 2009); and *Linguistic landscape in the city* (ed. w/ Ben Rafael and Barni, 2010). Elana is the editor of the journal *Language Policy* and the winner of the 2010 lifetime achievement award granted by ILTA in Cambridge, UK.

Tim McNamara has worked on major international tests such as IELTS, TOEFL-iBT and the OET over the last 35 years, and was co-founder of the Language Testing Research Centre at The University of Melbourne. His work focuses both on improving test quality to enhance test fairness and on the broader justice issues involved in the policy functions of language assessments, for example in contexts of education, immigration and citizenship. He is the co-author (with Carsten Roever) of *Language Testing: The Social Dimension* (Blackwell, 2006) and (with Ute Knoch and Jason Fan) of *Fairness and Justice in Language Assessment* (OUP, 2019).

Local needs and global priorities in ensuring fair test use: synergies and tension in balancing the two perspectives

Organizer: **Jamie Mark Dunlea**, Assessment Research Group, British Council

Discussant / Moderator: **Barry O'Sullivan,** Assessment Research Group, British Council

8:30 a.m. to 10: 35 a.m. Location: Decatur B

While there is still no consensus on a single model of validation for language testing and assessment, or on a single model of language acquisition which could underpin validation, there is agreement on some key principles. Tests are not validated, rather uses and interpretations are for specified purposes, and the test takers who will be the subject of those decisions should be at the centre of that process.

As with many aspects of our fluid and globalizing societies and economies, language testing is the subject of rapid change, often in superficially contradictory directions: standardized, international models for economies of scale, or locally developed, contextually-dependent testing systems.

But as is often the case, digging into details reveals a more varied picture than such a simple dichotomy represents, and the choice(s) are often better represented as being on a continuum of change rather than a black and white decision of either local or global. Equally, the choice of local and global is not easily explained by economic competition and marketing potential. Locally developed tests can become dominant market players within a particular context, and the same arguments that are played out regarding international, large scale commercial testing systems swamping local diversity can be reflected at the local level within one particular national or regional context. O'Sullivan (2016) expanded the perspective of localization beyond the test to consider localization in terms of the argumentation and approach to reporting of validation evidence that makes sense for local stakeholders.

Localization, then, encompasses a range of dimensions, from test content through test use and interpretation to the way we gather and present evidence to support those uses. This symposium aims to draw on a number of important local test development projects in East Asia to present the various tensions and synergies that test developers in a range of local contexts faced in meeting local needs while also taking account of global priorities in terms of established international best practice. Each of the presenters will speak from the experience of working with a national-level testing program in a particular context while being active members of the international language testing research community. The presenters will dig into their experiences and offer insights, both theoretical and from their hands-on practical experience, of managing these sometimes conflicting, sometimes convergent demands.

The symposium will open the floor for discussion to elicit the experiences and perspectives of our community in dealing with these issues. As language assessment expertise has developed and become established in diverse international contexts, it is particularly pertinent for us to take stock as a community and consider how we facilitate the interaction between local and international perspectives of test design, development and validation going forward. The symposium aims to honor Professor Cyril Weir, who played a fundamental role in mentoring language testing professionals and advising and contributing to their diverse work in local contexts, including the testing programs represented in this symposium.

Paper 1: The glocal approach to monitoring quality in language testing in Asia

Jessica Wu, Language Training and Testing Centre

As a tribute to Prof Cyril Weir, this presentation focuses on the glocal approach, which is also the main theme of a book jointly edited by Prof Weir and Wu (in press). A local test like the GEPT can be viewed as a glocal test based on Dendrino's (2013) views of glocalization, which "involve locally operated schemes, set up to serve domestic social conditions and needs, which are informed by international research and assessment practice".

I will suggest that the glocal approach can be useful for a local test such as the GEPT in establishing and maintaining test quality by presenting selected GEPT validation studies in terms of the socio-cognitive validation framework (Weir, 2005). However, the glocal approach can also be relevant beyond the GEPT. One urgent need is to apply the glocal approach when developing guidelines for language assessment practice in Asia for ongoing quality and fairness. While English language assessment in Asia is still heavily influenced by Western testing theories and international standards, it is imperative to develop guidelines for language assessment practice which are more locally appropriate for the Asian context. Jin and Wu (2008) proposed that future Asian guidelines be glocal by nature, which should reflect the distinctive features of the Asian testing context and at the same time maintain the universal concepts and principles for good language assessment practice stated in the current international guidelines. In this presentation, I will follow up on a method used to achieve this goal.

Paper 2: Features unique to a localised English language test

Yan Jin, Shanghai Jiao Tong University, **Wei Wang**, National Educational Examinations Authority, **Haoran Yang**, Shanghai Jiao Tong University

In 2017, a total of 18.25 million tertiary-level English language learners registered to take the College English Test (CET), a national English language testing system in China. These students were primarily motivated to take the test in order to fulfil the requirements of their institutions: College English is a compulsory course required of students in higher education institutions, and the CET is an exit test to assess the English language proficiency of tertiary-level learners upon their completion of the course. The sheer volume of the test indicates that its impact on individuals, institutions, and the educational and social systems is deep and wide-ranging. In this presentation, we will discuss the features unique to this "large-scale language test with distinctive Chinese characteristics (Zheng & Cheng, 2008: 416). We will first describe the intended purposes of the CET and the way in which the test serves its purposes. We will discuss the banded system of English language teaching and testing and the interface between the CET and College English teaching and learning, focusing on the alignment of assessment criteria and curriculum requirements. We will also analyse the micro-level features of the localised test as reflected in its choices of lexis, language functions, and topics. Finally, as a case study, we will compare the CET speaking test with the speaking tests of two international tests, TOEFL iBT and IELTS. Through the presentation, we hope to demonstrate the value and the challenges of developing and using a contextualized local test to promote teaching and learning.

Paper 3: The politics of glocalization

Quynh Nguyen, University of Languages and International Studies, Vietnam National University

Vietnamese Standardised Test of English Proficiency (VSTEP) is a new-comer, so its developers were able to learn a great deal from the previous work and literature in language testing and assessment. In particular, the experience of other localised English proficiency tests in East Asia provided. perspectives, models, strategies, and options in balancing local needs and maintaining global norms: the do-s and the

don't-s in test glocalisation to ensure both fairness to their local test-takers and international recognition of their test results. However, like any other localised English proficiency test in the region, their choices are driven by and specific to the local political, social and educational, ideological, and economic contexts. This presentation provides an insider's reflection of the rationales for the choices made for VSTEP, explicating the developers' vision of localization, their view of the advantages and disadvantages of adaptation and adoption of the CEFR, their operationalization of localised features in the test, their strategies to promote the test, and their efforts to maintain standardization. I will compare the account of VSTEP with those of some other national-level localised English tests in East Asia in order to show that there exist some common dimensions to the underlying politics that shape the glocalisation of such tests. Some of these dimensions can be found common to most local tests. However, I will highlight those that are specific to tests developed for national goals and used at a national level. I will then call for further empirical research to investigate these recommendations.

Paper 4: Localization as validation

Jamie Mark Dunlea, Assessment Research Group, British Council

All tests constitute a set of compromises in which developers set out to balance the sometimes competing demands of the testing context. This presentation examines the solutions proposed by two testing programs, produced by the same test developer in and for the EFL context of Japan. The two testing programmes provide an interesting case study for the issues associated with developing local, or localized, tests precisely because of the different solutions that the test developers have arrived at in response to different testing purposes within the same overall EFL context. These examples suggest that the concepts of local and global are not black-and-white choices between opposing extremes. These examples also illustrate the important intersection between proficiency levels targeted and the need to "localize" content for specific populations. This has implications for the adaptation of frameworks such as the CEFR for local contexts (whether in or outside Europe). In the CEFR, for example, lower levels such as A1-A2 can be said to be always local, with key words such as everyday, routine, and familiar defining them. From the pivotal B1 level upwards, however, the level descriptions, and intended uses, become more outwardly focused, away from only one local context. The appropriate balance, then, between local and global is not necessarily determined by the geographical context in which a test is developed, but is better seen as deriving from the relationship between the test and its intended uses within that context, a "traditional" validity paradigm.

Language Proficiency Assessment and Social Justice in the US K-12 Educational Context

Organizer: **Mark Chapman**, WIDA at UW-Madison
Moderator: **Margo Gottlieb**, WIDA at UW-Madison
Discussant: **Micheline Chalhoub-Deville**, University of North Carolina at Greensboro

3:20 p.m. to 5:20 p.m. Location: Decatur B

Approximately 5 million students in the US Kindergarten through grade 12 (K-12) educational system are in the process of acquiring English language proficiency to access and engage with challenging classroom content in order to reach academic parity with their peers. Key elements of K-12 federal educational policy in the US mandate that multilingual students be provided with English language instructional programs and, for accountability purposes, take an annual large-scale English language proficiency test. These requirements are enshrined in federal legislation that is supported by Supreme Court case law and litigation (ESEA, 1994, 2001 2015; *Lau v. Nichols*, 1974; *Castañeda v. Pickard,* 1981). These civil rights for multilingual learners have emerged from a long-fought struggle to ensure that all students are provided with equitable access to educational opportunities.

Emerging from these legal requirements is the need for methods to validly and reliably identify the multilingual learners who receive English language support, and to evaluate the point at which these students have reached sufficient levels of language proficiency to be reclassified as no longer in need of these services. Questions remain regarding whether there have been unforeseen consequences from the use of these assessments and the extent to which the promise for educational equity has been fulfilled (Abedi & Linquanti, 2012; Shohamy, 1998).

This symposium examines the historical and social contexts of social justice for multilingual learners and their impact on language proficiency assessment in the US K-12 educational arena. Its goal is to examine ways in which policies and practices in language proficiency assessment have and have not fulfilled their intended purposes of increasing educational equity for multilingual students. In doing so, presenters within the symposium will address the consequences, both intended and unintended of the uses of high-stakes tests for students and educators in the US K-12 assessment context. Questions to be addressed include: To what degree has the implementation of large-scale standardized English language proficiency assessment supported the social justice promise of ensuring increased educational opportunity for multilingual students? And, to what extent have there been unintended consequences from high-stakes testing for these students?

The panel will address the interweaving of specific aspects of social justice with language assessment. The first talk will set the stage with an overview of US legislation and litigation that addresses issues related to assessment of multilingual learners and their implications for social justice. The focus will then turn to examining scores from ACCESS, an English language proficiency test administered to two million English learners annually, in relation to the students' content assessment to determine reclassification decisions, that is, when students are no longer eligible to receive supplemental support. The next presentation will hone in on reading assessment and its consequences for 3rd grade English learners who are developing reading proficiency in English yet are held to the identical state policies as their proficient English peers. The final panelist will discuss the negative impacts of high-stakes assessments and their influence on stakeholders within a specific local context; in particular, bilingual learners and their teachers.

The moderator will deliver a short, framing introduction, based on the overall proposal (above) of no more than five minutes. The moderator will then introduce each paper in turn, with each presenter allowed a maximum of 20 minutes. After all four papers have been delivered, the discussant will present

comments on the papers and a summary of the overall symposium theme (10 minutes). This will leave 15 minutes for Q&A with the audience.

Paper 1: The Socio-Political Contexts of K-12 English Language Proficiency Assessment in the United States: Are We Bending the Arc Toward Justice?

Keira Ballantyne, Center for Applied Linguistics

The United States conducts large-scale standardized testing of the English language proficiency of "English learners" (ESEA, 20 USC § 7801(20)), testing almost five million K-12 students annually (US Department of Education, 2018). This presentation traces the historical and social roots of US K-12 language testing from a social justice perspective. A pastiche of legislative, judicial, and federal funding efforts have led to the current system of required assessments.

Advocates for culturally and linguistically diverse (CLD) learners have employed legislative and legal means to push for enhanced educational equity for students from minority language backgrounds. Cases such as *Lau vs. Nichols* (1974), as well as further legal and legislative decisions in the 70s and 80s, cemented the US socio-political commitment to provide English language supports to English learners.

The legal mandate to identify and provide language instruction to English learners requires a uniform means to identify these children. Additionally, protections are in place against prematurely removing children from instructional programs. Language testing is a primary tool used for these purposes, and tests are state-supported via statutory obligation and funding streams.

In the spirit of Shohamy's (1998) call for a "critical language testing," a key question at this juncture is whether widespread institutionalized language proficiency testing of children supports the social justice aims of ensuring access to educational opportunities. Critiques of the status quo question burdensome testing requirements for children, the prioritizing of English over other languages (e.g. Menken 2008), and the continued delegitimization of minority linguistic practices (Flores & Rosa, 2015).

Paper 2: The use of ACCESS for ELLs scores in determining EL reclassification: relationships between English language proficiency assessment scores and academic content assessment performance

H. Gary Cook, WIDA at UW-Madison

One primary purpose for English language proficiency (ELP) assessments in U.S. public schools is to ascertain whether English learners (EL) have gained sufficient language proficiency to be reclassified as fully English proficient students. Once reclassified, ELs are no longer required to receive English language instruction support. Accordingly, reclassification is a high stakes decision for students. How does one determine that an EL is proficient in academic English? One way that states establish how students meet "academic standards" is through academic content assessments. Hence, understanding the relationship between ELs' English proficiency and academic content performance provides insight into when ELs are or are not English proficient or may or may not be reclassified.

Cook, Linquanti, Chinen, and Jung (2012) examine this relationship empirically and provide a series of statistical procedures that outline how the ELP and content assessment relationship might be understood in the context of EL reclassification. In their paper, Cook, et al provide three types of analyses that could be used to support reclassification decisions: Descriptive box plots, decision consistency analysis, and logistic regression. Subsequent to this paper's publication, many states have used these types of analyses in support decisions on EL reclassification.

This paper will highlight examples of how these types of analyses have been used to support reclassification decisions. It also provides lessons learned in applying these empirical approaches to support decision-making. The aim of this work is to introduce and/or expand empirical methods available to policy makers in making "language proficiency" decisions based on language assessments.

Paper 3: Third-grade reading assessment and retention theories and policies in Michigan and in the USA – consequences for ELL students, families, and educators

Paula Winke, Michigan State University

Lawmakers in 16 states plus DC have enacted laws that require 3rd graders to pass a state 3rd-grade reading test or they cannot automatically progress to 4th grade. The laws are based on a theory that 3rd graders undergo a transition from learning-to-read to reading-to-learn (Chall, 2000). Beginning in 4th grade, reading curricula shift from code-based teaching to content-based teaching. Consequently, if 4th grade children cannot read at grade level, they will be less able to learn the content-based curriculum, which will contribute to academic failure (see Invernizzi & Hayes, 2012, for arguments against this theory upon which the laws are based).

The effects of the laws have been explored in states where they have been in place the longest, such as Florida (Schwerdt, West, & Winters, 2017; Winters, 2012). However, researchers have not investigated how the laws have affected ELLs specifically. Michigan's "Read by Grade 3" begins in 2020 (codified at Mich. Comp. Laws § 380.1280f (2016)). In Michigan, ELLs take two reading tests, the M-STEP English language arts (ELA) test and the WIDA ACCESS for ELLs reading proficiency test.

We show 8,323 ELLs' M-STEP ELA and ACCESS reading test scores from spring 2016. We show the data on a heatmapped-scatterplot and explain the varying, non-linear relationship between the two reading assessments. We discuss what the results mean for Michigan ELL-retention-law interpretations. We propose more contextualized data for understanding ELLs' reading development.

Paper 4: The Impact of High-Stakes Standardized Assessment on Bilingual Learners and their Teachers

Todd Ruecker, University of New Mexico

Policymakers often ignore research on language acquisition and the appropriate use of assessments, which can lead to policies that ultimately harm students, their teachers, and their schools (e.g., Booher-Jennings, 2005; Holloway & Brass, 2017; Hursh, 2007; Wilcox & Lawson, 2018). In recent years, the New Mexico Public Education Department has placed increasing weight on the role standardized assessments play in teacher evaluations. Instead of simply being used as measures of student proficiency, tests like WIDA ACCESS and the Common Core-aligned PARCC are being weaponized in a way that can harm teachers and students. Students are being asked to take advanced tests too early, with students who have been in the U.S. for more than two years being required to take the state tests in English at grade level in all subjects.

The presenter draws on qualitative research at six different school sites. During these visits, he observed ESL, English, and Spanish classes, conducted interviews, and kept detailed notes and an ethnographic journal. The presentation will explore how the oppressive assessment culture created by neoliberal-informed state policies have exacerbated teacher shortages, ultimately hindering schools' abilities to build up local communities of practice in order to better serve bilingual learners. He will also discuss the role large-scale assessments play in upholding an English-only culture in schools. He ends by exploring ways assessment experts can design more inclusive assessments while working to ensure their assessments are used in line with their intended design.

Toward Social Justice in L2 Classroom Assessment Theory and Practice: The Potential of Praxis

Organizers: **Matthew E. Poehner**, The Pennsylvania State University, **Ofra Inbar-Lourie**, University of Tel Aviv
Discussant: **Constant Leung**, Kings College London

8:35 a.m. to 10:35 a.m. Location: Decatur A

Assessment as a regular feature of classroom activity has long been understood as indispensible to monitoring student progress and informing instructional decisions (Genesse & Upshur, 1996; Harlen & James, 1997). In both the general education and L2 fields, a number of proposals have been put forward by researchers that are intended either to document teacher current assessment practices or to guide them toward new ways of approaching assessment. These include, for instance, assessment-for-learning (Black & Wiliam, 2009), dynamic assessment (Poehner, 2013), and learning-oriented assessment (Purpura & Turner, 2014). At the same time, research concerned with language assessment literacy (e.g., Inbar-Lourie, 2017) has helped to conceptualize the knowledge teachers need both to interpret external indicators of learner performance and to design their own assessments (Davison & Hill, 2016) while also increasingly recognizing that assessment needs and goals must be understood relative to local contexts (Scarino, 2016). The aim of this symposium is to consider how the theoretical expertise and concerns of assessment researchers, on the one hand, and the local, practical concerns of classroom teachers, on the other hand, might be understood in relation to one another, and in particular how the notion of *praxis* might offer an orientation to developing both theory and practice together.

Praxis emerges from the tradition of Critical Theory and therefore departs from approaches to research that attempt either to control for potential influences from values and perspectives on the object of study or that alternatively take such individual perspectives and meanings to be the focus of investigation (Delanty & Strydom, 2003). Following Carr & Kemmis (1986), Critical Theory adopts a dialectical method whereby understanding is achieved as part of a process of engagement intended to bring about change. In educational contexts, praxis positions theory and practice in dialectical relation, with theory providing principles and concepts that allow teachers to build their practice in a reasoned, reflective manner that goes beyond firsthand experience while practice serves as a testing ground for theory, pointing to areas in need of revision and expansion. The process involves partnership and trust between researchers and teachers. Moreover, praxis exemplifies the concern in Critical Theory with social justice by emphasizing commitment to *phronesis,* that is, to making better the lives of real people.

The symposium begins with a 10-minute conceptual introduction to praxis as a framework for L2 classroom assessment theory and practice. Twenty minutes are then allotted to each of 4 papers reporting projects undertaken from the vantage of praxis across 4 continents and in different languages and educational settings. Each of the papers evidences the potential of praxis to advance L2 classroom assessment theory and practice. Finally, 10 minutes are given to the discussant to draw attention to areas of overlap among the papers, points of tension, and possibilities for continued development. The remaining twenty minutes are open for general questions and discussion with the audience.

Paper 1: Assessment literacy as praxis: Mediating teacher knowledge of assessment-for-learning practices

Ofra Inbar-Lourie, University of Tel Aviv, Israel, **Tziona Levi**, Ministry of Education, Israel

This paper presents the evolving implementation of a praxis-based language assessment initiative as manifested in school cultures (Deal & Peterson, 1999). Drawing on Vygotsky's perspective of praxis, wherein theory and practice are tied together and mutually substantiated and on understandings as to

the need to consider unique contextual traits that form language assessment cultures, the researchers set out to probe assessment literacy development and practice among teachers in five schools in Israel. Following the delivery of an in-service course on formative assessment and assessment literacy facilitated by one of the researchers, the study examined how and to what extent the knowledge and practice gained in the course were integrated into the schools' assessment culture. The sample included heads of secondary school language departments and school administrators. Interviews were used to elicit possible changes in assessment knowledge, decisions, and practices in the school community that reflect formative socially attuned assessment, as may be attributed to the in-service course by participants. The data yielded four major themes which originally appeared in an alternative assessment study (Hargreaves, Earl, & Schmidt, 2002): cultural, technical, political and post-modern. A link was established between the schools' overall profiles (characterized in the study as either conservative or innovative), and the extent to which the theories, principles and practices introduced in the course were integrated into the school community discourse, perceptions and actions. The study makes a case for working with the school community to create differential dialogically constructed assessment scripts to match local school cultures.

Paper 2: Trajectories of language assessment literacy in a teacher-researcher partnership: Locating elements of praxis through narrative inquiry

Luke Harding, Lancaster University, **Tineke Brunfaut**, Lancaster University

In 2011, a researcher-teacher partnership was set-up between Lancaster University (UK) and a team of secondary school teachers in Luxembourg. The aim of the partnership was to (a) explore the potential for redesigning the national end-of-secondary-school English exam to ensure alignment with current approaches to language teaching in the classroom, and (b) help to develop the teaching team's language assessment literacy, and their capacity to carry out high-stakes language test development work (Brunfaut & Harding, 2018). The partnership provides fertile ground for exploring the concept of "praxis" (Lantolf & Poehner, 2014) and its relationship to current understandings of language assessment literacy (LAL). In this paper we explore these issues through narrative inquiry; specifically, an analysis of written narratives produced by two teachers and two researchers reflecting on their experiences of the project over the past six years. Drawing on Pavlenko's (2007) approach to the analysis of autobiographical narratives, we will demonstrate how narrative inquiry can provide evidence of trajectories of language assessment literacy over time, as well as reveal relations between key characters, and identify complicating factors within overarching plots. The paper will conclude with a reflection on the usefulness of narrative inquiry as a method for exploring a praxis perspective on language assessment literacy.

Paper 3: Elaborating L2 Dynamic Assessment Through Praxis

Matthew E. Poehner, The Pennsylvania State University, **Remi van Compernoll**, Carnegie Mellon University

Dynamic Assessment (DA) derives from L. S. Vygotsky's (1987) argument that engaging with learners dialogically when they have reached the limits of their independent functioning brings to light underlying sources of difficulty and creates opportunities for them to stretch their abilities (Haywood & Lidz, 2007). This presentation reports the implementation of DA with U. S. university learners of L2 French as part of an effort to help teachers understand learner readiness to move from basic language courses to ones focused on literary and culture studies through the L2. Specific features of French were targeted based upon curricular objectives in the program and areas of common struggle as reported by teachers. Implementation of DA followed a pre-test – intervention – post-test design, which allows for differentiation of learners according to their initial performance and degree of improvement manifested

through the procedure (see Sternberg & Grigorenko, 2002). However, to better serve teacher interest in aligning instruction to learner needs, it was determined that information provided by sets of scores needed to be supplemented with details of mediating moves that proved beneficial to individuals. Discussion highlights the praxis orientation of the project, wherein DA is understood as a conceptual framework to address local assessment needs while, in dialectic fashion, being itself elaborated. Specifically, a shift in perspective is proposed from viewing pre- and post-test scores as 'snapshots' of learner abilities at discrete moments in time to instead understanding them as parts of a picture of development as it occurs *through* time.

Paper 4: Mediation in the assessment of language learning within an interlingual and intercultural orientation: Reciprocal interpretation and emergent understandings

Angela Scarino, University of South Australia, Australia

Over the past two decades, consideration of the nature and role of culture in foreign language learning has gained renewed prominence in the educational work of teachers. The expanded understanding of language learning to include an interlingual and intercultural orientation to language teaching and learning has presented major challenges for assessment on the part of teachers. In this presentation I discuss the collaborative objectives and processes of a three-year collective case study that investigated teacher assessment of student language learning within this orientation. A team of researchers worked with fifteen highly experienced teachers of a range of languages in both primary and secondary school settings (i.e., kindergarten through twelfth grade) over the course of two year-long assessment cycles. The researchers accompanied the teachers through their work on conceptualizing the nature of language learning within an interlingual and intercultural orientation, designing assessment experiences to elicit language learning; gathering classroom learning and assessment data; analysing and judging samples of student work; and evaluating the overall process. Through ongoing *facilitated dialogues* researchers mediated collaborative, interpretive analyses together with the teachers as a group. These analyses focused on identifying and explaining the emergent characteristics of the teachers' assessment designs as well as evolving understandings – on the part of both teachers and researchers – of the expanded construct of language learning and specifically the phenomenon of assessment of language learning within an interlingual and intercultural orientation.

Aligning language tests to external proficiency scales: validity issues

Organizer: **Sha Wu**, National Education Examinations Authority, Ministry of Education, China
Moderator: **Lianzhen He**, Zhejiang University, China
Discussant: **Jianda Liu**, Guangdong University of Foreign Studies, China

8:35 a.m. to 10:35 a.m. Location: Decatur B

The past decade has seen emerging interests in aligning tests to external performance scales or frameworks, e.g. the CEFR. Such alignment gives more meaning to the test scores through association with the descriptions of proficiency that comprise the external framework. Test score alone is not very informative for test users to make decisions, but a test score with information locating learners at a performance level defined by "can do" descriptions can provide a clearer understanding of what that test score really means. Test alignment is to make a claim on the interpretation of test scores in relation to performance levels. To support such a claim, test alignment is a process of validation, requiring multiple sources of evidence.

Aligning an international examination to a localized framework can help make the meaning construction of the test score more relevant to the context of use. This symposium attempts to illustrate the process, discuss the challenges and explore the implications of such a process with two parallel research projects on linking IELTS and TOEFL to China's Standards of English Language Ability (CSE), i.e. a set of comprehensive performance scales of English proficiency specifically designed for the Chinese EFL context. With the joint efforts from test providers and academic institutions from UK, USA and China, similar research methodologies are employed for these two research projects. By adapting the steps of familiarization, standardization, specification, standard setting and validation proposed in the Manual for Relating Examinations to the CEFR (Council of Europe, 2009), the two projects put special emphasis on constructing a chain of validity evidences, including internal validity, procedural validity and external validity.

This symposium includes four talks. The first talk will discuss the significance of conducting the test alignment research and the role of educational and social context in the alignment process. The second talk will focus on how to build up the concept of "just qualified candidates", which is the most important part of standard setting. The third talk, based on the AUA framework, will discuss how to evaluate the quality of the standard setting process and the credibility of the results with internal and procedural validity evidences. The fourth talk will address the value and ways of collecting evidence of external validity.

The symposium will start with a brief introduction, leading to four 20-minute presentations, followed by a 15-minute discussion and a 20-minute Q&A session.

Paper 1: Contextualized considerations in aligning tests to external performance levels

Sha Wu, National Education Examinations Authority, Ministry of Education, China, **Han Yu**, National Education Examinations Authority, Ministry of Education, China

It's a common practice to report the test results in scores, either in the form of individual numbers or bands. The test score alone is insufficient to inform test users of what test takers with that score (or a score within the defined band) can do. This leaves test users with inadequate support to make meaningful decisions about selection, placement, etc. Aligning tests to external performance levels is believed to provide meaning to the scores (Kane, 2012). Test results, however, are often used within an educational or social context; and test takers' cognitive processing also interacts with contextual factors. Thus, context should be taken into consideration when interpreting the test scores in relation to external performance levels.

The China's Standards of English Language Ability (CSE), released by the Ministry of Education and National Language Commission of China in 2018, includes comprehensive English proficiency scales covering the full range of EFL learners in China. Built upon a large-scale empirical study, the CSE was designed within and for China's specific context of use. I will draw upon two parallel research projects on aligning IELTS and TOEFL iBT to the CSE to illustrate how score meaning may be constructed and the role of educational and social context in the alignment process. It is envisaged that these research projects will promote the validation of these tests in Chinese context, as well as of the CSE.

Paper 2: Standard Setting: Constructing Descriptions of the Just Qualified Candidate

Richard J. Tannenbaum, Educational Testing Service, **Spiros Papageorgiou**, Educational Testing Service, **Ching-Ni Hsieh**, Educational Testing Service

Scores from tests of English as a foreign language (EFL) are often used to classify students into different levels of English skill or proficiency. The levels are progressive, representing increasing skill or proficiency as one moves from one level to the next higher level. For example, if there are three levels of English reading proficiency, Level 2 represents greater reading proficiency than Level 1, and Level 3 greater proficiency than Level 2. In order to know which students are performing at what levels, the lowest test score needed to classify a student at each level must be identified. The process followed to accomplish this is known as standard setting. Standard setting relies on the informed judgment of one or more panels of experts. For EFL testing, experts include teachers of English who also teach students within the grade span targeted by the test. A critical step in standard setting is working with the experts to build descriptions of the skills expected of a student who deserves to enter that level. This student has the minimum skills needed to be classified into the particular level. This student is sometimes referred to as the "borderline student" or the "just qualified candidate." In this presentation, we discuss how "just qualified candidate" descriptions are developed, drawing upon a recent standard setting workshop conducted to map TOEFL iBT® test scores to levels 4 through 8 of China's Standard of English Language Ability (CSE).

Paper 3: Internal validity issues in linking language assessments to reference levels

Shangchao Min, Zhejiang University, **Hongwen Cai**, Guangdong University of Foreign Studies, **Jie Zhang**, Shanghai University of Finance and Economics

The central concept of all linking practices has been validity, mainly encompassing procedural validity, internal validity, and external validity (Tannenbaum & Cho, 2014). Among these, internal validity is a core issue that deserves special attention, as it addresses the issues of the accuracy and consistency of the linking results. Despite an overall favorable result regarding internal validity in linking practices, previous linking attempts have met considerable challenges, for instance, divergent cut scores yielded by different standard-setting methods (Green, 2018; Kaftandjieva, 2010), divergent usage of the scales and descriptors by different panel members (Harsch&Hartig, 2015), and inadequately specified language frameworks for testing purposes (Lim, 2014). These rebuttals weaken the internal validity claims of linking practices, leading to a lack of meaningfulness of the score interpretations of the test aligned, as well as a lack of "comparability" between different tests aligned to the same level. In this presentation, I will first, following Papageorgiou and Tannenbaum's (2016) work, situate the refined and augmented procedural and internal validity issues into the assessment use argument (AUA) framework. Then we will illustrate its use through two case studies on linking two international EFL assessments, TOEFL and IELTS, to the China's Standards of English (CSE), with comparable groups of panelists, following similar standard-setting methodologies. I will conclude by presenting some of the preliminary findings of the two linking studies and discussing broader implications for integrating linking procedures into the AUA framework.

Paper 4: External validity: supporting alignment claims with multiple strands of evidence

Jamie Mark Dunlea, British Council, **Richard Spiby**, British Council

Supporting claims of alignment between a test and standards is a complex endeavour. In the body of literature accumulated over the last two decades on linking tests to the Common European Framework of Reference (CEFR), standard setting has been “at the core of the linking process” (Kaftandjieva, 2004), often implemented through the traditional approach of panels of judges applying either test-centred or examinee-centred methods. In developing a methodology for linking exams to the CSE, O’Sullivan (2017) suggested an integrated approach to collect multiple strands of evidence, with panel-based standard setting playing a central, but partial, role. The proposed linking framework draws on the three categories of evidence proposed by Cizek and Bunch (2007), procedural, internal, and external, and adds the important a priori step of construct definition. This presentation explores the external validity stage. While Cizek and Bunch (2007) suggest that multiple standard-setting methods can create problems in interpretation, Kane (2001) suggests replication with different methods “would provide an especially demanding empirical check.” In piloting the CSE linking methodology with IELTS, various forms of external validity evidence were collected, including an application of the Contrasting Groups procedure, drawing on an approach demonstrated by Dunlea and Figueras (2012). In addition to bolstering empirical evidence, this method benefits from the widespread engagement of teachers in the alignment process. This is particularly relevant for the CSE, an ambitious project with the goal to drive language education and assessment reform in China.

Transformative teacher-researcher partnerships in language assessment

Organizer: **Beverly Baker**, University of Ottawa

8:30 a.m. to 10:30 a.m. Location: Decatur B

This symposium collects together work that responds directly to Shohamy's (1998) call for "more democratic models of assessment where the power of tests is transferred from elites and executive authorities and shared with the local levels, test takers, teachers, and students." The contributors to this symposium have all attempted to involve non-language testing specialists, specifically teachers and graduate student instructors, as equal partners in language test development or research.

We will discuss our projects in terms of participatory or emancipatory research approaches, as well as through the lens of language assessment literacy (LAL). We consciously avoid a narrow deficit-based interpretation of LAL (which risks focusing only on what teachers need to learn about language assessment) and include an examination of how researchers also develop their LAL during these projects—especially in terms of knowledge of the specific assessment context and of appropriate assessment content.

Enacting democratic principles in educational settings is not without its challenges, obstacles, and setbacks, which must be acknowledged. Any two groups of stakeholders will necessarily approach a project with different perceptions and preoccupations. Divergence can arise at various stages of the project, including setting objectives and planning, execution, and dissemination of findings and recommendations. This divergence can compromise the success of the project and has ethical implications as well. As Winkler (2003) states, "The validity and ethical defensibility of collaborative research ultimately depends on the critical acknowledgment of multiple realities…" (p. 400).

In this symposium, we therefore take a self-reflexive stance, where we attempt to make explicit the multiple realities of the different stakeholders. We do more than trumpet our successes; we critically examine the extent to which the projects truly realized their democratic intentions, and the extent to which language assessment researchers were able to position themselves successfully as equals with their partners during project decision-making.

As opposed to the traditional symposium format, we will not have a discussant. Instead, we plan to each include in our individual contributions comments on the common threads that connect all our work as discussed above, and that align with the LTRC theme of social justice. Therefore, following our presentation of our research work, we each offer a few critical reflections on the LAL development of all parties, as well as the successes and challenges we had in creating a truly equal and inclusive research partnership.

Paper 1: A critical examination of the success of two researcher teacher partnership projects in language assessment

Beverly Baker, University of Ottawa

In this contribution, I report on two recent attempts to incorporate inclusive and emancipatory practices in language teaching and assessment projects: one with a professional learning community of English teachers in Haiti, and one with an Indigenous immersion classroom teacher in her home community.

Inclusive and emancipatory research are umbrella concepts include critical, participatory, and transformative research approaches, such as community-based participatory research, or decolonizing research. Whatever the preferred term, researchers in these areas are all concerned with transforming the research process by recognizing of the equal importance of all stakeholder contributions—especially those historically marginalized by traditional research practices—and eliminating hierarchies by actively resisting the traditional researcher-research subject dynamic.

After a brief summary of each project, I will compare the two projects and discuss them in terms of the LAL development of all parties, using an adapted form of the framework for the elements of LAL introduced by Taylor (2013). This framework enables a nuanced portrait of the complementary nature of the differential expertise of each stakeholder in the assessment process.

In addition, I will engage in a critical reflection of the extent to which the research process in each project was truly democratic. While some successes in these areas were apparent, there was evidence that the position of power of the researcher in the Haitian project may mask some sources of tension, and that negotiation and compromise in goal-setting is a constant preoccupation in the Indigenous project.

Taylor, L. (2013). Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections. Language Testing, 30(3), 403-412.

Paper 2: Graduate student perspectives of transformational teaching and language assessment literacy

José Manuel Martínez, Department of Teacher Education, Michigan State University, **Ni-La Lê**, Department of Linguistics and Languages, Michigan State University, **Erika B. Kraus**, Department of Forestry, Michigan State University, **Azad Hassan**, School of Planning, Design and Construction, Michigan State University, **India C. Plough**, Assistant Professor, Residential College in the Arts and Humanities, Michigan State University

Language Assessment Literacy (LAL) is developing as "a dynamic loose framework… that sets general guiding principles for different assessment literacies [and] is aware of local needs and is loose enough to contain them and allow them to develop from theory to practice, [and] also from practice to theory" (Inbar-Lourie, 2017). Contributing to the "situated differential conceptualizations of LAL," we describe a small-scale, exploratory study into the teaching and assessment components of a localized Cultures and Languages Across the Curriculum (CLAC) program from the perspectives of all participants.

Based in an undergraduate residential college in the arts and humanities, the CLAC program is grounded on the premise that the processes of teaching, learning, and assessment are interdependent and symbiotic. Following from this, and guided by practices of Critical Language Testing (Shohamy, 2001a, 2001b), all participants enter the CLAC Program with the understanding that they are the creators of the program. While disciplinary 'expertise' naturally varies among colleagues, the evolution of the program intimately depends upon the voices of all stakeholders.

To provide context, we briefly describe the program, highlighting participant development of key components. We then focus on the perspectives of the graduate students who co-lead instruction and assessment. Drawing on questionnaires, feedback discussions of video recordings, and guided interviews, graduate students reflect on engagement with the core methods of transformational teaching (Slavich and Zimbardo, 2012), resources relied on, challenges that hindered engagement, influential contextual factors, and LAL opportunities through engagement with undergraduate students and colleagues in the CLAC program.

Inbar-Lourie, O. (2017). Language Assessment Literacy. In Shohamy, E. Or, I. G., & May, S. (Eds.). *Language Testing and Assessment*, Encyclopedia of Language and Education, DOI 10.1007/978-3-319-02261-1_19.

Shohamy, E. (2001a). Democratic assessment as an alternative. *Language Testing, 18*(4), 373–391.

Shohamy, E. (2001b). *The power of tests: A critical perspective*. London: Longman.

Slavich, G.M. & Zimbardo, P. G. (2012). Transformational Teaching: Theoretical Underpinnings, Basic Principles, and Core Methods. *Educational Psychology Review*, 24, (4), 569-608.

Paper 3: The development of a profile-based writing scale: How teachers and testers develop language assessment literacy through collaborative assessment practices within a post-admission ESL writing program

Xun Yan, University of Illinois at Urbana-Champaign, **Ha Ram (Hannah) Kim**, University of Illinois at Urbana-Champaign, **John Kotnarowski**, University of Illinois at Urbana-Champaign, **Hyunji (Hayley) Park**, University of Illinois at Urbana-Champaign

This study reports on the revision of a rating scale for an ESL writing placement test and demonstrates how teacher-tester collaboration can help both groups to develop language assessment literacy and enhance the assessment practice within the ESL writing program. Following a data-driven approach, teachers participated in a three-stage scale revision process, where they (1) reflected on the range of writing performances in ESL courses, (2) evaluated sample test essays and revised the scale descriptors, and (3) pilot-rated essays using the new scale.

During the first stage, both teachers and testers recognized that test-takers display different strengths and weaknesses in argument development and lexico-grammar. However, when evaluating sample essays, teachers weighted argument development more heavily whereas testers stressed lexico-grammatical accuracy. Additionally, when rating argument development, some teachers relied heavily on surface/structural rhetorical features rather than essay content. The contrasts stemmed from the testers' unfamiliarity with the ESL curriculum as well as the teachers' unwillingness to assess students' writing needs as reflected in essay performances. These contrasts resulted in conflicting ratings on certain essay profiles, making it difficult to arrive at agreed-upon placement levels and descriptors.

Through several rounds of discussion, these differences were eventually mitigated by creating separate criteria for argument development and lexico-grammar. The revision process standardized the conceptualization and operationalization of writing quality, shifting teachers' focus from surface rhetorical features to essay content. Meanwhile, collaboration with teachers enhanced testers' understanding of the local instructional contexts. Teachers' involvement promoted collaborative assessment-related dialogues and practices within the ESL program, strengthening the alignment across curriculum, instruction and assessment.

Paper 4: Collaborative research approaches to classroom-based assessments with linguistically diverse communities

Jamie L. Schissel, University of North Carolina at Greensboro, **Constant Leung**, Kings College London, **Mario López-Gopar**, Universidad Autónoma "Benito Juárez" de Oaxaca (UABJO), **Julio Morales**, Universidad Autónoma "Benito Juárez" de Oaxaca (UABJO), **James R. Davis**, University of North Carolina at Greensboro

For linguistically diverse communities, researchers have questioned assessment approaches that disregard learners' multilingualism (Authors, 2018; Otheguy, García, & Reid, 2015, 2018; Shohamy, 2011). While little research exists on multilingual language assessments, our study offers potential approaches that center contextualized, multilingual knowledges and skills within classroom-based language assessments. Our paper delves into the methodological concerns that are at play when designing and integrating such an assessment approach with one teacher and his English foreign language courses for pre-service English teachers in Oaxaca, Mexico.

We combined critical ethnographic and participatory action research (PAR) methodologies to foster the collaboration with the teacher and pre-service students, which shaped the assessment design and implementation. In this paper, we detail our methodological approaches and include views about the

resulting assessment as reported by the pre-service teachers in post-assessment surveys and stimulated recall-interviews, and weave reflections on the different roles taken on by the authors to foster the collaboration throughout the article. Our work points to the promise of adopting these approaches that positioned researchers as individuals who support participants. We discuss the implications of our approach not only in the creation of assessment instruments, but how this collaboration also impacted teaching approaches and laid the foundation for sustaining and expanding the collaborative work with the teacher. We conclude with recommendations for adapting this contextualized collaborative approach in other settings.

Understanding teacher educators' language assessment literacy practices while training pre-service English teachers in Chile

Salomé Villa Larenas, Lancaster University

10:55 a.m. to 11:25 a.m. Location: Mary Gay C

Arguably, in order to promote fairness and social justice in language testing it is vital that key stakeholders possess a multidimensional level of language assessment literacy (LAL). This entails possessing language assessment knowledge and skills, being aware of and adhering to guidelines for good practice, and being able to place these within broader social and political contexts (Pill & Harding, 2012). For future English language teachers, developing their language assessment literacy may largely depend on the quality of assessment knowledge and practices they see modelled by those who train them, i.e. teacher educators (Graham, 2005). Hence, the language assessment practices adopted by teacher educators, when training pre-service teachers, may prove to be pivotal in ensuring fair language assessment practices by the next generation of teachers. Additionally, they may be crucial in enabling teachers to critically evaluate assessment policies within their own professional context.

However, to date, little research has focused on the LAL of teacher educators, while greater attention has been given to language teachers themselves and to a range of other stakeholders (e.g., Hasselgreen et al., 2004; Pill & Harding, 2012). Also, the majority of studies have focused on stakeholders' gaps in language assessment knowledge, and have investigated this by means of questionnaires (e.g., Berry & O'Sullivan, 2016; Vogt & Tsagari, 2014). Less research attention has gone to assessment practices and uses of assessment materials. Therefore, scholars have started to call for a richer research agenda on LAL – moving beyond needs and self-reported knowledge, and towards the complex interactions between language assessment beliefs, knowledge, and contextual constraints (e.g., Levy-Vered & Nasser-Abbu, 2015; Xu & Brown, 2016).

In response, I designed a mixed-methods study to explore the LAL of teacher educators in terms of their theoretical knowledge and practices and the extent and nature of the LAL training they conduct with pre-service teachers. A LAL 'test' was developed to evaluate teacher educators' LAL knowledge, while their LAL practices were investigated by means of interviews and an analysis of the assessment materials they use in their teacher training. The study was conducted with teacher educators working in English as a foreign language (EFL) teacher education programmes in Chile.

In this talk, I will report on the interviews that were conducted with 24 EFL teacher educators from seven Chilean universities. Following the Teacher Assessment Literacy in Practice model by Xu and Brown (2016), the aim of the interviews was to explore: the macro socio-cultural and the micro institutional contexts in which teacher educators work, teacher educators' cognitive and affective conceptions of assessment, and the compromises they make in their practices to find a balance between contextual constraints and their perceptions of good standards. The interview findings provide insights into the LAL of teacher educators and the LAL training they conduct with pre-service teachers. The study has implications for the design of teacher education curricula and for policy making in Chile with respect to teacher education. More generally, the study sheds light on the profile of the underresearched group of teacher educators.

Reading Self-Concept and Reading Achievement in Monolingual and Multilingual Students: A Cross-Panel Multiple-Group SEM Analysis

Christopher Douglas Barron, University of Toronto

10:55 a.m. to 11:25 a.m. Location: Henry Oliver

A major debate within the self-concept literature is whether reading self-concept (RSC) is a product, rather than a predictor, of reading achievement (skill development hypothesis), or whether RSC and reading achievement mutually influence each other (reciprocal effects hypothesis). The majority of articles analyzing RSC and reading achievement have inadequate research designs to support the causal assumptions imbedded in research questions. Without utilizing longitudinal structural equation modeling, there is minimal justification in making inferences regarding causality (Little, Preacher, Selig, & Card, 2007; Pedhazur & Schmelkin, 2013).

Despite growing numbers of students with a non-English home language background, few studies have investigated the measurement and structural invariance of RSC and reading achievement measures among monolingual and multilingual students (Niehaus & Adelson, 2013). Without sufficient measurement invariance, the validity of instruments measuring RSC and reading achievement among multilingual students is questionable. The purpose of present study was to determine the causal relationship between RSC and reading achievement, and evaluate the extent to which RSC and reading achievement assessments were measured equivalently for students with multilingual backgrounds.

Secondary data from a province-wide standardized reading assessment and student background questionnaire was utilized. A cohort of students were compared on their Grade 3 and Grade 6 RSC and reading achievement scores. Reading achievement was measured using two English reading assessments that included four reading passages and 32 items, integrating both multiple-choice and open-response question formats. RSC was measured at each timepoint using a 3-item measure. Students' language status was calculated through self-report of home language use. Students who reported only speaking and hearing English at home were coded as monolingual, while students who reported both speaking and hearing another language as often or more often than English at home were coded as multilingual.

Analysis involved two cross-panel structural equation models with two assessments between reading self-concept and reading achievement. The first model included monolingual students (n= 18, 953) and while the second included multilingual students (n= 8695). Model specification followed the guidelines suggested by Marsh and Martin (2011). Measurement invariance was then calculated by comparing configural, metric, scalar and residual invariance across the two models (Meredith, 1993).

Both models had excellent model fit (RMSEA<.05, SRMR <.05, and CFI>.9). Results provided evidence for the reciprocal effects model among monolingual students, but the skill development model among multilingual students. For monolinguals, RSC was both a predictor of, and predicted by, reading achievement across the three years. In comparison, multilinguals had a significant relationship between reading achievement and subsequent RSC, while RSC failed to predict subsequent RA. Measurement invariance results indicated a significant drop in model fit when testing metric invariance. Thus, monolingual and multilingual students had significantly different factor loadings on reading self-concept, invalidating cross-group comparisons regarding mean or covariance structures.

Implications of this study emphasize the need for the development of culturally-sensitive measures of reading self-concept and reading achievement, and highlight the importance of testing measurement invariance before comparing group means or regression coefficients across monolingual and multilingual students.

Rater cognition and the role of individual attributes in rating speaking performances

Kathrin Eberharter, University of Innsbruck

10:55 a.m. to 11:25 a.m. Location: Decatur A

Rating quality is a fundamental aspect of validity in performance assessment. The argument-based approach to test validation (Kane, 1992; 2006) places rating scale design and rating procedures almost at the beginning of the interpretive argument (Knoch & Chapelle, 2017). As raters interpret and apply a test's criteria, they become the arbiters of the test construct as expressed in the scores they award (McNamara, 1996). Thus, any systematic weakness of the rating process will potentially impact on the meaning of test scores and on the decisions that are made on basis of these scores. However, as is evident from research in our field, the rating of language performance is highly complex (McNamara, 1996), influenced by numerous factors (Lumley, 2002) and may function as a source of construct-irrelevant variance (Bachman and Palmer, 2010). In order to reduce variance caused by the complex nature of human judgement, there has been an increased interest in understanding rater cognition; investigating the effects of rater attributes on scores or raters' mental processes during rating (Bejar, 2012; Suto, 2012). However, rater cognition research in language testing is yet to engage fully with wider research on judgment and decision-making – important areas of scholarship in cognitive psychology, economics and management science – despite clear synergies between decision-making behaviour in these fields. Findings from this literature suggest that complex decision-making tasks are greatly influenced by a wide range of factors including processing capacities, perception, the interplay of deliberate and automated thinking, and metacognitive control (Newell and Bröder, 2008). Exploring these factors would prove useful in understanding the complex task of assessing second-language speech in real time.

The current study investigated influences on rater cognition in the context of a national English school-leaving speaking examination. Specifically, the project aimed to establish the influence of decision-making style, preferred cognitive style, working memory and executive function on rater accuracy and consistency. Following recruitment and training, 39 pre-service English-as-a-Foreign-Language teachers rated a set of video-recorded speaking performances (N=30), filled out two validated psychological questionnaires (targeting decision-making and cognitive style), and completed a battery of cognitive tests (targeting working memory and executive function). Data were analysed using a combination of Many-Facets Rasch Measurement (MFRM) and inferential statistics to determine (a) rater fit and alignment with a set of benchmark ratings, (b) rater profiles on the decision-making and cognitive attributes, and (c) relationships between individual attributes and measures of rater quality. Findings demonstrated that there were considerable individual differences among raters on some variables despite recruiting from a homogeneous group. Relationships between variables suggest a complex association between the attributes measured and rater quality. The paper will also discuss the second phase of the study, in which raters representing different profiles will be selected to complete further rating tasks with eye-tracking and stimulated verbal recalls. Finally, the need for expanding the range of cognitive attributes considered as influential on rater judgment will be discussed, and recommendations will be made for rater training.

Test Consequence on Rural Chinese Students: Investigating Learner Washback of National College Entrance English Exam

Yangting Wang, The University of Texas at San Antonio, **Mingxia Zhi**, The University of Texas at San Antonio

10:55 a.m. to 11:25 a.m. Location: Decatur B

Washback, an important aspect of consequential validity (Messick, 1996), refers to the effect, or consequence of a test on language learning and teaching (Abeywickrama & Brown, 2010). National College Entrance English Exam (NCEEE) is a mandatory high-stakes test for high school students to enter higher education in China. However, little attention was paid to examine the impact of this exam, let alone its impact on the economically disadvantaged students from rural areas of China. The current study used a concurrent partially mixed equal status research design (Leech & Onwuegbuzie, 2007) to investigate the consequence of NCEEE on student affects, learning process, and outcomes in a rural high school in Southeast China.

The present study followed Bailey's (1999) "washback to learners" and Hughes (1993) tripartite model of washback. A pre and post-survey were administered with senior students (N = 140) before and after taking NCEEE. The pre-survey included 115 Likert-scale items regarding students' motivation, anxiety, learning attitudes, learning activities, test-taking strategies, and one open-ended question. The post-survey included five Likert-scale items and eight open-ended questions. NCEEE scores were obtained. Quantitative and qualitative data were analyzed independently and then triangulated for interpretations (Leech & Onwuegbuzie, 2007).

Descriptive results indicated that low-socioeconomic status (SES) participants had limited outside resources in learning English. Most participants reported being nervous (62.6%) and stressed (82.1%) over this exam. Their learning activities involved vocabulary memorization (98.6%) and grammar drilling (90%) and rarely engaged in authentic materials (29.9%). Regression analysis demonstrated that NCEEE scores reflected students' self-perceived listening and reading scores, but not speaking or writing scores. After controlling for covariates and in-class learning activities, family income significantly and positively predicted the test scores, but the parent education yielded a non-significant effect. Before-test stress and the stress-about-results negatively predicted NCEEE scores, while during-test stress yielded positive association. In addition, "hours spent on English not directly related to test-preparation" negatively predicted the scores.

Qualitative analysis was concluded using Emerson, Fretz, and Shaw's (2011) open and focused coding techniques. Qualitative data support the quantitative results to some extent. First, six students felt "painful" and "exhausted" preparing NCEEE and ten students commented on the "teach to the test" education system. However, eleven students reported that the test also helps "understand the importance of English" and "felt motivated in learning English". Second, in terms of stress, some students believed that some pressure can make them be more focused during the test, but too much anxiety leads to distraction and lower scores. Third, consistent with quantitative results, students commented that NCEEE does not assess oral skills and is "not real-life English". However, contradicting to the regression results, eight students believed that this test help to "reduce the gap between poor and rich" and is a fair test that offers a chance for students with low-SES to go to universities.

The findings indicated negative washback effects of NCEEE in developing communicative skills and its mixed effects on students' perception and attitude towards English learning. Suggestions on NCEEE test reform from the students' perspectives are discussed.

Using the Language Assessment Literacy Survey with an Under Researched Population: The Case of Uzbekistan EFL Teachers

David Lawrence Chiesa, U.S. Department of State

11:30 a.m. to 12:00 p.m. Location: Mary Gay C

One subject matter area that is central to a teachers' professional practice is that of assessment. Scholars in the area of language assessment have long been interested in investigating what is known as language assessment literacy (LAL) – the level of a teacher's engagement with constructing, using, and interpreting a variety of assessment procedures to make decisions about a learner's language ability (Taylor, 2013). Particularly, Kremmel and Harding have been developing a data collection tool called the Language Assessment Literacy Survey in order to "create a comprehensive survey of language assessment literacy which can be used for needs analysis, self-assessment, reflective practice, and research" (LTRG, 2018). The study used Kremmel and Harding's Language Assessment Literacy Survey and focused on a group of 96 in-service university English as a Foreign Language (EFL) teachers in Uzbekistan. The overarching research question is: To what extent does the Language Assessment Literacy Survey provide valid and actionable information about teachers' language assessment literacy? To answer this inquiry, the following three subquestions were addressed: A. What assessment skills and knowledge do Uzbekistan EFL teachers believe teachers need to possess? B. Is the theoretical basis for the survey (Taylor's 2013 framework of LAL) faithfully implemented in the survey? C. What is the factor structure present in the Uzbekistan EFL teachers' responses?

From the EFL teachers' responses, descriptive statistics were calculated for each item and the overall survey. These results generally explained Uzbekistan EFL teachers' cognitions about which assessment knowledge they believe language teachers (in general) need to know, and what assessment skills they think language teachers should possess. To answer the second question, an external review of the journal issue Language Testing, 30(3) – a special journal issues focusing on language assessment literacy – was conducted to identify possible definitions that Taylor (2013) used to create her Language Assessment Literacy Framework, and which Kremmel and Harding used to write their survey items. From this review, operational definitions of each language assessment literacy dimension (Taylor, 2013) were created and subsequently, the survey items were coded with two other coders (a = 0.91). These results provided information about the English, Uzbek, and Russian versions of the survey and whether they adhered faithfully to Taylor's (2013) framework. To answer the third question, an Exploratory Factor Analysis (EFA) was conducted to identify the factor structure inherent in the Uzbekistan EFL teachers' responses. In terms of the results, Kremmel and Harding's Language Assessment Literacy survey shows evidence of validity. In its current state, it can support future research in language assessment literacy, and it can provide support as a data collection tool to be used by teacher educators with new or different populations.

What did the Reading Assessment Miss?

Elizabeth Lee, Iowa State University

11:30 a.m. to 12:00 p.m. Location: Henry Oliver

Genre, task types, and question types influence how readers process and respond to reading assessments (Grabe, 2009; Read, 2000). An evaluation of these factors enables researchers to survey the extent to which a reading assessment measures the targeted reading construct. Previous studies have found that L2 readers use both academic reading skills and test-taking strategies when taking the TOEFL reading test (Thomas & Upton, 2006), and that uses of reading strategies vary between testing and nontesting contexts (Chou, 2013). These studies show that reading assessments may not perfectly measure the reading construct, however, these findings have rarely been examined with locally developed reading assessments. A possible reason may be that these assessments are considered low stakes and consequences are not deemed as severe as large-scale standardized tests. As a result, test developers may not endeavor to conduct validation studies or make serious improvements to these assessments (Johnson & Riazi, 2017). To address this oversight, this presentation evaluates a locally developed reading assessment that was used at a large midwestern university, in an attempt to demonstrate the consequences of neglecting validation studies. The reading passages and items are evaluated to determine the degree to which the interpretations of the reading test scores reflect the ESL reading course domain. To guide this evaluation, the following questions were raised: First, what linguistic features (i.e., lexicogrammar and genre) appear in the course and assessment texts? Second, what types of tasks and questions are found in the course and assessment texts? Corpus and SFL analyses were used to examine features of lexicogrammar, genre, task types, and question types found in course and assessment texts. To identify lexicogrammatical features, specific types of grammar common in academic texts (Biber, 2006), including nouns and verbs, were investigated using AntConc (Anthony, 2018). Genre was analyzed in terms of the text organization and linguistic features characteristic of text features (Derewianka, 2016). Task types were determined using Alderson's (2000) categorization of reading tasks. Question types were analyzed using Gerot's (2005) five categories of questions, where questions are defined in terms of how the answer is encoded within the text. Results show that the lexicogrammatical features did not vary between course and text tests. On the other hand, genre, question, and task types showed variation. A broader range of assessment techniques and readings were used in the course compared to assessments, which were limited to uses of multiple choice and gap-fill tasks and argument texts. The results point to a weak validity of the use of this reading assessment and that serious revisions were needed. An analysis of discourse feature need to be considered along with lexicogrammatical features when measuring a test's validity. This study demonstrates the value of qualitative evaluations of reading assessments in local contexts and recommends greater use of qualitative methods in the field of language testing. Clearly, attending to the validity of small-scale locally developed assessments matter as it affects the quality of education that ESL students receive at the tertiary level.

The Effect of Raters' Perception of Task Complexity on Rater Severity in a Second Language Performance-based Oral Communication Test

Yongkook Won, Iowa State University

11:30 a.m. to 12:00 p.m. Location: Decatur A

Even with the benefits of second language performance-based oral communication tests (Katz & Gottlieb, 2013; McNamara, 1996), a plethora of variables, including raters and interviewers, affect test scores and the chances of having construct-irrelevant variance in the test scores increase accordingly. Many studies have tried to identify and reduce the effects of raters and/or interviewers on the test scores in performance-based oral communication tests [such as raters' first language (Zhang & Elder, 2010), experience and training (Davis, 2016), familiarity with certain pronunciation varieties (Carey, Mannell, & Dunn, 2011; Yan, 2014), and gender (Brown & McNamara, 2004; O'Sullivan, 2000)]; however, most of these studies sacrificed their test authenticity by not allowing interviewers to adaptively choose interview questions depending on examinees' performance. Little attention has been paid to the use of performance-based oral communication tests while maximizing test authenticity. The present study investigates how raters evaluate examinees' performance in performance-based oral communication tests when interviewers can adaptively choose their questions, in terms of task complexity, responding to examinees' performance.

A sequential explanatory design with a mixed-methods approach (Creswell et al., 2003) was used in the current study. For the initial quantitative data analysis, 16 experienced raters judged 299 speech samples with four task complexity levels produced by international teaching assistants in the oral English communication test at a Midwest U.S. university. To investigate raters' behavior in a more controlled situation, nine novice raters were trained to judge 80 speech samples with two task complexity levels. A task-complexity-related three-facet (i.e., examinees, task complexity, and raters) partial credit model of the data with a many-facet Rasch measurement (MFRM) was used to analyze raters' use of an evaluation rubric depending on task complexity. Raters were also interviewed on their use of the evaluation rubric when they had the knowledge of task complexity. Friedman's test (Friedman, 1937) and Wilcoxon's test (Wilcoxon, 1945) were used to analyze the influence of task complexity on examinee scores and language use.

Comparison of Rasch-Andrich threshold (Masters, 1982) in the MFRM indicated that raters did not use the rating scale consistently across different task complexities. Friedman's tests indicated that fluency indices, but not holistic scores, were lower with more difficult tasks. This discrepancy between fluency indices and holistic scores suggests that raters were more lenient when they evaluated examinees' performance on more difficult tasks. Rater interviews also supported the claim that increasing the task complexity could decrease raters' severity.

Implications for MFRM modeling of the performance-based tests and rater training can be gleaned. As task difficulty in the MFRM model can only be correctly measured when raters do not change their rating severity depending on task complexity, the findings of this study enhance the understanding of task difficulty in the MFRM model when raters are exposed to task complexity levels. This understanding would provide empirical evidence of a new relationship between task types and raters in the oral communication assessment model (Ockey & Li, 2015) and would lead to the discussion of including a task complexity factor in rater training.

Examining Washback on Learning from a Sociocultural Perspective: The Case of a Graded Approach to English Language Testing in Hong Kong

Chi Lai Tsang, University College, London

11:30 a.m. to 12:00 p.m. Location: Decatur B

Washback on learning has been examined in language testing research for over a quarter of a century due, in part, to a growing awareness of the consequential effects of high-stakes testing for stakeholders, including in perpetuating or reinforcing social inequality (Shohamy, 2014). However, most studies have either focused narrowly on observable washback effects or on discrete mediating factors at the individual test-taker level in isolation of the wider sociocultural context (e.g., Xie & Andrews, 2012). The present study, therefore, examines the washback on learning construct more holistically by linking learner perceptions as a result of changes to the test, learners' individual characteristics, and macro-level socioeducational and societal factors operating within and beyond classroom contexts. The focus is on the large-scale Hong Kong Diploma of Secondary Education English Examination (HKDSE-English), which acts as gatekeeper in shaping test-takers' educational futures. Since 2012, HKDSE-English has incorporated a graded approach that allows test-takers to choose between easier or more difficult reading and integrated skills tasks. The study, thus, aims to uncover (1) the washback effects that learners identify following the introduction of this graded approach, (2) learners' self-reported intrinsic and extrinsic mediating factors shaping such effects, and (3) the paths of influence between the types of washback effects and the categories of mediating factors identified. Using an exploratory sequential mixed methods design, semi-structured focus groups involving 12 Hong Kong secondary students recruited from across all three school bandings were inductively coded to generate 15 learner perceived washback effects and 30 intrinsic and extrinsic mediating variables. Participants' comments were then built into a questionnaire, which was administered to another 150 learners. Two sets of exploratory factor analyses revealed that four major types of negative washback effects took place within and beyond the classroom, and the mediating variables fell under eight categories pertaining to the learners themselves, other stakeholders (classmates, teachers, family), and societal influences. Finally, four sets of simultaneous multiple regressions conducted between the eight mediating factors categories and each of the four macro-level washback effects showed that the strongest predictors varied for each washback effect type, but they were all significantly affected by at least one intrinsic and one extrinsic factor. Hence, the washback on learning construct appeared to be socially-situated and driven by a complex array of intertwining and potentially competing forces.

Taken together, the results suggest that the graded approach could disempower test-takers if they are coerced into a particular option, thereby reinforcing the power imbalance of Hong Kong's academically-streamed educational system. The study argues for the adoption of a fairer, more scientific way to inform test-takers' decision-making that takes into account their individual differences and seeks to empower (not subjugate) test takers' voices. The presentation concludes with recommendations for advancing the washback on learning construct, including rejecting the notion that washback on learning is rooted only in classrooms, expanding the scope of extrinsic mediating factors to include not only human agents, but also societal factors, and reconceptualizing washback on learning as a negotiated construct that involves the interplay between multiple influences.

Investigating the validity of a writing scale and rubric using a corpus-based analysis of grammatical features

Susie Kim, Michigan State University

3:20 p.m. to 3:50 p.m. Location: Mary Gay C

To make a reliable inference of language ability from a test score, the rating rubric for the test needs to include relevant constructs and corresponding descriptors. To be valid, these constructs and descriptors need to be identifiable in the examinees' response (Knoch & Chapelle, 2017). Therefore, one way to support test validity is to demonstrate if (1) the linguistic ability descriptors in the rating rubric reflect the examinees' actual language use, and (2) the linguistic features of examinee-produced texts differentiate between score levels. Previous validation studies on large-scale English exams (e.g., Banerjee et al., 2007; Biber & Gray, 2013) uncovered what aspects of linguistic features differed depending on the rater-assigned scores. However, they provided little discussion on the relationship between the rating scale and the linguistic characteristics observed at different score levels. To that end, this study investigates test validity and usefulness with regards to a rating scale, rubric descriptors, and learner language, focusing on the grammatical features identified in learner English. Specifically, this study aims to examine how well test candidates' use of a set of grammatical features and their frequencies, ranges, accuracy, and lexis differentiate between score levels, thus evaluating how well the rating rubric and rater-assigned scores reflect learner language characteristics.

The context of this study is a large-scale English exam, Certificate of English Language Competency, which certifies English B2 level of the Common European Framework of Reference (CEFR; Council of Europe, 2001; 2018). The writing section of the test requires examinees to write an opinion essay. Two raters rate each essay, using an analytic rubric with four categories (i.e., grammar, vocabulary, task completion, and genre appropriateness) on a scale of five performance levels. The rubric descriptors generally follow the CEFR, with the grammar category alluding to syntactic complexity, range, and accuracy. In order to provide context-relevant and concrete grammatical features, this study used literature and materials from the English Profile (CUP, 2015; UCLES/CUP, 2011) as a guide for selecting the target grammatical features. Data included 200 texts written on three different topics, ranging across all five levels of performance. The occurrences of 16 grammatical features were extracted from the corpus using Natural Language Processing tools and analyzed in terms of their frequencies of occurrence, length of structure, and the number of different features that appeared in each text. Additionally, the occurrences were manually annotated for errors and profiled according to the lexical items associated with each grammatical structure (e.g., used with frequent/infrequent words). These measures were then used as predictor variables in logistic regression analyses to examine the extent to which various aspects of grammar use differentiated between score levels. The findings will be discussed in light of the validity of the writing scale and scale descriptors as well as the linguistic features that influence writing scores. This study has implications for improving writing scale descriptors that are crucial for test management: explicating the descriptors of syntactic complexity, range, and accuracy with specific examples; and improving materials for raters and examinees such as benchmark samples.

Assessing Workplace Listening Comprehension of Thai Undergraduates in English as an Asian Lingua Franca Contexts

Panjanit Chaipuapae, Northern Arizona University

3:20 p.m. to 3:50 p.m. Location: Henry Oliver

Considering the status of English as a Lingua Franca (ELF) in Asia, Thai graduates need to be able to understand various English accents to be successful in their future careers. To determine students' readiness before they enter the workforce, a workplace listening test which includes Thailand's major trading partners—America, China, and Japan—is needed. However, no such standardized listening test exists. Accents, accent familiarity, and attitudes may also cause listening difficulty (Kang & Rubin, 2009; Ockey & French, 2016). Previous studies investigated these issues, but the findings were inconclusive (Harding, 2011). The purpose of this dissertation is to investigate the effects of accents on Thai undergraduates' listening comprehension. This examination includes three aspects. First, the study aims to provide justification for using a listening test as a readiness measure. Second, the effects of accented speech are investigated. Third, the roles of accent familiarity and attitudes on listening performances are examined.

The data collection takes place between August and December 2018. Participants include 140 Thai undergraduates divided into four mixed-proficiency groups. Four instruments are used. First, the English Learning Questionnaire considers students' initial ability levels, recent English courses, and course grades to determine their proficiency levels. Second, the Workplace Listening Test (CEFR level B2-C1) measures the ability to understand monologic talks delivered by speakers of American, Chinese, Japanese, and Thai accents. The test comprises eight listening passages; each passage has six 4-option-multiple-choice questions resulting in a total score of 0-48 points. Listening comprehension is measured at local and global levels (Becker, 2016). Thai experts' judgement is used to validate the items' difficulty level and to set a cut-off score. Third, the accent familiarity questionnaire, a 5-point Likert scale, adapted from Harding (2011) and Ockey and French (2016) measures students' exposure to four accents. The questionnaire consists of two portions—immediate and overall perceptions. The first portion is embedded in the listening test. After answering the questions of each passage, students immediately rate their familiarity with the speaker's accent they just heard. The second portion is administered at the end of the listening test. Last, the accent attitudes questionnaire, a 5-point Likert scale, adapted from Abeywickrama (2013) measures how much students like four accents in listening tests. The questionnaire consists of two portions—immediate and overall perceptions. After each listening passage, the immediate attitude portion is placed after the immediate familiarity question. The overall attitude portion is put together with the overall familiarity questionnaire. Using a replicated 4x4 Latin square, each group listens to eight speakers with different order and takes the listening test followed by the accent familiarity and attitudes questionnaires.

Results are reported in three main aspects. First, the interpretation/use argument framework is used to evaluate score interpretations and use (Chapelle, Enright, & Jamieson, 2008; Kane, 2013). Second, using three-way ANOVA, results of the effects of accents are reported. Last, using correlation analyses, the roles of familiarity and attitudes on listening comprehension are reported. Expected positive washback to stakeholders as well as implications for teaching and testing L2 listening with accents are addressed.

Justifying the Use of Scenario-Based Assessment to Measure Complex Constructs of Communicative Language Competence

Heidi Liu Banerjee, Teachers College, Columbia University

3:20 p.m. to 3:50 p.m. Location: Decatur A

With the vast development of digital technology and the widespread use of social network platforms, the competences required for academic and career success in the 21st century have expanded to include complex skills where individuals need to demonstrate their abilities to think critically, reason analytically, and problem-solve strategically (Shute et al., 2010). Consequentially, there has been a call to broaden the constructs of communicative language ability in L2 assessment to better represent the everyday language use in the modern society (e.g., Bachman, 2007; Purpura, 2015, 2016; Sabatini et al., 2014), so that the results of an assessment can yield interpretations that are aligned with the contemporary views on L2 knowledge, skills, and abilities (KSAs).

Scenario-based assessment (SBA), an innovative, technology-based assessment approach, shows great affordances for expanding the measured constructs of an assessment. Initiated by the CBALTM project (Bennett, 2010; Bennett & Gitomer, 2009) to address the limitations of traditional language assessment, SBA is designed in a way that learners can demonstrate their KSAs in a context that simulates real-life language use. Through the utilization of a sequence of thematically-related tasks along with simulated character interaction, SBA offers opportunities to examine L2 learners' communicative competence in a purposeful, interactive, and contextually meaningful manner.

The purpose of this study is to utilize SBA to measure high-intermediate (CEFR B2) L2 learners' topical knowledge and their L2 KSAs as part of the broadened constructs of L2 communicative competence. To fulfill the scenario goal, learners are required to demonstrate their listening, reading, and writing abilities to build and share knowledge. In addition, learners' prior topical knowledge was measured and their topical learning tracked using the same set of topical knowledge items.

118 adult EFL learners participated in the study. The results showed that the tasks embedded in the SBA served as appropriate measures of high-intermediate learners' communicative competence. The topical knowledge items were found to function appropriately, supporting the use of SBA to measure topical knowledge as part of the broadened constructs of communicative competence. In addition, most learners exhibited substantial topical learning over the course of the SBA, suggesting that with proper contextualization, learning can be facilitated within an assessment. In sum, this study demonstrates the potential value of SBA as an approach to measure complex constructs of communicative language competence in L2 contexts.

French Learners' Use of Sources in an Integrated Writing Assessment Task

Anna Mikhaylova, University of Iowa

3:55 p.m. to 4:25 p.m. Location: Mary Gay C

Traditionally, writing assessment in foreign language (FL) classes is based on independent writing tasks, which require second language (L2) learners to compose essays in response to a short prompt. However, independent writing tasks are often critiqued for being inauthentic in academic contexts and for providing little insight into the full range of L2 learners' writing abilities (Cumming, Kantor, Baba, Erdosy, Eouanzoui, & James, 2005; Weigle, 2004). Concerns with the limitations of independent writing tasks have resulted in increased interest in integrated writing tasks among L2 researchers, test developers, and practitioners (Plakans, 2015; Yu, 2013).

Integrated writing assessment tasks are based on language-rich sources and require L2 learners to search for ideas in the sources, select, synthesize, organize, and connect the ideas, and modify and acknowledge source text borrowing in their own writing (Knoch & Sitajalabhorn, 2013). In general, integrated writing tasks are ubiquitous in education: Students are rarely required to produce academic writing without any references, but are rather expected to summarize, synthesize, and respond to sources in their writing (Cumming, 2014; Hirvela & Du, 2013). Integrated writing tasks are, thus, worthwhile for FL education and should be widely incorporated in FL classes as well.

It is, then, essential to understand how L2 learners approach integrated writing assessment tasks and what their integrated essays look like as a result. However, the overview of the research on integrated writing assessment tasks reveals a number of significant gaps. First, the existing research has been limited to English as a second/foreign language contexts and has yet to expand to include other foreign languages. Second, a limited number of studies have investigated how L2 test takers engage with modalities (e.g., audio and visual sources) other than print sources in integrated writing tasks. This study addresses the identified research gaps by investigating L2 French learners' essays written in response to an integrated writing assessment task that contains print, audio, and visual sources.

This study investigates what language (e.g., copies, quotations, and paraphrases) L2 French learners integrate from the sources into their essays and why. The integrated writing task under investigation is a persuasive essay: L2 French students read a passage, look at a graph, listen to an audio recording, and then write an essay in which they integrate different viewpoints from all the sources and present and defend their own viewpoint on the topic. Data in the study include 38 L2 French students' integrated essays, responses to the retrospective source integration questionnaire, and interviews. This study followed Shi's (2004) and Weigle and Parker's (2012) procedures to analyze the language that the L2 French writers integrated from the sources into their writing and Petrić and Harwood's (2013) procedures to analyze the L2 French writers' motivations for their use of sources. An in-depth qualitative analysis was performed in terms of the L2 French learners' written products and perceptions of the integrated writing assessment task. Finally, the study provides pedagogical implications for student writing in upper-division language courses as well as content courses in literature, culture, and film.

Item-level analyses of a listening for implicature test: Evidence against an implicature subskill construct?

Stephen O'Connell, University of Wisconsin-Madison

3:55 p.m. to 4:25 p.m Location: Henry Oliver

Assessments of second-language listening, whether they are academic in nature or general proficiency, often include conversational implicature as part of the listening construct, alongside other subskills such as listening for main ideas or details. Much research has been conducted on listening subskills (e.g., Goh & Aryadoust, 2010; Kostin, 2004; Sawaki & Nissan, 2009), from which a fairly consistent finding has been that testing inferences increases item difficulty in comparison to non-inference items. There has additionally been research into whether inferencing items show a construct distinct from more explicit listening, although these results have been less clear (Eom, 2008; Wagner, 2004).

What has been lacking in the listening assessment literature, however, is research that tries to isolate assessment of conversational implicature with an instrument designed expressly for that purpose (as Bouton (1988) and Taguchi (2009) did in language learning and acquisition contexts). With this goal in mind, a 60-item listening test was created. Specifications for all items were identical except for the subskill classification of implicature (k=30) and general non-implicature listening (k=30). The items were

developed following standardized testing best practices (i.e., three independent content reviews) and were administered in two formats (multiple choice and constructed response) to 251 learners of English whose proficiency was in the low intermediate to lower advanced range (i.e., B1–C1 on the CEFR).

Rasch analyses showed that while the implicature items were more difficult than the general items in both multiple choice (MC) and constructed response (CR) formats, the differences were not statistically meaningful. Furthermore, logistic and linear regressions on a subset of participants (n=84) whose CEFR level was obtained by an external measure (Michigan Language Assessment's MET) showed that general ability was a better predictor of higher-level proficiency (defined as C1 on the CEFR) than implicature ability.

To try to investigate the question of item subskill more granularly, logistic regressions were conducted on each of the 30 implicature items. To run these analyses, two global scores based on raw scores were generated per participant: their total MC implicature correct and their total MC general correct. These analyses tested whether the probability of a correct answer on each implicature MC response was more predictable from other implicature correct or more predictable from the general MC items sum. If the former, there is potentially evidence of an implicature subskill; if the latter, evidence is provided that the probability of understanding an implicature is influenced more by having higher general proficiency than a separate identifiable implicature subskill. The analyses showed that for only 6 of the 30 implicature items performance on other implicature items was predictive of getting the item correct, compared to 15 cases of 30 where performance on the general items was predictive.

The Rasch analyses and the regressions on CEFR level will be summarized before focusing on the item-level logistic regression results and discussing the implications for the inclusion of conversational implicature items on general or academic proficiency tests of listening.

Differential Item Functioning in GEPT-Kids Listening

Linyu Liao, University of Macau

3:55 p.m. to 4:25 p.m Location: Decatur A

Differential item functioning (DIF) analysis examines whether test items function differentially towards two test taker groups after controlling for the overall ability level of the two groups. Many scholars and test standards have advocated this method for the purpose of detecting construct-irrelevant or biased test items so as to improve test validity and fairness (e.g., Camilli & Shepard, 1994; ITC, 2000; Kunnan, 1997, 2000, 2004; TCTP, 1988, 2005; Xi, 2010). Only when no test groups are favoured or disfavoured because of item bias can test fairness and social justice be preserved. Existing studies mainly focus on DIF effect towards test taker groups classified by native languages, gender, age, and academic majors in tests for mature learners (e.g., Aryadoust, 2012; Aryadoust, Goh & Kim, 2011; Banerjee & Papageorgiou, 2016; Geranpayeh & Kunnan, 2007; Grover & Ercikan, 2017; Oliveri, Lawless, Robin & Bridgeman, 2018). Very few, if any, have examined DIF in tests for children.
The main research question of this study was whether GEPT-Kids Listening has DIF items in terms of Grade (Grade 5 and 6), Gender (male and female), and Region (Mainland China and Taiwan). To address this issue, the current study examined DIF from these three perspectives. Data (i.e., test scores) were collected from 791 test takers in eight Chinese speaking cities on the GEPT-Kids Listening test items, which were all in multiple choice. Two R packages ‘difR’ and ‘difNLR’ (Magis, Beland, Tuerlinckx & De Boeck, 2010; Drabinova & Martinkova, 2017) were used to perform four types of DIF analysis (IRT based Lord's and Raju’s tests, Mantel-Haenszel, and Non-Linear Regression tests) on the test performance data. ShinyItemAnalysis package (Martinková, Drabinová, Leder & Houdek, 2017) was used to draw item characteristic curves (ICCs) in RStudio to visually present which test group was favoured by DIF items.

The Grade, Gender, and Region related DIF analyses altogether flagged 14, 5, and 15 test items respectively, but only four, one, and two of them were flagged by at least three methods. By checking these items flagged by multiple methods through ICCs, it was found that in terms of Grade DIF, there were four uniform DIF items favouring either Grade 5 or Grade 6 test takers. In terms of Gender DIF, the DIF effect could be ignored as it was very minor. In terms of Region DIF, there were only two uniform DIF items favouring Taiwanese test takers. Based on these findings, implications for test development and the use of the test were discussed.

Formative Assessment through Automated Corrective Feedback in Second Language Writing: A Case Study of Criterion

Giang Thi Linh Hoang, The University of Melbourne

4:30 p.m. to 5:00 p.m. Location: Mary Gay C

Automated written corrective feedback systems are increasingly used in classroom settings to provide formative feedback to learners, yet research into the efficacy of this type of feedback is neither ample nor sound in research approaches (Stevenson & Phakiti, 2014). Therefore, well-designed studies to examine learners' engagement and response to automated feedback are needed (Storch, 2018). The current study adopted a pre-post quasi-experimental design to examine the effectiveness of automated corrective feedback as a formative assessment tool for L2 learners' writing development. Seventy-five English majors divided into two groups participated in the study. The students were undertaking a 15-week writing course during which both groups had three practice sessions to compose essays on three writing prompts. The control group wrote their essays on paper and submitted them to the instructor for feedback. The experimental group, however, had access to the automated writing instructional program Criterion during their practice sessions and revised their drafts in response to its feedback before submitting revised drafts to the teacher for feedback.

Beside test essays, other data included first and revised drafts from Criterion practice sessions, recorded think-aloud protocols (TAP) conducted with 14 students as they revised essays using Criterion corrective feedback, follow-up interviews with four students, and end-of-term focus group interviews. Error analysis was conducted on the test scripts of the control and experimental groups to investigate any changes in accuracy over time. Two accuracy measures were used: a holistic measure based on Foster and Wigglesworth's (2016) weighted clause ratio approach, and analytic measures of accuracy in using specific grammatical morphemes following the obligatory occasion analysis.

Findings from the quantitative pre-post test analyses were triangulated with the qualitative data of how the students engaged with the feedback from Criterion and their revision practices. Students' revisions were analysed for the type of changes made in response to Criterion corrective feedback. Analyses of TAP recordings employed Schmidt's (1993) key concept of "noticing" to shed light on students' cognitive engagement with the automated feedbacktoclarify revision practices found. Further triangulation came from students' perceptions of automated feedback in TAP follow-up and focus group interviews which are thematically analysed.

The study found that automated corrective feedback did not facilitate significantly improved accuracy in students' L2 writing. This was substantiated by TAP analyses which revealed Criterion's failure to deeply engage students, resulting in a mixture of correct and incorrect revisions in their revised drafts. Interestingly, several extensive engagement episodes with the feedback led to no revisions, while most correct revisions resulted from perfunctory noticing. However, students generally expressed satisfaction with Criterion, perceiving its corrective feedback as an additional resource rather than replacement of teacher feedback.

The findings have implications for formative feedback practices in the classroom, including how best to use Criterion automated feedback to support L2 writing instruction and classroom-based assessment. Moreover, Criterion corrective feedback should be designed to be more adaptable to focus learners' attention on the relevant issues for their developmental stage.

Content-rich videos in academic L2 listening tests: A validity study

Roman Olegovich Lesnov, Ball State University

4:30 p.m. to 5:00 p.m. Location: Henry Oliver

Standardized L2 listening assessments have been predominantly operationalizing this language skill as visual-free despite abundant counterevidence (Buck, 2001; Kang, Gutierrez Arvizu, Chaipuapae, & Lesnov, 2016). This study has attempted to clarify the nature of the L2 academic listening assessment construct regarding the role of visual information. This goal was achieved by developing a validity argument for including video-based visuals in L2 academic listening tests. Using Kane's validity framework, the explanation inference was of primary concern to this study because it is used to justify the measured construct (Kane, 1992; 2004; 2006; 2013).

The explanation inference was supported by two types of evidence. First, the performances of 143 English as a second language (ESL) and English as a foreign language (EFL) students on an academic English listening comprehension test were quantitatively analyzed for the effect of delivery mode (i.e., audio-only vs video-based) and its relationships with test-takers' listening proficiency (i.e., lower vs higher), item video-dependence (i.e., whether or not an item was cued by video), item type (i.e., local vs global), and viewing behavior (self-reported on a scale from 1-did not watch the video to 5-watched all of the video). Analyses were based on both classical test theory (i.e., ANOVA and correlations) and item response theory (i.e., Rasch analysis). In the video-based version of the test, content-rich videos were used, defined as videos containing relevant graphical content-related visual cues for 60% of the video length.

The findings showed that video-dependent items were easier with videos than without for both lower-level and higher-level test-takers, regardless of item type. Video-independent items were unexpectedly harder with videos in general. In particular, video-independent global items were harder in the video-based mode than in the audio-only mode for lower-level test-takers. Viewing behavior had a weak positive relationship with listening comprehension, regardless of proficiency.

Second, stakeholders' perceptions about using content-rich videos were investigated. Using a questionnaire, the same 143 test-takers provided their perceptions of (1) test difficulty, (2) motivation towards listening, (3) listening authenticity, and (4) whether content-rich videos should be used in high-stakes academic listening tests. The effects of mode and proficiency on these perceptions were examined. Similarly, 310 ESL and EFL teachers provided their opinions about the effects of content-rich videos on the same four perception constructs.

Test-takers found the video-based mode easier than the audio-only mode; however, their perceptions of motivation, authenticity, and using videos in tests were not affected by mode. Teachers were more favorable towards the video-based mode than the audio-only mode in terms of listening difficulty, motivation, authenticity, and using videos in L2 academic listening tests.

The study has discussed how these findings supported the validity argument for including content-rich video-based visual information into the assessment construct of L2 academic listening comprehension. Challenges revealed by the findings were also addressed, with limitations acknowledged. The study offered theoretical and practical implications for the field of L2 assessment. As its primary implication,

the study recommends that test developers start using content-rich visual information in L2 academic listening tests.

Raters' Perceptions and Operationalization of (In)authenticity in Oral Proficiency Tests

John Dylan Burton, British Council

4:30 p.m. to 5:00 p.m. Location: Decatur A

The misuse of high-stakes language tests has far-ranging consequences for a range of stakeholders. While most studies have considered this misuse from the perspective of decision and policy makers, it is also equally essential to consider this issue in terms of test-takers. Most testing boards implicitly assume that test-takers approach language tests in an authentic manner, but as Messick (1982) warned, certain test-preparation practices may lead to skewed inferences about language ability. The few studies that exist in this area (e.g., Lam, 2015; Luk, 2010; Spence-Brown, 2001) have found evidence that construct-irrelevant test-preparation practices may lead to test-takers engaging inauthentically with oral proficiency tests. Nonetheless, to date, little work has considered human detection of this test misuse.

This study aims to address this gap by exploring how raters perceive and operationalize inauthentic discourse. Three main areas were explored: 1) whether raters accurately classify inauthentic speech samples, 2) whether experience rating these types of samples aids in its detection, and 3) the characteristics of inauthenticity that raters find salient. This issue was approached by conducting and recording mock IELTS speaking tests with two groups of Chinese students: one was an unexposed control group and the other was an experimental group which was exposed to a mock speaking test one week prior to taking the test. The recorded samples were then rated by 58 trained IELTS speaking examiners based in five countries using a scale of semantic differentials covering various aspects of authenticity and proficiency. A small subset of these examiners participated in stimulated verbal recall protocols to explore their perceptions and thought processes in more detail.

Statistical analysis revealed that raters overall were able to identify inauthentic engagement in the exposed samples, suggesting that for this group of test-takers, extensive exposure to the test prompts resulted in discourse that was noticeably different from that of the control group. However, only the examiners experienced with Chinese test-takers were able to categorize these consistently with their exposure. The verbal protocols revealed a taxonomy of features that raters associate with both authenticity and inauthenticity. These findings highlight the need for increased test security in order to prevent exposure to task prompts prior to speaking tests and have implications for rater training programs. The findings furthermore have potential applications for the computer detection of construct-irrelevant response strategies.

Individualized Feedback to Raters: Effects on Rating Severity, Inconsistency, and Bias in the Context of Chinese as a Second Language Writing Assessment

Jing Huang, The University of Hong Kong, **Gaowei Chen**, The University of Hong Kong

5:05 p.m. to 5:35 p.m. Location: Mary Gay C

Performance-based language assessment commonly requires human raters to assign scores to language leaners' performances. The subjectivity in human ratings inevitably introduces rater variability that has been identified as a main source of construct-irrelevant variance. Individualized feedback has been used in rater training to limit such variance and increase the reliability of performance- based language assessment. This training method, however, has not yet proved to be effective in improving raters' rating outcomes (i.e., severity, inconsistency, and bias towards a rating scale category). Moreover, few previous studies have investigated the effects of feedback frequency at a given time period on raters' rating outcomes. The present study examined the immediate and retention effects of individualized feedback on raters' rating outcomes, as well as the effects of raters' perceptions of the individualized feedback on rating outcomes. The participants were 93 native Chinese speakers without previous rating experience, and they were randomly assigned to one of three treatment groups. The three groups differed in the way of receiving individualized feedback at a given time period: (a) control group receiving no feedback, (b) single-feedback group receiving the feedback once, and (c) double-feedback group receiving the feedback twice. Each participant rated 100 writing scripts on Day 1 as the pre-feedback ratings, and received one of the feedback treatments on Day 2. The post-feedback ratings were conducted immediately after the feedback session on Day 2 by assigning each participant 100 new writing scripts to rate. This was followed by a questionnaire and interview for the double-feedback and single-feedback groups on the same day. Raters' retention of the feedback was measured by assigning each participant 100 new writing scripts to rate as the delayed post-feedback ratings after one week. Raters' rating outcomes were yielded through multi-faceted Rasch modeling. One-way ANCOVAs, repeated measures ANOVAs, Kruskal-Wallis H tests, and two-way ANCOVAs were run to test the hypotheses. Based on the results of this study, the following main conclusions were made. First, individualized feedback significantly affected raters' rating severity and inconsistency. The rating severities of the double-feedback and single- feedback groups were superior to the control group. In addition, the rating inconsistency of the double-feedback group was superior to the single-feedback group. Second, with regard to the retention effects, individualized feedback was found to be beneficial to raters from the double-feedback group to retain their improvements in rating severity. It also helped raters from the single-feedback group to retain their improvements in rating inconsistency. Third, in terms of the effects of raters' perceptions of the individualized feedback, the results revealed that raters' perceptions of usefulness of the individualized feedback affected rating severity, inconsistency, and bias towards coherence. Furthermore, raters' perceptions of recall of the individualized feedback during subsequent ratings affected rating inconsistency. In addition, raters' perceptions of incorporation of the individualized feedback into subsequent ratings affected rating severity. These findings may shed light on the application of individualized feedback in the designs of face-to-face and online training programs.

Examining the effects of foreign-accented lectures on an academic listening test at the item level using differential item functioning analysis

Sun-Young Shin, Indiana University, **Ryan Lidster**, Indiana University, **Senyung Lee,** Indiana University

5:05 p.m. to 5:35 p.m. Location: Henry Oliver

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 differential performance against some test takers who do not share the first language (L1) with the speaker. Unfortunately, this dilemma for ESL listening test developers and users has not been resolved yet because prior research on the effects of shared L1 on L2 listening test scores has shown conflicting results vis-à-vis the conditions under which a shared-L1 effect takes hold. In addition, most previous research on the effects of non-native accents and shared L1 benefit on listening comprehension test has focused on comparing test takers' performance on overall listening test scores across a group of listeners from different L1 backgrounds (Abeywickrama, 2013; Kang, Moran, & Thomson, 2018; Major et al., 2002; Ockey & French, 2016), whereas an item level analysis is necessary in order to test hypotheses concerning whether foreign-accented speech affects L2 listening comprehension in specific ways.

This study investigated the potential for a shared L1 effect on listening test scores by conducting Differential Item Functioning (DIF) analyses in an effort to understand how accented speech is related to a source of different performance at the item level, while controlling for key variables including listening passages and the degrees of intelligibility and accentedness of speakers. The Mantel-Haenzel (MH) procedure (Holland & Thayer, 1988) was used to detect DIF items because it has been most widely used in the field of language testing and is appropriate for small sample sizes (McNamara & Roever, 2006) generating relatively small Type I errors (Fidalgo, Ferreres, & Muniz, 2004; Muniz, Hambleton, & Xing, 2001). Three native speakers each of English, Hindi, and Mandarin, were selected to represent different varieties of world Englishes. Their relative levels of intelligibility and accentedness of speech were measured in separate transcription and rating tasks. A total of 386 undergraduate and graduate students enrolled in a large Mid-Western university, whose L1 is Mandarin, or Korean, or Hindi, took two listening tests. They first took a retired standardized listening comprehension test comprising only texts recorded by native English speakers, then they took another listening comprehension test consisting of counterbalanced sets of American English-, Hindi-, and Mandarin-accented lectures.

Results show that a shared-L1 effect is minimal and exists only in a few items in which Mandarin or Korean listeners performed relatively poorly when listening to Hindi speakers. In most cases, items exhibiting DIF focused on narrow, detail-oriented answers, and some DIF items were not clearly linked to specific accent features. Since speakers were carefully vetted for being roughly equally intelligible to native listeners, the shared L1 effect was minor, limited to some detail-oriented items, and not wholly related to accent alone. This finding suggests that we can tentatively move towards using more representative speech without unfairly advantaging a particular group of listeners. However, on high-stakes tests, since even small effects could have large consequences, using non-native speech must be evaluated carefully.

Investigating variance sources and score dependability of an ITA speaking test for construct-related validity and fairness: A mixed method G-theory study

Ji-young Shin, Purdue University

5:05 p.m. to 5:35 p.m. Location: Decatur A

Given the complexity of speaking performance tests, examining sources of score variability and dependability is important to validating speaking test use and interpretation (Bachman, Lynch, & Mason, 1995). Generalizability theory (G-theory) provides construct-related evidence, by decomposing variance sources (G-study) and informing score dependability under various conditions (D-study) (Brennan, 2001; Cronbach, Rajaratnam, & Gleser, 1963; Shavelson & Webb, 1991). Previous G-theory studies have investigated various person-, task-, and rater-related sources of variance (e.g., Brown, 1999) but have paid less attention to understanding the variables within and across the source categories simultaneously and/or seeking qualitative evidence from rating and training sessions to support the G-theory results. The comprehensive approach is useful for language proficiency tests for international teaching assistants (ITAs), where raters' and examinees' various backgrounds and diverse tasks interact. The present study systematically investigated the relative effects of variance sources on a local ITA English speaking test scores within and across person-, task-, and rater-related source categories (Step 1) and the score dependability to evaluate the current test and rating design (Step 2), while providing qualitative information from rater comments (Step 3).

The data consist of the 8 item-level scores on an ITA oral proficiency test from 673 examinees at a large, mid-western public university, including examinees' and raters' metadata. First, using the R lmer package, three separate G-studies were conducted, identifying the major variance sources within each category: person-related (person, person L1, sex, academic discipline), rater-related (rater, rater L1, rating confidence, rating experience), and task-related sources (item, skill-integration, prompt-type). Next, 6 best-representative sources that the within-category G-studies identified constructed three confirmatory G-studies across the categories, experimenting different potential interactions that literature informs. The best-fitting model was selected using standard fit indices (e.g., BIC, AIC). Second, the phi (Φ) coefficient, the dependability index for criterion-based testing, was calculated for the scores from the variance sources of the best-fitting model. Also, D-studies with the varied number of tasks and raters were compared to evaluate the current test and rating design. Third, using NVivo, the rater comments from rating and training sessions were analyzed, especially focusing on ratings where the raters disagreed and where raters were of different L1 backgrounds to support the quantitative results.

The within-category G-study results identified 6 best-representative sources: item-type, person, person L1, rater, rater L1, and rating experience. The best-fitting model indicated that person (35.2%) and person L1 (22.6%) are the largest sources of variance to the present ITA test item scores, with a high dependability. Two interaction effects—person L1 and rater L1 (7.4%), and person L1 and rating experience (6.7%)—and rater effect (4.17%) were relatively small with a moderate dependability, D-studies for test design improvement suggested adding third rating to increase dependability, which is the current procedure to deal with rater disagreements. Required rater justifications made during rating and training revealed raters' varied perceptions towards the interaction between ITAs' accent, intelligibility, and English proficiency while highlighting the importance of rater training that explicitly addresses performance profiles of different L1 populations to achieve test fairness.

Analyzing stakeholders' voices in the aviation context: a glocal perspective

Natalia de Andrade Raymundo, University of Campinas/ Brazilian Air Force

3:20 p.m. to 3:50 p.m. Location: Mary Gay C

The International Civil Aviation Organization (ICAO) introduced the Language Proficiency Requirements (LPRs), a language policy that stipulates that pilots and air traffic controllers (ATCOs) must demonstrate a minimum proficiency to be licensed to either fly airplanes or control air traffic on international (cross-border) flights. Despite the high-stakes nature of the assessment tools embedded in this context, validity evidence for the constructs and criteria in relation to the goal of the ICAO LPRs is lacking, especially within the South American context. Although the ICAO rating scale references the ability to communicate effectively on what they call work-related topics, there is little research on the actual target language context of pilot-controller communication, where English is used as a Lingua Franca (ELF) (Douglas, 2014; Kim 2017). Considering Shohamy`s (1998) claim for more democratic models of assessment, this paper aims at bringing into discussion the policy established by ICAO from a glocal perspective (Canagarajah, 2002), utilizing data generated through the ongoing validation of EPLIS_ the Aviation English Proficiency Exam for Brazilian Air Traffic Controllers. This paper reports the two first phases of a broader multistage mixed-methods study. The first phase investigated episodes of non-routine situations in air traffic facilities in Brazil, i.e. situations where standard phraseology was not enough to communicate effectively. I analyzed radiotelephony communications where non-native speakers and English native speaker pilots interacted with Brazilian ATCOs who were assessed as Level 4 (operational) or Level 3 (pre-operational) in EPLIS. In the second phase, aimed at analyzing stakeholders` perspectives, highly experienced Brazilian ATCOs participated in a focus group interview to discuss the findings of the first phase. Results for the two phases suggest that while aviation English proficiency is undoubtedly important, there are two other critical factors determining the success of communication, which are taken for granted in the ICAO`s language policy: professional expertise and experience. Results also questioned ICAO's definition of proficiency in relation to the goal of communicative effectiveness for pilots and air traffic controllers and highlighted the importance of revising the policy of not assessing native speakers` ability to communicate in the international context of aviation.

Positioning students as active learners: An examination of student-generated question quality in literacy assessment

Hyunah Kim, University of Toronto, **Megan Vincett**, University of Toronto, **Samantha Dawn McCormick**, University of Toronto, **Melissa Hunte**, University of Toronto, **Xue Lin**, University of Toronto

3:20 p.m. to 3:50 p.m. Location: Henry Oliver

Traditional language and literacy assessments position students as passive test-takers, expected to answer questions provided. In recent literature, scholars argue that by taking a learning-oriented (Carless, 2007) and learner-centered (Duncan & Buskirk-Cohen, 2011) assessment approach, more meaningful student learning is emphasized. Student-generated questioning (SGQ) has been used in educational settings to immerse students in the educational process and learning (Chin & Brown, 2002). An SGQ approach prompts students to think critically about their learning experience, while advancing their understanding of educational concepts (Chin, 2002). Previous research suggests that SGQ facilitates students' autonomy, interest, and motivation (Taboada, Bianco, & Bowerman, 2012). However, SGQ has not been used as widely in language and literacy education as in content subjects (e.g., science, history). To narrow this gap in the literature, we examined the extent to which SGQ is associated with students' reading comprehension ability, their interest in passage topics, attitude to

reading and writing, and reasoning ability. We further tested the relationships by text type (i.e., narrative and expository).

Approximately 240 grade 5 and 6 students in a metropolitan city in North America participated in this study. The data used in the present study were from 100 students after excluding those who did not complete all assessment tasks. SGQ was elicited as part of reading comprehension assessment in which students were asked to indicate the level of interest in a given passage topic and generate three questions for the passage writer or an expert in that relevant field. Subsequently, students completed multiple-choice comprehension questions associated with each passage. Students' attitude to reading and writing was assessed using a literacy orientation instrument. Finally, their fluid reasoning ability was measured through a matrix reasoning task (Condon & Revelle, 2014). The quality of each SGQ was evaluated using a six-point rubric from 'one' representing the most basic, factual questions (e.g., "Did she survive the swim?") to 'six' indicating higher-level critical thinking (e.g., "Can bats eventually make mosquitoes extinct?").

Multiple regression analyses showed that reading comprehension ($p < .000$) and positive attitude towards writing ($p = .009$) are statistically significant predictors of the SGQ quality, whereas the effects of interest, attitude towards reading, and reasoning ability were not (adjusted R-square = .476). Once data were analyzed by passage, differing results emerged. For the narrative text, the effects of reasoning ($p = .006$) and attitude towards writing ($p = .041$) were statistically significant, while reading comprehension ($p = .035$) was the only statistically significant predictor for the expository text.

These findings highlight important differences in the relationships of the quality of SGQs with the factors investigated, by text type. High-quality SGQs in narrative texts appear to be associated with student reasoning ability and positive orientation to writing, whereas in expository texts, they are associated with comprehension. The findings provide implications for teachers to support students to be more active and critical readers through SGQ. The study concludes with a call for more research to elucidate the use of SGQs in language and literacy assessments across genres.

Enhancing the Interpretability and Usefulness of TEPS Section Scores Through Alignment with CEFR

Heesung Jun, Seoul National University, **Euijin Lim**, Seoul National University, **Yong-Won Lee**, Seoul National University

3:20 p.m. to 3:50 p.m. Location: Decatur A

Messick's (1989) unified definition of validity and Kane's (1992, 2006, 2013) argument-based approach to validation emphasize the evaluative judgement of not only score-based inferences and decisions but also consequences of test use as integral components of assessment validation. In particular, the implication and utilization inferences in the validity argument frameworks call for evidence showing the interpretability and usefulness of test scores for intended test score uses (Chapelle, Enright, & Jamieson, 2008; Kane, 2006). One way to achieve these goals is to set important cut scores indicating different proficiency levels and establish score meanings for each score level in the form of verbal descriptors that can be easily understood by score users. The main goals of the current study are to: (a) conduct a standard-setting study to align the Reading and Listening section scores of a large-scale standardized English proficiency test, TEPS, to the widely-used CEFR scale; and (b) find ways to revise (or augment) the existing verbal descriptors of the 10-score band system for TEPS, in such a way that the meaningfulness and interpretability of the scores are enhanced for the major stakeholders of TEPS.

For the present study, two standard setting methods were used—the Modified Angoff Method, which was used as the main method, and the Bookmark Method, which was used to provide supporting results for external validation of the main method. The panel consisted of 11 and 13 experts in English education and assessment for the listening and reading standard setting sessions, respectively. The 5-stage linking procedure outlined in the Manual for Relating Language Exams to the CEFR (Council of Europe, 2009) was closely followed. The panels made four to five rounds of judgments using both methods. At the end of each session, panelists completed a survey questionnaire on the standard setting procedure and their confidence in the final cut scores.

The two standard setting sessions produced final cut scores for the TEPS Listening Comprehension and Reading Comprehension sections corresponding to A2, B1, B2, C1, and C2 levels in the CEFR. The standard deviations of judgments decreased across rounds, and similar cut score results were derived from the two methods. These findings show that the current standard setting study has both internal and external evidence to support its validity (Tannenbaum & Cho, 2014). Survey results revealed that the panelists were mostly confident about the final cut score recommendations. The procedures of the standard setting sessions were also perceived by the panelists as being clear and helpful for making judgments. Interestingly enough, the panelists reported being more comfortable with the cut scores based on the Modified Angoff Method than those based on the Bookmark Method. In addition, they pointed out the issue of test-takers' guessing as a source of confusion in the process of making judgments using the Modified Angoff Method. Avenues for further investigation, including schemes to deal with various issues arising in standard-setting studies, are discussed in the presentation along with the implications of the study results for the utility, fairness, and consequences of assessment.

Assessing textual sophistication and linguistic complexity in L2 writing

Jianling Liao, Arizona State University

3:20 p.m. to 3:50 p.m. Location: Decatur B

In the current practice of L2 writing assessment, organizational features are often included marginally in the rating scales used to evaluate writing, often expressed as cohesion and coherence (Chen & Baker, 2016). A more effective evaluation of organizational features, however, is crucial not only because appropriate organizational features contribute to writing quality but also because a better knowledge of discourse performance will facilitate an understanding of how L2 writers allocate their cognitive resources and what learners achieve in a specific writing task on both local and higher levels. Nevertheless, limited studies have been devoted to the assessment of discourse features in L2 writing. To acquire a better understanding of these issues, this study investigates the effects of using various measures to evaluate organizational features in L2 writing as well as the relationships between organizational quality and linguistic complexity. Three questions guide the study:
1. What kinds of organizational quality and linguistic complexity exist in beginning, intermediate, and advanced L2 Chinese learners' descriptive essays, respectively? And how do they differ between levels?

2. What are the relationships between the performances of organizational quality and linguistic complexity in beginning, intermediate, and advanced L2 Chinese learners' descriptive writings, respectively?

3. Which organizational and linguistic complexity metrics are the strongest predictors of subjective evaluations of organizational quality and holistic writing quality?

The dataset comprised 75 Chinese descriptive essays on the same topic, produced by 25 beginning, 25 intermediate, and 25 advanced English-speaking college L2 Chinse learners. Organizational quality was evaluated by the usages of local, global, and textual cohesive devices, as well as usages of interactional metadiscourse markers in different categories (Hyland, 2005; McNamara, 2016). The holistic organizational quality was also assessed with a rating scale. Linguistic complexity was evaluated in both lexical and syntactic complexities. Lexical complexity was analyzed with metrics of lexical diversity, density, and sophistication. Syntactic complexity was evaluated with metrics on sentential, T-unit, clausal, as well as phrase levels. The holistic writing quality of the essays was also scored with a rating scale.

The findings show that several of the organizational measures and most of the syntactic complexity measures demonstrated growth patterns. Lexical complexity had only limited growth. While positive correlations were found between some of the linguistic and organizational measures, other measures did not display a parallel developmental pattern. Essays with more effective organizational performance did not necessarily have linguistically more sophisticated features. The lexical complexity and organizational measures predicted the human ratings more effectively than the syntactic measures. Moreover, the textual and linguistic indices that predicted human scoring of organizational and writing quality were in discrepancy with those that showed growth. The findings suggest that discourse knowledge and skills do not seem to be by-products of the development of L2 linguistic proficiency, and they need to be nurtured through language-specific and culture-specific rhetoric training.

The domain expert perspective on workplace readiness: Investigating the standards set on the writing component of an English language proficiency test for health professionals

Simon Davidson, Monash University

3:55 p.m. to 4:25 p.m. Location: Mary Gay C

As part of the prerequisite to obtain professional registration and practice in Australia, International Medical Graduates (IMGs) need to demonstrate satisfactory English language proficiency. Concerns have been raised that the specified minimum level on tests used for this purpose (including the Occupational English Test (OET), a specific-purpose language (LSP) test for health professionals), might be inadequate for successful workplace functioning. To better understand the validity of these concerns, a study was conducted to review minimum standards on the OET via the process of 'standard setting' – a procedure for drawing insights from appropriate stakeholders (in this case health professionals with experience of workplace communication demands) about levels of proficiency viewed as satisfactory for a particular purpose. The study sought to determine the minimum levels of competence deemed appropriate for effective performance in the workplace, and also to understand the basis for the decisions made and how closely these corresponded to the construct of communication that the OET is designed to measure. A previous study (Manias & McNamara, 2016; Pill & McNamara, 2016) explored these issues in relation to the OET speaking subtest, whereas the current study focuses on writing – a thus far neglected area.

The writing task on the OET is a letter of referral, based on a set of provided case notes. 18 doctor participants (all with experience of working as medical educators, GPs or specialists) were recruited to participate in standard-setting workshops designed to elicit decisions about what level of performance on this task was deserving of a passing grade and why. Five OET writing samples representing different performance levels were used in the workshops and an additional 30 take-home samples were issued.

To gain further insight into the basis for the standards set, verbal reports in the form of a think-aloud protocol (TAP) were employed with five of the 18 workshop participants. These participants were separately asked to say 'out loud' what they noticed while assessing the competence of 12 further samples. A FACETS analysis (Linacre, 2017) was used to calculate new passing standards and compared with current OET cut-scores. In addition, the doctors' comments were thematically coded and intercoder reliability checks were conducted.

The quantitative analysis yielded a slightly more stringent passing standard than the current one. The qualitative findings however showed that some decisions were influenced by perceptions of candidates' clinical competence, extending beyond the construct of communicative competence as defined by the OET. The stricter passing standard set by the domain experts could be interpreted as supporting anecdotal evidence that the current cut-score is too low and that some IMGs are entering Australian workplaces without satisfactory communication skills. However, the qualitative findings indicate a possible misapprehension by some domain experts about what the OET is designed to measure, raising questions about whether they are equipped to judge language proficiency independently of other professional skills. The validity implications of these findings for the OET, and for LSP testing more generally, are considered

Do test accessibility features have the intended effect for K-12 English learners?

Ahyoung Alicia Kim, WIDA, University of Wisconsin-Madison, **Meltem Yumsek**, University of North Carolina at Greensboro, **Mark Chapman**, WIDA, University of Wisconsin-Madison, **H. Gary Cook**, WIDA, University of Wisconsin-Madison

3:55 p.m. to 4:25 p.m. Location: Henry Oliver

English learners (ELs) with disabilities face a variety of challenges to meaningfully participate in educational institutions, with over 10% of the Kindergarten to Grade 12 (K-12) ELs in the US being identified with one or more disabilities (USDE, 2014-15). The Every Student Succeeds Act entitles ELs with disabilities to one or more accommodations when taking English language proficiency (ELP) assessments. Test designs should allow ELs with disabilities to meaningfully participate in the assessments, so they may demonstrate their ELP in a manner consistent with their peers and assessments should employ features of universal design that improve accessibility for all test-takers, such as a text highlighter and magnifier. These accessibility features, also known as universal tools, are designed to provide the necessary support for the general EL population, including ELs with and without disabilities (Willner & Monroe, 2016). It is important to understand how ELs with and without disabilities use accessibility features, as both groups are student populations with special needs. However, no known studies have investigated this topic in the ELP assessment context, suggesting the need for research.

This study examined the use of online accessibility features of an ELP assessment by Grades 1-12 ELs with and without disabilities. The test studied is ACCESS for ELLs (hereafter ACCESS), an annual ELP assessment administered across 39 US states and territories and measures the four language domains of listening, speaking, reading, and writing. Test scores are used for making high-stakes decisions about ELs (e.g., placement, reclassification). Approximately 1.3 million ELs' test telemetry data (i.e., records of test-takers' online interactions during the test) were analyzed and ELs with disabilities comprised 11% of these data. The study focus was on ELs' use of several accessibility features, embedded in the online ACCESS platform: Colored Overlays, Color Contrast, Help Tools, Line Guide, Highlighter, Magnifier, and Sticky Notes. To explore the degree to which ELs use the accessibility features, descriptive and frequency analyses of the telemetry data were conducted for each feature across all ELs, and ELs with

and without disabilities. In comparing the use between ELs with and without disabilities, Mann-Whitney U tests were conducted due to non-normal distributions. Effect sizes were reported, in addition to the significance of group differences.

Findings show that ELs as a whole generally used the Line guide, Highlighter and Magnifier more frequently than other accessibility features, but used the Help Tools the least. Use of accessibility features was more common in the selected response listening and reading domains, which were administered prior to the constructed response speaking and writing sections. The comparison between ELs with and without disabilities revealed that higher percentages of ELs with disabilities activated the accessibility features across all domains, than ELs without disabilities. Although the difference in the use of some features between the two groups was statistically significant, effect sizes were small. The findings indicate the usefulness of some accessibility features that are embedded in online tests, suggesting that such features may provide the intended supports for special populations of students, including ELs with disabilities.

High-stakes tests can improve learning - Reality or wishful thinking?

Jessica Wu, The Language Training & Testing Center (LTTC), **Judy Lo,** The Language Training & Testing Center (LTTC), **Anita Chun-Wen Lin**, The Language Training & Testing Center (LTTC)

3:55 p.m. to 4:25 p.m. Location: Decatur A

Having been influenced by the trend of globalization, education across levels in Taiwan requires students to demonstrate their English ability at a specified Common European Framework of Reference (CEFR) level through taking a standardized English language test (e.g., GEPT, IELTS, TOEFL iBT, TOEIC). This practice has been increasingly criticized for failing to achieve its intended goals of enhancing students' English language proficiency, given the fact that these tests are mostly used for summative purposes rather than formative ones. In students' minds, tests are used to judge how good one is rather than to help one to learn better. Therefore, there is an urgent need to alter such mindset resulted from using standardized tests as a means to improve learning.

This study sought to improve the use of the GEPT in facilitating learning by reforming the current score reporting practices. For this study, the information provided in the score report was expanded to include not only test scores but also diagnostic feedback for GEPT test-takers based on the principles of Learning-Oriented Assessment (Jones & Saville, 2016). We believe that providing learners with detailed and personalized analysis of their test performance can help them understand what their strengths and weaknesses are (feed-back) and what they can do to improve their learning more effectively (feed-forward).

To provide effective feedback to test-takers, detailed can-do statements underlying the ability measured in the test were first developed and each test item was then tagged with its corresponding can-do statement by a panel composed of five researchers who are familiar with the GEPT specifications. To investigate the effectiveness of the new score reporting practices, around 1,000 learners were invited to take the GEPT listening and reading test in June 2018. Besides test scores, they received diagnostic feedback including: 1) a general description of the ability underlying the scores attained, 2) an analysis of individual strengths and weaknesses in listening and reading ability, and 3) personalized learning suggestions. In addition, an online questionnaire was administered to investigate test-takers' perception of the new score reporting service. Among them, about 5% of the test-takers were selected for a follow-up interview. Questionnaire responses and interviews including both quantitative and qualitative data were analyzed.

Results showed that in general the test-takers were highly satisfied with the new score reporting. More than 95% of the questionnaire respondents reported that the diagnostic feedback increased their understanding of their ability and was useful for their future learning. The findings suggest that with the new score reporting service, the GEPT can be used to improve learning. Therefore, we call for a paradigm shift from test orientation to learning orientation among learners, teachers, and testers in view of the consequences of using high-stakes standardized tests to improve learning.

Linguistic Tools in Writing Assessment: Their Impact on Test-takers' Writing Process and Performance

Saerhim Oh, Pearson

3:55 p.m. to 4:25 p.m. Location: Decatur B

Linguistic tools, such as spelling, grammar, and reference tools, have become an integral part of many L2 writer's writing process. In order to confirm that their spelling, grammar, and chosen words and/or phrases are accurate, L2 writers regularly refer to spell checks, grammar checks, dictionaries, and thesauruses. This is especially the case in academic contexts, in which having the ability to write in order to communicate with others in a digital form while making use of available linguistic tools is crucial. Accordingly, it is important to understand how these writing behaviors in real life can be simulated in an assessment setting if we are aiming to generalize test-takers' writing performance from these assessments to their real-life writing ability (East, 2008; Weigle, 2002).

To achieve this overarching goal of understanding test-takers' use of linguistic tools and its effects on their scores, the current study investigated 120 adult L2 test-takers' use of linguistic tools (i.e., spelling, grammar, and reference tools) in an academic English writing assessment setting. Three highly contextualized tasks which reflect the tasks L2 learners may encounter in an academic context were used, and the responses were scored for lexical form and meaning, morphosyntactic form and meaning, cohesive form and meaning, topical meaning, functional meaning, and implied meaning produced (Purpura, 2017).

The scores were analyzed using classical test theory and many-facet Rasch measurement to understand the extent to which the test-takers' writing performance differs across their language proficiency as a function of having access to different linguistic tools and to investigate the systematic interactions between the assessment conditions (i.e., have access to no linguistic tools, spelling tool, grammar tool, or reference tool) and 1) the components of the rubric, 2) the test tasks, and 3) the proficiency levels. To better explain the findings from the quantitative analyses, screen recordings of the process of completing the tasks with access to a certain linguistic tool were analyzed based on a coding scheme.

Through the comparison of the writing scores across the proficiency level, it was found that for all three tasks, the intermediate and the proficient test-takers were consistently distinct in their scores in all assessment conditions. However, in distinguishing the intermediate from the advanced test-takers and the advanced from the proficient test-takers, having access to the spelling tool or the reference tool proved helpful. Additionally, with access to the spelling tool or the reference tool, test-takers systematically received a higher score on the components of the rubric that focus on form and meaning than on those that focus only on meaning, but with access to the grammar tool, contradictory results were found. It was also found that the group of test-takers with access to the reference tool received a higher score than expected for the task that involved writing more descriptive language compared to the other tasks. These findings are elaborated by discussing what was observed from the qualitative analyses. The presentation will conclude by discussing the implications of the use of linguistic tools in language assessment.

How valid are language tests used in the overseas-trained nurse registration processes?

Ute Knoch, University of Melbourne, **Sally O'Hagan**, University of Melbourne

4:30 p.m. to 5:00 p.m. Location: Mary Gay C

Language tests for specific purposes, in particular occupation-specific language tests, are designed to represent the target language use domain in such a way as to allow inferences about the language ability of a test taker in a specific work context. In the case of internationally qualified nurses (IQNs), registration processes in a number of English-speaking countries require evidence of language proficiency in the form of either academic English tests (e.g. IELTS or TOEFL) or specific-purpose tests such as the Occupational English Test (OET). Despite rigorous standard-setting, there is evidence that nurses who have fulfilled the language requirements and have entered the workplace are struggling with certain aspects of workplace communication (Crawford, 2013; Cummings, 2009; Chege & Garon, 2010).

The aim of this study was to determine whether the difficulties displayed by IQNs admitted to the profession might be attributable to the failure of the relevant screening test to capture the communication demands that nurses encounter in the workplace.

To explore this possibility, we compared the test content of those tests currently used in a number of English-speaking countries (UK, Australia, New Zealand, US, Canada) to what emerged from the literature on nursing communication, in particular from those studies detailing areas of difficulties for IQNs. The literature review drew on Whittemore and Knafl (2005)'s method and stages of integrative review. In the first stage, we conducted a data base search aimed at identifying papers on nursing communication in general and in particular on communication problems encountered by IQNs. After screening out irrelevant articles, the content of the remaining papers was coded according to key themes and sub-themes. We then conducted a mapping exercise comparing the areas emerging from the review with the public test specifications of the tests currently used for language screening of IQNs in key English-speaking countries.

The literature review findings revealed key areas in which IQNs are shown to particularly struggle when working in English-speaking workplaces. These included (1) reception and production of spoken features, (2) non-verbal communication (in particular the lack of this during telephone conversations), (3) medical language (both understanding and producing medical terms and explaining these to patients in accessible language), (4) social cultural communication (understanding cultural references, idioms, euphemisms and conducting small talk with colleagues and patients), and (5) using language to enact the local mode of patient care. This last theme included a number of problems displayed by IQNs, for example strategies for understanding key aspects of patient-centred care such as active listening and pausing as well as showing assertiveness in the workplace, speaking up on group settings and delegating when required.

The findings revealed important discrepancies between the content coverage of tests currently used for IQN screening purposes and workplace communication demands, pointing to substantial construct under-representation. This has implications for the validity of inferences drawn about nurses entering the workplace, for test design and for the formulation of testing policies in the health workforce registration process more generally.

Empowering K-12 Teachers to Make Better Use of High-Stakes Summative ELP Assessments

Alexis Lopez, Educational Testing Service

4:30 p.m. to 5:00 p.m. Location: Henry Oliver

As part of the latest re-authorization of the Elementary and Secondary Education Act, known as the Every Student Succeeds Act (ESSA, 2015), states that receive federal funds for the education of English learners (ELs) must administer a standards-based summative English language proficiency (ELP) assessment annually to every student with the EL designation. The assessment needs to be tied to state English Language Development (ELD) Standards that "define progressive levels of competence in the acquisition of the English language" (U.S. Department of Education, 2016, p. 16). These high-stakes summative ELP assessments are utilized for significant accountability purposes. For example, the scores are reported to the federal government to determine state funding; employed at the state and district levels to evaluate the quality and future funding of bilingual/ESL programs; and used at the district level to determine individual student placement in bilingual education/ESL services and all-English classrooms (Wolf, Guzman-Orth, & Hauck, 2014).

However, there is relatively little research on how U.S. K-12 teachers use these summative ELP assessments and their associated resources to shape their instructional and assessment practices. Thus, a study was conducted to examine how classroom teachers from three different states use a summative ELP assessment to make decisions about their students' academic language and to facilitate their language development. Semi-structured interviews were conducted with 18 ESL teachers in grades K-12 in the United States using a web-based video conferencing platform. Interview questions targeted participants' educational context (students, instruction, assessment), how the content of the ELP assessments (e.g., skills being tested, format, item types) and their associated resources (e.g., score report, ELP standards, can do descriptors, rubrics) impact their instructional and assessment practices, and how they use the scores on the summative ELP assessments to make instructional decisions. Interviews lasted approximately one hour and were audio recorded and transcribed. Interview transcripts were analyzed qualitatively using open-ended coding to identify recurring patterns in the way the teachers used the summative ELP assessment and the type of instructional decisions they make.

In this presentation, I will summarize the study findings and will also discuss the challenges and opportunities involved in teachers' use of summative ELP assessments. The findings indicate that the scores on the ELP assessment do not have a great impact on the teachers' decision-making process. It was also found that both the format of the ELP assessment and some of their associated resources practices (e.g., ELD standards, can do descriptors, speaking and writing rubrics) shape the teachers' classroom practices. Finally, it was found that teachers lack the power to make important instructional decisions such as making changes to the curriculum, changing the timing of the tests, and changing the information in the score reports. These decisions could potentially improve their instructional practices. I will conclude with recommendations to improve the relationship between summative ELP assessments and classroom practices.

Source Use Behavior and Raters' Judgement in L2 Academic Writing

Pakize Uludag, Concordia University, **Heike Neumann**, Concordia University , **Kim McDonough**, Concordia University

4:30 p.m. to 5:00 p.m. Location: Decatur A

The integration of reading into writing tasks has become a common practice in both large-scale L2 proficiency tests and classroom-based assessment tools (Cumming, 2013). Researchers have investigated the impact of test takers' interaction with the source text on their writing performance in large-scale integrated writing assessment tasks in terms of linguistic (Guo, Crossley & McNamara, 2013), lexical (Gebril & Plakans, 2016), and organizational features (Plakans & Gebril, 2017). However, few have examined the relationship between source-text use behavior and scores on a classroom-based, integrated writing assignment. Therefore, the present research explores L2 writers' source-text use behavior as a predictor of holistic scores on a classroom-based, reading-to-write assignment.

Cause-and-effect essays (N = 51) written by English L2 writers in an English for Academic Purposes course at a university in Montreal were collected and rated by two independent raters using a holistic rating scale. Adopting coding frameworks from previous research (Shi, 2004; Gebril & Plakans, 2013), source-use behavior was analyzed in terms of instances of direct copying (quotations vs. plagiarized text), lexical revision (words added, subtracted, substituted, or word form changes), syntactic revision (minimal or substantial sentence structure changes) and substantial revision (paraphrases with lexical and syntactic modifications). The writers' purposes for including source information was also coded following Wette (2017) as either being used to support an idea or claim (providing an example or fact) or creating content (defining a concept or idea).

Correlation analyses were performed first to identify which source-use behaviors were related to the holistic scores. Variables that reached the benchmark for a small correlation coefficient in applied linguistics research (±.25, Plonsky & Oswald, 2014) were entered into a multiple regression model to determine whether they predict holistic scores. The findings will be presented, and implications for teaching L2 integrated writing and assessing source-based writing ability will be discussed.

Unpacking the textual features, vocabulary use, and source integration in integrated listening-to-write assessments for adolescent English language learners

Renka Ohta, University of Iowa, **Jui-Teng Liao**, Kirkwood Community College

4:30 p.m. to 5:00 p.m. Location: Decatur B

One goal for English language learners (ELLs) is to develop balanced linguistic skills across all language domains as real-life communication often requires ELLs to use multiple language skills simultaneously. For instance, reading and writing skills are often integrated into academic activities, and both listening and speaking skills are necessary for everyday interpersonal communication. Integrated writing tasks have increasingly gained popularity in standardized, institutional-, and classroom-based assessments for adult ELLs. Previous research has found that successful test takers demonstrate the balanced use of multiple skills (Plakans & Gebril, 2017), and that test takers actively interact with input materials, in a process called discourse synthesis, to produce their written texts (e.g., Ascención, 2005, Plakans, 2009, Spivey, 1990, 1997). However, this type of task has rarely been introduced to adolescent ELLs, potentially due to its multifaceted nature of language demand. In addition, these tasks require teachers and test developers to prepare input materials suitable for this population of learners. For those test takers, it is important to provide a scaffolding material during the test. To date, it is still unknown how adolescent ELLs tackle these challenging tasks and how test results can meaningfully be interpreted by

stakeholders. This study addressed these issues by investigating the relationship between textual features and writing performance, as well as how scaffolding materials (i.e., vocabulary support) affected test performance.

The purpose of this empirical research was to examine textual features derived from listening-to-write tasks completed by 198 high school English as a foreign language (EFL) learners. The students were required to listen to academic lectures and respond to writing prompts. Students were provided with a list of vocabulary to aid their listening comprehension. Students' essays were holistically scored, which were then categorized into three levels of writing proficiency, and analyzed for discourse features, including syntactic complexity, lexical sophistication, fluency, and grammatical accuracy. To understand the impact of vocabulary support on writing performance, we examined the ratio of the borrowed words from the list and rated the overall success of the use of borrowed words. We also scrutinized the effect of successful synthesis of input materials in their performance. Multiple regression analysis was conducted to investigate how well these discourse features predicted students' performance on the listening-to-write tasks.

Results showed that fluency and grammatical accuracy were significant predictors of students' listening-to-write scores. Moreover, the successful use of vocabulary from the list was a stronger predictor of writing scores than the ratio of borrowed words. In short, the ability to effectively use the given vocabulary words was important to receive higher scores. Qualitative results indicated that higher-scoring students tended to reiterate all the lecture's key points and supporting details in their essays than their lower-scoring counterparts. This study offers validity evidence for listening-to-write scores as text features, the use of given vocabulary, and source integration affected ELLs' performance. The presentation concludes with implications for stakeholders to adopt integrated writing assessments along with vocabulary support.

Assessing clinical communication on the Occupational English Test: The intersection of cognitive and consequential validity

Brigita Séguis, Cambridge Assessment English, **Barbara Ying Zhang**, Cambridge Boxhill Language Assessment, **Gad S. Lim**, Michigan Language Assessment

5:05 p.m. to 5:35 p.m. Location: Mary Gay C

In her account of critical language testing Shohamy (2014) calls for a shared responsibility between the tester and test stakeholders in collecting data about the quality of the instruments and their uses. She further explains that in making decisions about the instrument quality, consideration must be given to several aspects of validity, which are related not only to accuracy standards, but also social consequences. The purpose of the present paper is to investigate how these complementary aspects of validity were addressed in relation to the revised speaking component of the Occupational English Test (OET), a specific-purpose English language test for the healthcare context.

Following stakeholder concerns that the OET speaking test was not capturing important aspects of health professional-patient interaction, clinical communication criteria were developed in consultation with healthcare professionals, thus achieving a closer alignment between what was assessed on the test, and the communicative skills valued by those in the field (Elder, et al., 2013). This paper focuses on the subsequent steps in the process of validating clinical communication criteria and pursues two interrelated aims: 1) to investigate the cognitive validity to ensure that the cognitive processes elicited from test takers correspond to the processes they normally employ in the real-life context (Weir 2005);

2) to analyse the consequential validity, i.e. the social consequences by focusing on test takers' perspective.

The data for the study come from recordings of test performances, featuring general medicine candidates from one of the first administrations of the revised test. Recordings were transcribed, coded and qualitatively analysed using the Conversation Analytic approach. In addition, follow-up interviews and survey responses were collected, which were then analysed quantitatively and qualitatively.

The discourse analysis provides ample evidence of the candidates producing linguistic structures that constitute verbal manifestations of their thought processes and can be deemed representative of similar cognitive behaviours in a real-world setting, e.g. reacting to patients' cues, showing empathy, or checking understanding. Survey responses and interviews indicate that the majority of candidates had some familiarity with clinical communication but wanted to develop these skills further.

One of the important findings in terms of instrument quality and its social consequences that emerges from the study is that preparing for and taking the test gives candidates the opportunity to revisit how clinical communication is performed within the patient-centred approach, which encourages development of skills that will benefit them beyond the immediate test context. The study also provides evidence of how a test developed in consultation with key stakeholders can benefit the social sites it interconnects in terms of the impact on the learning opportunities, and the impact such learning carries into healthcare workplaces.

Strategies Used by Young English Learners in an Assessment Context

Lin Gu, Educational Testing Service, **Youngsoon So**, Seoul National University

5:05 p.m. to 5:35 p.m. Location: Henry Oliver

The importance of collecting evidence based on response processes was highlighted in Purpura (2014), as such evidence not only provides support for purposed score interpretation and use but also has the potential for informing construct definition and for influencing test design. Cohen and his associates (e.g., Cohen 2012, Cohen & Upton, 2006) also argued that understanding the strategies test takers use when interacting with test items illustrates whether or not the processes elicited by test items are construct-relevant, and therefore would allow us to evaluate the appropriateness of the inferences made based on the test performance. This line of investigation becomes especially crucial when designing tests for populations whose assessment behaviors are not well understood. Few studies have examined the strategies used by young English language learners in an assessment context. As a result, knowledge of the response processes engaged in by young learners is limited.

To address this lack of information, in this presentation we report on an empirical study that examined strategies used by young learners of different proficiency levels when taking TOEFL Primary listening and reading comprehension items. In particular, we wanted to understand whether their reported strategies are relevant or irrelevant to the assessment construct and to what extent strategy use relates to their proficiency level.

Sixteen Chinese speakers, aged 6-11, took the TOEFL Primary test and participated in a one-on-one cognitive interview. Following Cohen's framework (Cohen, 2012; Cohen & Upton, 2006) we classified the reported strategies into three categories: language learner (LL) strategies, test management (TM) strategies, and test-wiseness (TW) strategies. Both LL and TM strategies strategies were considered to be construct-relevant, whereas TW strategies were considered to be construct-irrelevant. We identified 12 LL strategies, 11 TM strategies, and 8 TW strategies, based on a total of 308 instances of strategy use

across listening and reading items. LL strategies were the most frequently used, accounting for close to 40% of all strategy use, followed by TW strategies (37%) and TM strategies (23%).

Our findings indicate that the majority of the strategies used (about 63%) were construct-relevant. We found that young learners could capably employ a wide variety of strategies, and that they were capable of deploying strategies beyond word or phrase level. We also found that there is a general alignment between construct-relevant strategy use and performance level. Low- and mid-performing learners used TW strategies more frequently than the high-performing group. In contrast, learners in the high-performing groups reported rare use of TW strategies.

The study demonstrates that investigating strategy use can help to inform construct definition. Practical implications on design assessment items for young learners will be discussed.

Cohen, A. D. (2012). Test-taking strategies and task design. In G. Fulcher & F. Davidson (Eds.), The Routledge handbook of language testing (pp. 262-277). London and New York: Routledge.

Cohen, A. D. & Upton, T. A. (2006). Strategies in Responding to the New TOEFL Reading Tasks. TOEFL Monograph Report No. MS-33. Princeton, NJ: Educational Testing Service.

Purpura, J. E. (2014). Cognition and Language Assessment. The Companion to Language Assessment.

Writing Assessment Training Impact and Mexican EFL University Teachers: A Proposed Categorization

Elsa Fernanda Gonzalez, Universidad Autonoma de Tamaulipas

5:05 p.m. to 5:35 p.m. Location: Decatur A

Assessment is a task that language teachers are required to conduct on a regular basis in their classrooms. In the Mexican English as a Foreign Language (EFL) context, language instructors need to select an assessment method that corresponds to their assessment purpose. In their regular assessment practice, they develop the assessment tool to use in the classroom, administer the tool, score students´ performance, interpret the score, make appropriate decisions, communicate the results to administrative offices and finally be aware of the consequences that assessment decisions may bring (Crusan, 2014; Fulcher, 2012; Stoynoff and Coomb, 2012; Weigle, 2007). When these high-demanding activities are combined with the nature of the assessment of EFL writing and the subjectivity entailed in this process, instructors may find themselves in a difficult situation. They may not have the theoretical knowledge or practical experience to assess writing in their classrooms. In the Mexican EFL context, teachers have expressed their lack of writing assessment training and their insecurity when conducting this specific type of assessment in their classrooms. Experts have come to suggest that providing teachers with assessment knowledge and practice, in other words providing teachers with tools to become assessment literate, may allow them to provide students with more reliable writing assessment and become more confident in their regular assessment practices. However, the benefit of writing assessment training and its actual impact on classroom assessment of writing has remained underexplored.

In an attempt to address this gap, the present study had the purpose of analyzing the impact that two sessions of writing assessment training had on eleven Mexican EFL university teachers who were tracked for a period of 12 months. The instrumentation included a participant background questionnaire, teacher focus group sessions, and two semi-structured interviews to the eleven teachers. Data obtained suggested that training had positive impact in three main areas: a) classroom teaching of writing, b) classroom assessment of writing and c) teachers´ self-awareness of assessment. However, the impact of

the training on teachers´ actual classroom assessment was quite shallow. Only two teachers reported to have changed their actual assessment processes and scoring tools after experiencing training while two more described they had managed to increase the amount of classroom activities dedicated to writing assessment. Greater impact was found in teachers' reflection processes and self-awareness of themselves as EFL teachers and assessors. All the participants reported to have become aware of their need to improve their assessment of the skill and above all to have reflected on their teaching of writing. Others stated to have reconsidered the importance that writing should have in their assessment process and for students´ language development. The Writing Assessment Training Impact Categorization is proposed, as a product of the results obtained, in an attempt to classify the innovations training may have encouraged in EFL teachers. The presentation finalizes with a discussion of possible research implications for EFL classroom assessment, teacher trainers and language program managers.

Japanese university students' paraphrasing strategies in L2 summary writing

Yasuyo Sawaki, Waseda University, **Yutaka Ishii**, Waseda University, **Hiroaki Yamada**, Tokyo Institute of Technology

5:05 p.m. to 5:35 p.m. Location: Decatur B

Writing a summary is a complex task involving various types of linguistic operations. Among them is to paraphrase the information in the source text into the summary writer's own words (Brown & Day, 1983; Cordero-Ponce, 2000; Hare & Borchardt, 1984; Winograd, 1984). Appropriate paraphrasing is particularly important for L2 learners in academic programs because they are expected to produce summaries in various forms, while avoiding plagiarism, in their course work. However, writing summaries with appropriate paraphrasing is not necessarily an easy task. For instance, a previous analysis of 99 summaries written in English by Japanese undergraduate English majors conducted as part of this project identified various types of suboptimal paraphrasing strategies, including verbatim copies of three-word strings or longer in approximately 16% of idea units that were coded as correct replications of the source text content across different summary content score levels. These results suggest the importance of understanding characteristics of EFL learners' paraphrasing behaviors and providing learners across levels with focused training on appropriate paraphrasing strategies in academic literacy training. The present study examines the relationship between summary content score level and paraphrasing strategy use for Japanese university students. Participants were English majors enrolled in an L2 academic writing course at a private university in Tokyo, who produced 80-word summaries of an expository text written in English in class. Paraphrasing strategies observed in the summaries analyzed in the above-mentioned study as well as additional summaries written by another cohort of students on the same text (175 summaries in total) will be analyzed by combining two types of text analyses based on previous paraphrasing strategy studies (Burstein, Flor, Tetrault, Madnani, & Holtzman, 2012; Keck, 2006, 2014). First, in order to identify instances of verbatim copying, learner-produced summaries will be annotated for exact string matches by conducting an automated analysis. Second, Keck's (2014) taxonomy for analyzing different types and degrees of paraphrases will be employed for two independent coders to categorize idea units in the learner-produced summaries into four paraphrasing types (Near Copies, Minimal Revisions, Moderate Revisions, and Substantial Revisions) with 20% double coding. Frequency data obtained from both analyses above will be tallied by level on a four-point summary content rating scale developed as part of this study to assess the accuracy and succinctness of source text content representation in learner-produced summaries. Chi-square analyses will be conducted to compare the above frequency data across the score levels. Finally, summary protocols identified as representing high degrees of verbatim copying and suboptimal

paraphrasing behaviors will be examined qualitatively to identify the context in which such behaviors occur within and across the levels. In this presentation, key results of the present study will be discussed along with their implications for designing performance feedback and instructional modules for paraphrasing strategy training for EFL learners in academic contexts.

Investigating raters' scoring processes and strategies in paired speaking assessment

Soo Jung Youn, Northern Arizona University, **Shi Chen**, Northern Arizona University

8:30 a.m. to 9:00 a.m. Location: Mary Gay C

Assessing pragmatic interaction and interactional competence using paired speaking tests has been increasingly common in the field of second language (L2) assessment (e.g., Plough, Banerjee, & Iwashita, 2018). One area that deserves further attention is an in-depth investigation of what strategies raters apply during scoring processes (May, 2009). Inconsistent raters' performances are irrelevant to the construct being assessed and, thus, threaten the validity of assessment. Previous quantitative research indicates that raters tended to show more variations (i.e., unreliable scoring) toward lower-level learners' paired speaking performances (Author, 2018). Nonetheless, as shown in the second language acquisition literature (Hall, Hellermann, & Doehler, 2011), lower-level learners are able to accomplish social actions in talk-in-interaction using various interactional resources, but ways in which such performances are scored and perceived by raters for assessment purposes are relatively unknown. Qualitative research on this research issue is still scarce and research into rater cognition can allow the examination of how raters make valid and reliable scoring decisions. In order to address these research gaps, this qualitative study reports 10 experienced raters' scoring strategies and processes involved when scoring six lower-level learners' performances that are culled from the database of 102 test-takers' paired speaking performances that measure L2 pragmatic interaction (Author, 2015). This qualitative study focuses on the sources of rater variations and different scoring strategies and processes. 10 experienced ESL teachers individually completed think-aloud sessions, which consisted of various phases, such as initial impression of paired speaking performances, training raters with rating criteria, and think aloud decision-making processes. Their audio-recorded data were transcribed for content and thematic analysis. The findings indicated that the raters' decisions go beyond understanding and applying the descriptions of the rating criteria, but influenced by distinct psychological burden associated with the process of rating. In addition, discrepancies were found between the factors that influenced scoring decisions during the impressionistic judgment and post-rater-training. For example, raters' initial scoring decisions were often influenced by either test-takers' accents or the extent to which the communicative goal was established or not, compared to their post-rater-training scoring decisions that reflect the detailed characteristics of L2 pragmatic interaction. Distinct patterns of paired speaking interaction stemming from each test-taker's characteristics (e.g., dominance, proficiency level) also influenced the raters' scoring decisions. Based on the findings, areas for improving the rating criteria for paired speaking performance based on the in-depth analysis of raters' scoring decision processes and strategies will be discussed. The findings have various implications for (a) broadening the construct of L2 speaking based on unique features of L2 pragmatic interaction and (b) addressing practical challenges that teachers face with scoring paired speaking performances.

Multilingual Assessment Reflecting Multilingual Educational Policy: Toward Assessment for Justice

Elana Goldeberg Shohamy, Tel Aviv University, **Michal Tannenbaum**, Tel Aviv University , **Anna Gani**, Tel Aviv University

8:30 a.m. to 9:00 a.m. Location: Henry Oliver

The lack of congruence between language assessment and language educational policies is a common phenomenon worldwide, whereby standardized national and international tests are often used to assess language proficiency as well as content knowledge, and in most cases do not reflect school teaching and learning practices nor do they address the full language repertoire of students. This conflicting situation can be viewed as unjust since students are tested on things they have never learned, and even more so since often the tests become the de-facto policies that students and teachers follow and the actual language policy is being neglected. One major component in the development of a new multilingual educational policy in Israel, as part of a recent proposal of the Israeli Ministry of Education, is to promote multilingual assessment, which relies also on the languages the students already know, i.e., their full language repertoire.

The research reported here is based on several previous initiatives (Schissel, 2018, de Backer, 2018; Shohamy, 2011) where it was demonstrated that students who are proficient in another language (usually their L1) can perform better on academic tests which are administered also in that language. Our aim was to assess the value of bilingual tests and whether they can be considered a viable tool for assessing academic knowledge of students who are both indigenous or immigrants. These students need to master the second language – Hebrew – for their academic and professional future, while their L1 is the language in which they are much more proficient, yet this is not acknowledged in the assessment procedures.

The study included 9th and 10th grade students whose home language is Russian or Arabic. They took a standardized science test as well as a writing task in a mono (Hebrew) and bilingual version (Hebrew and Russian/Arabic). In addition we conducted think aloud protocols with a smaller sample. Following the test we distributed feedback questionnaires to all participants, and conducted several focus groups with students.

The results indicated that overall, the option of bilingual assessment is quite appealing. Yet, several differences emerged between the groups; Arab students were comfortable in both languages, and yet attended to the two versions very often, especially in the writing task. The Russian students, on the other hand, have basically relied on the Russian version and explained their daily frustration from the lack of ability to express themselves properly in Hebrew and reveal their full knowledge, a phenomenon that tends to affect also their self-esteem and sense of belongingness.

Thus the study has important implication for applying the multilingual educational policy in Israel, whereby multilingual students will be able to express better their level of understanding, even if their proficiency in Hebrew is still not very good. It is not only a pedagogical issue but it reflects a positive socio-educational message to all students, and can improve minority students' achievements overall performance and well-being and enhance social justice and equality.

Developing lists of empirical English word difficulties specific to each L1

Steve Lattanzio, MetaMetrics, **Alistair Van Moere**, MetaMetrics, **Jeff Elmore**, MetaMetrics

8:30 a.m. to 9:00 a.m. Location: Decatur A

International English exams that claim to be standardized are challenged to ensure that test items are not biased against any one of 100+ language backgrounds that test-takers might represent. While some English words are easier to learn than others, it is not yet known how individual word difficulties differ for L1 and L2 students, nor how they vary by L1 or country. This presentation analyzes variation in individual word difficulties from country to country, which may help quantify the scope of the challenge facing globally "standardized" exams.

First, empirical difficulty for individual words was established via auto-generated cloze items. It is hypothesized that when a word is clozed out from a sentence, its difficulty can be disambiguated from context when it is presented in many different sentences to many different test-takers. In this study, 17,966 US-based L1 students engaged in 6.7 million item-level encounters, involving 59,740 unique words that appeared in 110,204 unique passages. A modified Rasch model was applied to cloze items where the item difficulty is a function of the word being clozed out and text complexity of the surrounding passage. Word difficulties were found by minimizing the logarithmic loss across the probabilistic outcomes. In other words, individual word difficulties are used to explain additional variance, as much as possible, than the Rasch model would with contextual text complexity alone. The word measures were placed on the equal-interval Lexile scale from approximately 0L to 2000L (the "L" denotes Lexile) and exhibited a normal distribution (mean=1150L, SD=510L). For validation, the resulting difficulty measures were compared with age-of-acquisition ratings (Kuperman et al, 2012) and EDL core vocabulary difficulties (Taylor et al, 1979), revealing corrected correlations of r=0.82 and 0.84 respectively.

Following this, another dataset was analyzed containing 3 million item-level encounters over 13,495 unique passages, from 20,096 learners of English from 199 different countries, who were predominantly preparing for the TOEFL exam. This dataset was used to compare word difficulties by country (a proxy for first language) for the 20 most well-represented countries in the data; 14,837 learners total.

A correlation matrix for pairings between each top-20 country's shared empirical word difficulties reveals coefficients in the range r=0.7 to 0.9 (e.g. Spain and Columbia, 0.90; Spain and Vietnam, 0.74). Principal component analysis (PCA) via singular value decomposition (SVD) was then performed to reduce the data to two dimensions, resulting in a plotted "landscape" of countries. The two-dimensional chart reveals a continuum of countries (and languages) that make intuitive sense: countries that share native languages with common roots are clustered together.

This research results in the first known list of country-by-country empirical word difficulties, together with standard errors, for 20 countries and over 9,000 words, which is underpinned by some degree of validation. The list could be utilized by test developers intending to assess English learners' reading ability, or curriculum developers considering content for EFL textbooks. The limitations of the study, including missing data, and possible next steps are discussed.

Rater behavior in a high-stakes L2 examination: Does test takers' perceived first language matter?

Ari Huhta, University of Jyväskylä, **Sari Ohranen**, University of Jyväskylä, **Mia Halonen**, University of Jyväskylä, **Tuija Hirvelä**, University of Jyväskylä, **Reeta Neittaanmäki**, University of Jyväskylä, **Sari Ahola**, University of Jyväskylä, **Riikka Ullakonoja**, University of Jyväskylä

9:05 a.m. to 9:35 a.m. Location: Mary Gay C

Language proficiency is an important gatekeeper between groups of people. Currently, many European countries have language requirements for citizenship, which is why language testing is a possible source of social inequality and why language testers should try to ensure that they do not discriminate against particular examinees.

We report on research investigating if raters working for a high-stakes second language (L2) examination exhibit different rating behavior with different first language (L1) background learners, and if any differences can be seen as unfairness. The research combines theories and methods of sociolinguistics, particularly (socio)phonetics and language ideologies, language testing and statistics. The project contributes to critical social perspectives on language testing for citizenship (Shohamy 2001; Hogan-Brun, Mar-Molinero & Stevenson 2009; Extra, Van Avermaet & Spotti 2009; McNamara & Roever 2006).

The study examined raters working for an L2 examination used as part of the citizenship process in a European country. The speaking part of the language examination in question takes place in a language lab and rating is based on audio recordings of examinees' performances. In the study, 44 certified raters assessed 50 examinees (one speaking task/performance per examinee). The examinees represented different migrant or minority languages, including Arabic, Russian, and Thai, that regularly face negative stereotyping in the country. Each language group comprised five male and five female speakers, mostly at CEFR levels B1 or B2. The raters judged the performances both holistically and separately for all the analytical criteria used in the examination (e.g., pronunciation, fluency, accuracy). The raters were also asked, through an open-ended question, to identify the speaker's L1 and to indicate how certain that identification was (ranging from 'certain' to 'uncertain').

Data were collected through an online platform that captures the rating process in detail, e.g., the order in which criteria are used and possible changes in ratings. Multifaceted Rasch analyses were conducted to discover any rater biases toward particular learners/groups or interactions between criteria and learner groups.

We analysed whether different L1 groups were rated differently and whether the raters' identification of the speaker's L1 affected assessments. We found that the raters used analytical criteria somewhat differently across the L1 groups; ratings of pronunciation, in particular, differed from the other criteria: for some groups, pronunciation was rated significantly higher vs lower than the other criteria. Comparisons of the raters shed interesting light on these differences: the raters who reported identifying the examinee's L1 differed from those raters who could not identify the L1 or who were uncertain about it. Again, the criterion most affected by L1 identification was pronunciation, and the effect was found for certain L1 groups but not for others. The findings suggest that if the raters know, or believe they know, examinees' L1, this may affect their rating of pronunciation, at least. Finally, we discuss possible reasons for the differences for why certain L1 speakers of the L2 were rated differently in this context, as well as the implications of the findings for fairness of assessment and need for further research.

Social justice and washback in language testing in Norway

Marte Monsen, Inland Norway University of Applied Sciences

9:05 a.m. to 9:35 a.m. Location: Henry Oliver

One key concern when it comes to social justice in language testing, is the washback effects the tests have on teaching, learning and on the society as a whole. In Norway today, the Common European Framework of Reference for Languages (CEFR, Council of Europe 2001) is the basis for curricula and assessment practices in all foreign- and second language learning. The framework is particularly prominent in assessment of adult learners of Norwegian. In order to get a job, to be able to study in Norway and to get citizenship or permanent residence, immigrants to Norway have to pass an oral and/or written CEFR-based test in Norwegian. Especially because of these gatekeeping functions of the CEFR, the tests have some quite high stakes. It is therefore important to investigate assessment according to the CEFR-scale. Based on a case study in two different learning centres for adult immigrants, this study explores washback effects on language teaching, especially concerning the teaching of writing. The study investigates writing tasks, student texts and teacher responses to student texts, as well as semi structured interviews with teachers and students about writing tasks and writing instruction, and the relationship between writing instruction and the language tests that the students take.

An important criterion for the CEFR is that it should "not irrevocably and exclusively [be] attached to any one of a number of competing linguistic or educational theories or practices" (Council of Europe 2001: 8). Many researchers have pointed out, however, that the framework in practice relies on specific theoretical and methodological considerations (see for example McNamara 2014, Leung & Lewkowich 2013, Martin 2013). This presentation will explore what assessment according to the CEFR looks like in a Norwegian context, and discuss how it affects both writing instruction and the language learners' participation as equal citizens in the Norwegian society. Findings from the study indicate that one of the washback effects of the tests is that teachers of L2 writing for adult learners adapt the content and the scope of their teaching to what is measured in the tests. In this way, what counts as writing in the CEFR-based tests affects all input that adult L2 learners get about writing in Norwegian. Teachers and the learners doubts that the tests covers the types of writing that the learners needs and wants to be able to do, for example participating in different kinds of digital communication. Nevertheless, because of time pressure and anxiety that the learners shall fail the tests, both teachers and learners want the instruction to focus on preparing for the tests.

Council of Europe 2001. Common European framework of reference for languages: Learning, teaching, assessment.

Leung, C. & Lewcowicz, J. 2013. Language communication and communicative competence: a view from contemporary classrooms. Language and Education 27 (5)

Martin, M. 2013. The complex simple – a problematic adjective in the CEFR writing scales. NORDAND 2/2013

McNamara, T. 2014. 30 Years on – Evolution or Revolution? Language assessment quarterly 11 (2).

Exploring the Impact of Bilingual Education Types on DIF: Implications for Vocabulary Test Development

Suchada Sanonguthai, Indiana University Bloomington

9:05 a.m. to 9:35 a.m. Location: Decatur A

Researchers in the field of language assessment have largely ignored contextual variables including socioeconomic status, teaching practices, and parental style as possible sources of differential item functioning (DIF) in DIF studies (Zumbo, 2007). To address this gap in DIF research, the present study aims to investigate the impact of bilingual/immersion education types across gender on differential item functioning (DIF) by analyzing a university-level English vocabulary subtest (25 dichotomously scored items) for detecting DIF items and for identifying possible sources of DIF. The current study is guided by three research questions: the first research question is whether test takers from the same gender, nationality, first language, and curriculum but in different types of bilingual/immersion education programs show differing probabilities of success after matching on their underlying lexical ability; the second research question is whether gender plays any role as an explanatory source of DIF; and the final research question asks if any other possible sources of DIF are identifiable. The study is conducted on data from a large-scale English assessment used to help inform undergraduate university admissions' decision making. Data in this study were collected from three groups of high school students (N=849) from differing types of bilingual schools in Thailand.

The methods used for DIF analysis were the Mantel-Haenzel (Holland & Thayer, 1988) and the Logistic Regression procedures (Swaminathan & Rogers, 1990). Both methods are suitable for the present study due to the relatively small sample size. Moreover, the Logistic Regression method was selected because it can detect both uniform and non-uniform DIF effects. A major finding in this study is that two uniform DIF items with the target words foe and curb favored male students from the dual language school in comparison to their male counterparts from the immersion school. There were no disagreements between the two methods, but the magnitude of uniform DIF effect was found to be moderate by Mantel-Haenzel ($p < .05$) and small by Logistic Regression based on Jodoin and Gierl (2001).

In addition, two non-uniform DIF items with the target words "invasion" and "invite" were flagged with significant moderate effect by Logistic Regression. Intermediate and advanced male students from the dual language school were more likely to respond correctly to "invasion" than their male student counterparts from the immersion school despite having comparable ability. However, the probability of responding correctly to "invite" was higher in intermediate and advanced male students from the immersion school. After factoring in female students from the dual language school, the conditional probability of responding correctly on "foe," "invasion," and "grip" was higher for female students from the dual language school than their male counterpart from the immersion school. The findings are discussed in reference to word frequency ranking (Davies & Gardner 2010) and differing rates of acquiring various senses or meanings that lexical entries can be used to express in a given context as potential explanatory sources of DIF in the context of L2 vocabulary test.

Not Unwarranted Concordances But Warranted Convergences: Approaches to Standard Setting and Maintenance Using Subject Experts

Gad S. Lim, Michigan Language Assessment, **Barbara Ying Zhang**, Cambridge Boxhill Language Assessment, **Brigita Séguis**, Cambridge Assessment English

9:40 a.m. to 10:10 a.m. Location: Mary Gay C

Shohamy's landmark 1998 article called for "assessment models which are more educational, democratic, and ethical in order to minimize misuses" (p. 343). One way to pursue this is by moving away from the current overreliance on general English tests and moving towards the use of English for Specific Purposes tests, which can assess more field-appropriate constructs, and indeed be developed with the participation of expert informants from those fields.

A case in point is the Occupational English Test (OET), a test of English for the healthcare-workplace context, the content of which was originally arrived at by observing healthcare workers in their workplaces to determine their communicative situations and requirements (McNamara, 1996), and whose content continues to be vetted by healthcare professionals, thus increasing its validity. More recently, the test has been revised to further expand construct coverage. In addition to doctor-patient interactions, professional to professional communication has been included on the Listening test, while "indigenous criteria" valued by healthcare professionals themselves were incorporated into the Speaking (and Writing?) tests (Elder, et al., 2013; Knoch, et al., 2017). Positive testing practice and use can further be promoted if, in addition to test content, standards on the test were also determined not by resorting to unwarranted concordances with other tests or with vague comparative frameworks, but informed by healthcare experts themselves.

This paper reports on the setting of standards on the original and updated versions of the OET, focusing on the Listening and Reading sub-tests, providing evidence in support of scoring validity across versions and implications for standard setting practice. While cut scores on the revised test could have been arrived at through purely quantitative means (i.e. IRT equating), judgment-based standard setting exercises were conducted involving panels of healthcare professionals and educators, for two reasons: (1) so that test content and test construct, and test takers and use context, are accounted for in the determination of the standard; and (2) so that the existing standards could be validated; while originally also informed by healthcare professionals, the existing standards have been maintained over the last 30 years using an equipercentile scoring method, which depends on strong assumptions about the characteristics of the test population, and may therefore not be warranted.

Analysis consisted of scoring the first session of the revised test using the existing equipercentile method, and then separately scoring the test using the cut scores derived from the Angoff-method standard setting exercises. The two approaches yielded the same cut scores on old and new tests for both Listening and Reading. The convergent outcomes do not just provide validity evidence for the test, but raise important questions for standard setting theory, which accepts divergent outcomes and does not see it as problematic (Camilli et al., 2001; Cizek, 1993; Zieky, 2001), and for methodology in standards maintenance beyond conventional IRT-based equating approaches in testing practice. More importantly, the study illustrates how an approach to testing informed by subject matter experts can lead to more valid, appropriate, and warranted outcomes.

Intended and unintended consequences of reforming a national school-leaving exam and their role for validation

Benjamin Kremmel, University of Innsbruck, **Carol Spoettl**, University of Innsbruck, **Veronika Schwarz**, University of Innsbruck

9:40 a.m. to 10:10 a.m. Location: Henry Oliver

With the political decision in 2004 to introduce a CEFR-based national curriculum for the modern foreign languages, Austria became one of the first European countries to implement the Common European Framework of Reference on a nationwide level. In 2007, this led to a subsequent incumbent reform of the secondary school-leaving examination towards standardized, professionally developed and CEFR-linked language tests. The exam was legally anchored in 2010 and was one of the first compulsory national school-leaving exam systems based on the European reference instrument.

While great care has been taken from the beginning to ensure professional exam development and delivery in this new exam, and considerable emphasis is put on checking the technical quality of the exam, little concern in validation efforts has gone to the consequences that the exam reform has brought about. In light of Chalhoub-Deville's (2016) call for the need to incorporate consequences systematically into the validation of exam reforms, which are specifically intended to bring about a certain impact, such as increased accountability, this talk outlines the initial efforts to frame validation research on this exam in a Theory-of-Action approach (Bennett, 2010) by making explicit the implicit intended and unintended consequences and motivations of various stakeholders behind the reform, so that they can be addressed and checked against actual consequences in future validation research.

The paper reports on four different studies that have been conducted to address this: a) a qualitative analysis of the transcripts of a key parliamentary debate prior to the legal anchoring of the reformed exam, b) a semi-structured interview with a former ministry representative, c) a semi-structured interview with the language testing expert coordinating the exam reform, and d) semi-structured interviews with language teachers (N=10). Coded themes emerging from these sources allowed for identification of intended and unintended consequences as seen by representatives of three key stakeholder groups: policymakers, language testers, and classroom teachers.

The findings show that while, for instance, language testing experts seem to have been successful in fostering a minimal level of assessment literacy in a few policymakers, the key motivations and intended consequences between the different stakeholder groups still diverged considerably. Policymakers see exam reforms as tools to implement international comparability, competitiveness and "functionalist values" (McNamara, 2012), whereas language testers and teachers think the intended consequences are communicative competence and increased assessment literacy. In any case, making all of these (often implicit) intended consequences explicit is necessary to evaluate and validate the reformed exam as to whether or not these have been brought about. The paper will therefore suggest ways forward as to how to take this into account in future validation research and concludes by echoing Chalhoub-Deville's (2016) demand for a stronger focus on political dimensions and consequences in validation research on accountability-based exam reforms.

A Knowledge-based Vocabulary List (KVL): German, Spanish, and Chinese Results

Norbert Schmitt, University of Nottingham, **Barry O'Sullivan**, British Council, **Laurence Anthony**, Waseda University, **Karen Dunn**, British Council, **Benjamin Kremmel**, University of Innsbruck

9:40 a.m. to 10:10 a.m. Location: Decatur A

Frequency of occurrence has been the main criterion for vocabulary test word selection. This approach has been useful, but the frequency-ranking is still somewhat crude. It is true that learners typically know more words in high-frequency bands then lower frequency bands (e.g. 1K>2K>3K, etc.). But while frequency ranking works relatively well for bands (at least up until about the 5K band), it does not work so well for individual words (e.g. #1,150 accounts will not necessarily be learned before #1,200 rose). At the level of individual words, there are numerous frequency misfits with what learners actually know. For example, pencil and socks will likely be some of the first words learned, but appear as relatively lower-frequency vocabulary on frequency lists.

What is needed to supplement frequency lists is a rank list of English vocabulary based on the actual likelihood of L2 learners knowing the words. This paper will report on the development of the Knowledge-based Vocabulary List (KVL). The list will provide rank knowledge information on the best-known 5,000 lemmas in English. The list was developed in two stages. In the first stage, we developed a target list of around 8,000 lemmas to test with a large of participants, and completed customized tests for speakers of German (mother language of English), Spanish (cognate language), Chinese (non-cognate language). In the second stage, we began testing in May, through both targeted classroom groups and through crowd-sourcing. The goal is to have 300 responses per target word by Spring 2019. The results will be analyzed through both classical and IRT approaches. The result of the overall analysis will be a list of relative difficulty of the first 5K lemmas, which can be interpreted as rank probability of learner knowledge of these lemmas. This list should be very useful in test development, and will probably be used in conjunction with frequency lists.

The presentation will illustrate the resulting lists, and discuss how the lists vary according to the three different L1s. It will then discuss the implications for English as a Second/Foreign Language assessment.

Understanding Writing Process of Adult EFL Learners in a Writing Assessment Context

Ikkyu Choi, Educational Testing Service

11:00 a.m. to 11:30 a.m. Location: Mary Gay C

How a writer arrives at a writing product has been a central topic in writing research for decades. Most writing process models (e.g., Flower & Hayes, 1981; Grabe & Kaplan, 1996; Kellogg, 2001) assume, often implicitly, naturalistic writing contexts in which writers can spend as much time as they want to draft, develop, and revise their writing. On the other hand, writers in most assessment contexts need to complete and submit their writing products within a set time limit. This difference in contexts led some researchers to skepticism about studying writing process in assessment contexts (e.g., Wolcott, 1987). However, several case studies of how test takers write their responses have provided evidence for distinctive writing activities and a positive association between the presence of such activities and the quality of submitted responses (e.g., Barkaoui, 2015; Hall, 1991; Khuder & Harwood, 2015; Worden, 2009). Moreover, it has been argued that writing process information can help understand writers' strengths and weaknesses and thus have a potential as diagnostic feedback for test takers (Deane & Quinlan, 2010). Previous case studies have been conducted at small scales where small- or varying-associations may not be detected in a stable manner. Larger studies are needed to robustly document the relationship between writing process and product in an assessment context.

In this study, we aim to further the understanding of test takers' writing process and the relationship between their process and product. Specifically, we investigated how test takers' writing pace changed as a function of writing time, and how such changes were related to different types of writing activities as well as the quality of submitted responses. Our data consisted of 380 adult EFL learners' responses to two writing tasks and the logs of all keystroke events during their writing. Each response was scored by two randomly assigned raters on a five point holistic scale, and the average of the two scores was used as the variable representing response quality. Using the keystroke logs, we reconstructed the trajectories of test takers' writing progression (in terms of word counts) at every 5 second interval, and examined what writing activities took place within each interval. The resulting trajectories, when considered together with the writing activities, showed that most test takers underwent at least three distinctive stages: planning, drafting, and revising responses. Moreover, the distribution of the stages differed substantially across different levels of response quality. We summarized the trajectories with logistic growth curves, and examined the relationship between the estimated curves and the rater scores. We found that the estimates were strongly correlated with the rater scores, and that they improved the prediction of the rater scores over a baseline model incorporating only length and writing time. In sum, our analyses showed that test takers went through distinct stages of writing and that the nature of these stages was related to the score awarded to the submitted response. These findings support earlier claims for the usefulness of writing process information for understanding writing quality.

How do raters learn to rate? Many-facet Rasch modeling of rater performance over the course of a rater certification program

Xun Yan, University of Illinois at Urbana-Champaign, **Hyunji Park**, University of Illinois at Urbana-Champaign

11:00 a.m. to 11:30 a.m. Location: Henry Oliver

In language testing, rater studies tend to examine rating performance at a single time point. Research on the effect of rater training either adopts a pre- and post-training design or compares rating performance of novice and trained raters. While those cross-sectional studies provide insights into the immediate results of rater training, little is known about how raters develop longitudinally as they learn to rate with a new scale. This study employs a mixed-methods approach to examine how rater performance develops during a semester-long rater certification program for an ESL writing placement test at a large US university.

The certification program adopted an iterative training approach, consisting of several rounds of face-to-face group meetings, individual rating exercises, and scale re-calibration based on rater performance and feedback. We tracked the performance of 30 novice raters throughout the certification program between 2015 and 2018. Using many-facet Rasch modeling, rater performance was examined in terms of rater agreement, severity and infit statistics. These measurement estimates of rating quality were compared across rounds and years. Rater comments on the essays were qualitatively analyzed to obtain a deeper understanding of how raters learn to use the scale over time.

The Rasch results showed a consistent three-phrase developmental pattern of rating behavior across years. At the start of the certification program, all rater performance indices were within the acceptable ranges, but their comments showed that they had not fully acquired the rating scale and used construct-irrelevant scoring criteria to score the essays. This misuse of the rating scale led to a quick drop in rater performance in the subsequent round, leading raters to become unsure and less confident about their ratings. However, through repeated rating exercises and discussion, raters became more adapted to the

rating scale; their performance improved, eventually reaching the target level on multiple indices of rating quality. Raters also became more indistinguishable from one another in the application of the rating scale.

Findings of this study suggest that rater performance does not improve in a linear fashion; instead, their developmental pattern resembles a three-stage U-shape learning curve (Strauss, 1982), suggesting that raters learn to rate in a similar fashion as one acquires a language and other skills. We argue that understanding the developmental pattern of rater behavior is crucial not only to understanding the effectiveness of rater training but also to the investigations of rater cognition and development. We will also discuss the practical implications of this study in relation to the effort and expectations needed for rater training for local language assessments.

Strauss, S. (1982). Introduction. In S. Strauss & R. Stavey (Eds.), *U-shaped behavioral growth* (pp. 1–11). New York: Academic Press.

Language assessment and student performance in South African higher education: The case of Stellenbosch University

Kabelo Wilson Sebolai, Stellenbosch University

11:00 a.m. to 11:30 a.m. Location: Decatur A

More than two decades into the new political dispensation, South African universities still have to grapple with low levels of student academic performance and the consequent high dropout and low completion rates. The country's graduation rates have, in the words of the Council on Higher Education (CHE) (2013:15), been "found to have major shortcomings in terms of overall numbers, equity and the proportion of the student body that succeeds". This has mainly been attributed to the mismatch that seems to exist between the knowledge that learners leave the high school with and the kind that academic education requires them to possess for success. This gap, also known as the "articulation gap", has been ascribed not only to the emotional and academic under preparedness of the students entering high education, it has also been seen as an outcome of a complex combination of the political and socio-economic factors that are unique to the country. Among the academic reasons often cited for this "articulation gap" are the levels of language ability that high school leavers bring to higher education (See CHE 2013; Coetzee & Coetzee Van Rooy 2015; Van Dyk 2015). While this is taken for granted by language teaching professionals, the tendency is for non-language teaching academics and students enrolled in disciplines that are traditionally less associated with language processing to be sceptical of the role of language assessment and intervention in academic performance. Not only has this been the case in South African higher education in general, it is also a challenge faced by language departments in other parts of the world (See Carkin 1997; Stoller 2012). The aim of this study was to investigate whether predetermined standards of performance on the school-leaving English examination and a standardized test of academic literacy related positively with academic performance. In order to determine this, Pearson Correlations and Analysis of Variance (ANOVA) were carried out on the scores obtained on these assessments by a total of 836 first year students enrolled in different faculties at Stellenbosch University. The English examination is part of the overall school leaving assessment that has traditionally been used by South African universities for taking decisions on student admissions. The standardized test was designed to measure students' ability to cope with the language demands of higher education and provide additional information about their preparedness for this education. The results of the study showed that overall performance on these two assessments related positively with first year academic performance. These results also showed, however, that the performance standards set for the standardized test of academic literacy also associated positively with first year academic

performance while the scores on the levels of performance set for the school leaving English examination did not. This justifies the language interventions that are often motivated by and follow the standardized test, a confirmation of the role that this assessment plays in ensuring access with success and promoting ultimate social justice.

The Impact of an External Standardized Test on Teaching and Learning for Young Learners: A Year 1 Baseline Study in Turkey

Mikyung Kim Wolf, Educational Testing Service (ETS), **Alexis Lopez**, Educational Testing Service (ETS)
Jeremy Lee, Educational Testing Service (ETS)

11:00 a.m. to 11:30 a.m. Location: Decatur B

Previous research on test impact or washback provides evidence that large-scale, high-stakes assessments have a substantial impact on teaching and learning, generating both positive and sometimes unintended negative consequences (e.g., Bachman & Palmer, 2010; Messick, 1996). On the other hand, relatively little research has addressed how standardized tests that are not designed for high-stakes uses influence teaching and learning for young learners in school settings. Considering the increasing use of standardized assessments for young students learning English worldwide, empirical investigations into the impact of such assessments will offer useful information for the test users and test developers.

The present study aimed at examining the impact of international, standardized proficiency assessments designed for young learners, namely the TOEFL® Young Student Series (YSS) tests, on teaching and learning. The TOEFL YSS tests, consisting of the TOEFL®Primary™ and TOEFL Junior® tests (for learners ages 8+ and 11+, respectively), are designed to measure students' foundational and communicative language skills and provide information to support teaching and learning[1]. To systematically collect the baseline data for this longitudinal study, we adapted a framework of washback from previous frameworks in language testing (e.g., Alderson & Wall, 1993; Henrichsen, 1989; Hughes, 1993; Shohamy, 1993; Wall & Horak, 2006), expanding the concept of three dimensions (participant, process, product) and mediating factors. This paper reports on our analyses of the baseline data regarding the characteristics of the tests, various stakeholders as participants, and process (instructional practices, assessment uses) in order to understand the types of consequences that the TOEFL YSS tests bring about.

In Year 1, we recruited five schools in Turkey, which had been using the TOEFL YSS tests for at least two years. From each school, we selected first-year and second-year users of the tests. This approach led to a sample from grades 3 to 6, as schools differed in the grade level at which they first administer the tests. In total, 18 teachers and 5 school administrators participated in the study, along with their students and students' parents. Data sources included surveys (teachers, students, and parents), interviews (school administrators, teachers), monthly instructional logs, instructional materials, TOEFL YSS test materials (test booklets, workshop and practice materials), and students' test scores. Descriptive statistics were computed for the surveys and test scores. The qualitative data were also coded with a protocol and summarized for patterns.

The results indicated that the TOEFL YSS tests were perceived and used as an "objective" measure to evaluate students' progress and the effectiveness of instructional programs. A strong demand for external, high-quality tests aligned to global standards was voiced by schools, reflecting their need to provide evidence on the effectiveness of their programs to parents. At the level of instruction, the impact of the tests was less apparent, with some variation across schools. In this presentation, we will

discuss our major findings and their implications regarding mediating factors that yield positive washback in our study contexts. We will also present our adapted framework of washback and lessons learned from the baseline study.

[1] https://www.ets.org/toefl_young_students_series

What aspects of speech contribute to the perceived intelligibility of L2 speakers?

Willam Bonk, Pearson, **Saerhim Oh**, Pearson

11:35 a.m. to 12:05 p.m. Location: Mary Gay C

Intelligibility can be defined in terms of the extent to which speech is considered easily understandable by listeners (Derwing & Munro, 1997; Munro, Derwing, & Morton, 2006), and likely is determined by a variety of variables such as the listener's familiarity with the accent (Ockey & French, 2016), the strength of the accent (Ockey, Papageorgiou, & French, 2016), morphosyntactic expression, lexical resources, pragmatic ability, etc. In practice, the perceived intelligibility of an L2 speaker is probably a holistic judgement based on a number of subconsciously perceived features that are relatively difficult for listeners to identify in a specific way. This research represents a quantitative attempt to narrow down the set of likely implicated features and to quantify the extent of their utility in predicting this intangible human construct of speaker intelligibility.

For the purposes of this study, intelligibility is measured via two parallel methods: listeners' judgments of how intelligible a particular set of utterances is (based on a scoring rubric), and the proportion of words in those utterances correctly identified by native speaker transcribers. For the former measure, a large number of naive English native speakers scored the responses on a holistic intelligibility rubric. For the latter, the same group of native speakers of English transcribed the speech from digital recordings.

The 11,052 responses analysed in this study were collected from a large-scale operational test in which candidates were incentivised to perform to the best of their ability. Two task responses were analysed for this research: elicited imitation of sentences, and retelling of narratives. The two tasks represent somewhat more constrained speech (elicited imitation, in which verbatim repetition is required), and less constrained speech (story retelling, in which the content of the story can be re-told in whatever words the speaker chooses). A large number of measures of phonology, fluency, and supra-segmentals was generated for each response by an artificial intelligence speech recognition and analysis system. The features produced by this system include measures related to rate of speech, amount of silence, pitch, disfluency, duration of particular phonological segments, quantity of response, pronunciation of particular phonemes, and energy, among others. Although listeners are thought to be sensitive to many of these speech characteristics, they are not necessarily able to verbalise their evaluations, so a computerised system is ideal for an objective measure of the variables' magnitude while maintaining their statistical independence from one another. The speakers' L1 and its phonological distance from English (Levenshtein, 1965) and gender are also included in the General Linear Model analysis.

The results of this analysis quantify the extent to which intelligibility as a construct 1) varies among naive native speakers of the target language, and 2) is predictable based on a large set of objectively produced low- and mid-level features. The results have implications for the automated evaluation of intelligibility, a major issue among test score users in an increasingly globalised world of communication.

Establishing a Validity Argument for a Rating Scale Developed for Ongoing Diagnostic Assessment in an EFL University Writing Classroom: A Mixed Methods Study

Apichat Khamboonruang, University of Melbourne

11:35 a.m. to 12:05 p.m. Location: Henry Oliver

In the classroom, assessment should take an integral part in providing productive information to support ongoing teaching and learning processes (Bachman & Damböck, 2018; Moss & Brookhart, 2010; Turner & Purpura, 2016). Diagnosing learners' strengths and weaknesses can provide useful assessment information to progressively promote learning and teaching in a language classroom (Elder, 2017; Knoch & Macqueen, 2017; Kunnan & Jang, 2009). Despite much recent interest in diagnostic assessment, more research is needed to understand how diagnostic assessment can best be integrated into language classrooms.

Following an argument-based approach (Kane, 2013, 2016), this paper reports on an investigation of evidence to support assumptions underlying Evaluation, Generalisation, Explanation, and Utilisation inferences in order to establish a validity argument for a novel 33-descriptor diagnostic binary rating scale. The investigation drew upon findings from the operationalisation stage of a three-stage study employing an exploratory sequential mixed methods research design (Creswell & Plano Clark, 2018) to utilise an array of qualitative and quantitative data to construct, refine, and justify the scale. The scale was intended to diagnose students' weaknesses and strengths in a five-paragraph academic writing product and to support teaching and learning in an ongoing Thai EFL university classroom context.

The final operationalisation stage aimed to justify the scale which was designed and developed building upon theoretical, intuitive, and contextual sources in the initial construction stage, and was evaluated and refined repeatedly based on experts' and teachers' feedback as well as statistical results in the subsequent trialling stage. The scale was operationally implemented in four academic writing classrooms over one semester in a Thai public university. Eighty English-major undergraduates were encouraged to use the scale to guide, self-assess, and revise their assignment essays before submission. Five teachers used the scale to diagnose students' essays and were encouraged to utilise diagnostic outcomes to support teaching and feedback. On course completion, semi-structured interviews were conducted to elicit teachers' and students' perceptions of the scale. To examine evidence for the validity argument, Many-Facets Rasch and qualitative content analyses were used to analyse scale data and perception protocol respectively.

Results showed that the assumptions underlying the inferences were reasonably supported by quantitative and qualitative evidence, thereby solidifying the overall validity argument for the intended interpretations and uses of the scale's diagnostic outcomes in the classroom context. To elaborate, psychometric results suggested that teachers were consistent, albeit with different degree of severity, in their diagnostic ratings and the scale had acceptable psychometric properties and well differentiated students' writing performances. Qualitative results revealed that overall the scale was user-friendly and useful for supporting teaching and learning in the classrooms although further revision of the scale was suggested for more practical and effective assessment in an ongoing classroom. The scale also helped teachers specify students' weaknesses and strengths and provide more targeted feedback and helped students realise the skills they needed to improve their essays. This study carries implications for teachers and researchers regarding the development, validation, and implementation of a diagnostic rating scale for a formative classroom language assessment.

Exploring teacher understandings and beliefs as a basis for benchmarking assessments for university foreign language programs

Noriko Iwashita, The University of Queensland

11:35 a.m. to 12:05 p.m. Location: Decatur A

It is well acknowledged that teachers' knowledge and beliefs guide their assessment practice in the classroom (e.g., Scarino, 2013), and that assessment literacy, is central to achieving and maintaining the overall quality of classroom teaching and learning (Popham, 2004). Teacher cognitions are shaped by teachers' prior learning and teaching experience which in turn influence decisions about classroom practice (Crusan et al, 2016). Studies have reported considerable variations in teachers' beliefs and knowledge about assessment and teaching methodology according to their experience and language background (e.g., Crusan et al, 2016; Kim, 2014; Malone, 2013), but relatively little research has been undertaken in university foreign language teaching context. Such research is critical given that language competence is given greater priority in the selection of instructors in the university foreign language program than is the case with secondary, primary school and university ESL teachers, with the result their understandings of teaching methodology and assessment are relatively unknown.

This paper reports on an investigation of how university teachers' understandings and beliefs about teaching and assessment guide their classroom and assessment practices. The study forms part of a project aimed at benchmarking assessment procedures in undergraduate language courses at a large urban university in Australia offering eight foreign languages with 2,000 enrolments annually. In order to ensure fairness and consistency in standardising learning outcomes, a large-scale investigation of current assessment procedures and academic achievement standards has been undertaken with the aim of revising and aligning assessment a) internally, across language programs, and b) externally, with an international benchmarking framework (the CEFR). Despite comparable linguistic goals across language courses, the initial benchmarking exercise has revealed that assessment practices vary significantly regardless of language typology (i.e., Asian or European languages).

To further investigate the source of this variation, 35 language instructors of a variety of languages offered at the university were invited to participate in a questionnaire survey and focus-group interview. The questions asked in both the survey and focus group interview elicited information about a) teachers' knowledge, experience and beliefs about classroom and assessment practices and b) aspects of assessment expertise deemed important for teachers to acquire, and c) general perceptions about the CEFR as well as d) teachers' backgrounds (e.g., L1, experience, teacher training).

The analysis revealed that teachers articulated their knowledge and beliefs about their classroom and assessment practices and their approaches to teaching in relation to the curriculum and their own learning experience. The teachers' education and language (L1) backgrounds were two of the most influential factors in the way their knowledge and beliefs were shaped. These findings confirm those of previous research in the primary and secondary schools context and university ESL programs (e.g., Davison, 2004). They also provide useful information for designing workshops to create a common 'culture of assessment' understood as shared attitudes, approaches, and understandings that support the evaluation of student learning outcomes. It is argued that only through building such a culture is possible to achieve successful alignment of university foreign language courses with an international benchmarking framework.

Investigating the consequential validity of the Hanyu Shuiping Kaoshi (Chinese proficiency test) by using an Argument-based framework

Shujiao Wang, McGill University

11:35 a.m. to 12:05 p.m. Location: Decatur B

In recent years, China's rising global power has led to an international increase in Chinese language learning. The national standardized test of Chinese language proficiency for non-native speakers, the Hanyu Shuiping Kaoshi (HSK), literally "Chinese Proficiency Test," has played a vital role in certifying language proficiency for higher education and professional purposes. The multiple uses of the HSK have generated growing concerns about its validity, especially the reformed HSK's (post-2009) consequential validity. Employing Shohamy's (1998) Critical Language Testing theory and adaptingBachman and Palmer's (2010) Assessment Use Argument (AUA) framework, this mixed methods sequential exploratory (MMSE) study investigates the HSK's micro- and macro- level consequences, how and to what extent the test affects Chinese as a second language (CSL) teaching and learning, as well as the relationship between the Promoting Chinese Internationally (PCI) policy and any micro or macro level consequences of test (HSK) use. In Phase I of this MMSE study the official HSK documents were analyzed by content analysis; interviews with 12 test stakeholders were then conducted and analyzed by a two-cycle qualitative coding approach (Saldaña, 2009). In Phase II, 136 CSL/CFL teachers and 512 HSK test-takers participated in a questionnaire, and the data were analyzed by using exploratory factor analysis (EFA) and structural equation modeling (SEM); classroom observations were also conducted and analyzed to contextualize the quantitative results. Phase III involved two exploratory questionnaires and interviews with 35 administrative personnel who use the HSK to inform academic and employment decisions, and the data were analysed through statistical (e.g., descriptive statistics) and qualitative methods (e.g., grounded theory). The results of the MMSE study highlighted the complexity of the HSK's consequences and washback effects. They indicated that although the HSK had limited effects on teaching, it was somewhat successful in its goal of promoting CSL/CFL learning. In general, HSK scores and other related information (e.g., score report, level interpretation) also provided users with relevant, useful, and meaningful data for candidate selection. Overall, based on the HSK's AUA conceptual framework, the findings provided evidence that Claim 1 (Consequences), Claim 2 (Decisions), and Claim 3 (Interpretations) were partially supported, in that the test developers' intended goals for the HSK were only achieved to a certain degree. This study helped unpack the consequential validity of the HSK in the CSL context, shed light on understanding the washback effects of the HSK, fleshed out the values underlying the multiple interpretations and uses of the test, and pointed to implications for the HSK developers and future consequence/impact/washback research.

Examination of test-taking strategies used for two item types during L2 listening assessment

Ruslan Suvorov, University of Hawaii at Mānoa

1:35 p.m. to 2:05 p.m. Location: Mary Gay C

Validation research in language testing has traditionally been outcome-oriented and relied on the use of statistical methods (O'Sullivan & Weir, 2011). However, there is a growing recognition of the need to utilize more process-oriented approaches to validation by gathering validity evidence based on the analysis of individual response processes (e.g., AERA, APA, NCME, Joint Committee on Standards for Educational & Psychological Testing, 2014; Messick, 1995; Wu & Stone, 2015). Following Shohamy's (1998) call for critical language testing that "challenges psychometric traditions and considers

interpretive ones" (p. 332), this study aims to critically examine the validity of interpretations of test scores obtained from a widely used 4-option multiple-choice item format. In doing so, this study leverages eye-tracking technology to explore and compare test-taking strategies used by L2 learners during the completion of traditional 4-option multiple-choice items vs. 4-option true-false items in a computer-based L2 listening test. Motivated by the results of the author's recent study (in press), in which L2 learners' use of test-wiseness strategies contributed to construct-irrelevant variance and had a statistically significant effect on their observed scores for 4-option multiple-choice items, this study also intends to examine whether a 4-option true-false format is less conducive to the use of test-wiseness strategies than a 4-option multiple-choice format.

Using the convergence model of the data triangulation design (Creswell & Plano Clark, 2007), this study analyzed test score data, eye-movement data, and verbal report data. These datasets were gathered during one-on-one 2-hour sessions with 40 participants who were non-native speakers of English at a public university in the US. During individual data collection sessions, every participant completed an online L2 listening test that consisted of six mini-lectures and 30 questions, with five questions per mini-lecture. To investigate whether participants' use of test-taking strategies varied depending on an item type, three mini-lectures were followed by questions presented in a 4-option multiple-choice format, whereas the other three mini-lectures were followed by questions presented in a 4-option true-false format. The distribution of the two item formats was counterbalanced among participants. During the test, participants' individual eye movements were recorded using a remote eye-tracker GazePoint GP3 HD (150 Hz). Immediately after the test, each participant's eye-movement recordings were used as a stimulus to elicit verbal data about test-taking strategies used during the test. Over 32 hours of verbal data were transcribed, coded, and analyzed in NVivo to identify the types of test-taking strategies reported by participants. Eye-movement data were analyzed manually using visual scanpath analysis and converged with the results of the verbal data analysis. Test score data were analyzed using a paired-samples t test. The results revealed that, compared to the traditional 4-option multiple-choice item format, 4-option true-false items appeared to be less conducive to the use of test-wiseness strategies such as guessing and resulted in observed scores that matched the true scores closer than the observed scores from 4-option multiple-choice items. Implications of these findings for future language test development and validation research efforts will be discussed in light of the conference theme.

Academic language or disciplinary practices? Reconciling perspectives of language and content educators when assessing English learners' language proficiency in the content classroom

Lorena Llosa, New York University, **Scott Grapin**, New York University

1:35 p.m. to 2:05 p.m. Location: Henry Oliver

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

Despite agreement on the existence of this overlap between language and content, the nature of the overlap is understood differently by language educators and content educators. From the perspective of

language educators, the overlap is represented by academic language in terms of lexical, grammatical, and discourse features. In other words, language accounts for content via academic language. On the other hand, from the perspective of content educators, the overlap is represented by disciplinary practices. Content accounts for language via language-intensive disciplinary practices, such as arguing from evidence or constructing explanations.

In an effort to reconcile these two perspectives, we propose an alternate conceptualization of English language proficiency at the intersection of language and content. We propose that to support English learners in the content areas, we start with the disciplinary practices, which provide the communicative purpose for which academic language is used. Specifically, we focus on (a) the nature of the disciplinary practices and (b) precision of the disciplinary meaning communicated through those practices. This approach highlights how academic language acts in the service of engaging in disciplinary practices, and thus goes beyond a conceptualization of academic language as a discrete set of features. Using this approach, content teachers can attend to language in meaningful ways by focusing on those aspects of language that are most crucial to engaging in the disciplinary practices of content areas. This approach is also more realistic for teachers who work with English learners but are not trained as language teachers. For example, by prompting students to be precise in the disciplinary meaning they communicate, teachers can indirectly address students' linguistic choices.

Using examples of science tasks aligned to K-12 science standards in the US, we will show how we operationalize this language construct and how this approach affords unique opportunities for rich formative assessment practices to support English learners in the content classroom.

Placement Testing: One test, two tests, three tests? How many tests are sufficient?

Kathryn Hille, Ohio University, **Yeonsuk Cho**, Educational Testing Service

1:35 p.m. to 2:05 p.m. Location: Decatur A

When students enroll in intensive English programs (IEPs), placement tests are typically administered to determine the level of English instruction that students should receive. Ideally, the use of a locally developed test may be desirable with respect to maintaining a close relationship between test content and the local curriculum so that test scores directly point to the level of instruction students should receive in a specific context. Nevertheless, according to one survey (Author, Year), the majority of institutions rely on commercially available English tests to support placement decisions. Despite the widespread use of commercially available English tests, however, there is little attention paid to the effectiveness of such assessments in aiding placement decisions. The purpose of the current study is to fill this gap by examining the usefulness of various English tests for placement decisions in one IEP at a Midwestern university in the United States.

The IEP in the current study uses a locally developed composition test, as well as two commercially available English tests—TOEFL ITP and Michigan EPT—for placement decisions. The IEP uses a two-step placement process, in which the local composition test and the TOEFL ITP are administered to all students, and the Michigan EPT is used as an additional placement test for students with lower proficiency levels. The Michigan EPT is assumed to help place low proficiency students more accurately than the TOEFL ITP. We evaluated the relationship of placement test scores with the accuracy of placement decisions and success in the assigned IEP courses, using 847 records consisting of test scores, beginning and end placement results, teachers' evaluations of placement accuracy, and average grades in assigned IEP courses. The results of the analyses indicated that the use of three placement tests together led to the most accurate placement decisions, followed by the use of a combination of the local

composition test and one of the commercially available tests. The assumption that the Michigan EPT is a better predictor of placements for low proficiency students than the TOEFL ITP was not supported by the results of the analyses. Based on the findings, we attribute improvement in the accuracy of placement results to the use of multiple measures, not that of a particular assessment. The relationships between placement test scores and GPAs in assigned IEP courses were generally weak. This result may indicate the effect of factors other than initial English ability on students' performance in English classrooms. Placement test scores from the beginning of a semester were also poor predictors of whether students were able to progress to a higher IEP course level at the end of that semester. Based on the findings, we will discuss the value of individual placement tests in the study context and the relevance of the results to other IEP contexts.

Developmental frameworks for writing in Denmark, Norway, and the US: A Cross-national comparison

Jill V. Jeffery, Leiden University, Netherlands, **Nikolaj Elf**, University of Southern Denmark, **Gustaf Bernhard Uno Skar**, Norwegian University of Science and Technology, **Kristen Campbell Wilcox**, The University at Albany, State University of New York

1:35 p.m. to 2:05 p.m. Location: Decatur B

Given an understanding of writing as a socioculturally situated activity, educational systems will necessarily vary with regard to 'the what, why and how writing is taught' (Graham and Rijlaardsam, 2014: 782). In this paper, we examine one domain of such variability: developmental pathways that are reflected in writing standards' grade level distinctions. To begin shedding light on how such pathways vary cross-nationally, we compare writing standards in three educational systems (Denmark, Norway, and the US) and examine their basis in theory and research regarding writing development. As such, we conduct a curriculum analysis to explore 'what should count as knowledge' (Deng and Luke 2008: 66) by attempting to make more explicit the presuppositions and framings of teaching and learning writing in different nationally-embedded school settings around the world. To contextualize our analysis, we discuss briefly how our research has been shaped by nationwide policies with regard to standards, curricula, and evaluation, based on our involvement with large-scale writing studies in different national contexts (Standards as a Tool for the Teaching and Assessing of Writing in Norway; Writing to Learn, Learning to Write in Denmark; the National Study of Writing Instruction in the US). We then provide side-by-side comparisons of how standards for writing are developmentally framed in each context. Finally, we compare the differing theoretical underpinnings of the standards and discuss their basis (or lack thereof) in empirical research regarding writing development. The analyses bring to light some important similarities and differences among educational systems, particularly with respect to the roles of student motivation and teacher expertise in various developmental framings.

Deng, Z. and Luke, A. (2008) Subject matter: Defining and theorizing school subjects. In M. F. Connelly, M. Fang, & J. Phillion (eds.) *The SAGE Handbook of Curriculum and Instruction* (pp. 66-88). Thousand Oaks: SAGE Publications.

Graham, S. and Rijlaardsam, G. (2014) Writing education around the globe: Introduction and a call for a new global analysis. *Reading and Writing: An International Journal 29*: 781--792.

Exploring the relationships between test value, motivation, anxiety and test performance: The case of a high-stakes English proficiency test

Jason Fan, University of Melbourne, **Yan Jin**, Shanghai Jiao Tong University

2:10 p.m. to 2:40 p.m. Location: Mary Gay C

Cognitive factors such as learning motivation and test anxiety have been shown to affect L2 learners' performance on high-stakes language assessments. Meanwhile, the social context in which a language test is developed and used is believed to affect learning motivation, test anxiety, and by extension, test performance. Few attempts, however, have been made to explore the relationships between and among these variables and how such interactions affect test performance (see e.g., Cheng et al., 2014; Wu & Lee, 2017 for exceptions). Drawing upon expectancy-value theory (e.g., Eccles & Wigfield, 2002) and self-determination theory (e.g., Ryan & Deci, 2000) in education research, this study explored the relationship between social variables (e.g., test takers' gender, attitudes towards test use, and test value), cognitive variables (L2 learning motivation and cognitive test anxiety or CTA), and test performance, in the context of a high-stakes English proficiency test developed and used in a first-tier university in China. The test is considered as high-stakes because passing it constitutes part of the graduation requirements for students.

The sample of this study consists of 318 students, with 146 males and 172 females. Methodologically, Rasch measurement theory and structural equation modelling (SEM) were used sequentially and in a complementary manner, following Bond and Fox (2015). First, the Rating Scale Model in Rasch measurement theory was used to examine the construct validity of the instruments in this study, and to calibrate person measures of test value, attitudes towards test use, L2 learning motivation, and CTA. Next, these person measures were imported into EQS 6.3, an SEM program, to model the relationships between the social variables (i.e. test value and attitudes towards test use), the cognitive variables (i.e. L2 learning motivation and CTA), and test performance. Finally, a multi-sample SEM analysis was performed to investigate whether the model was invariant across male and female test takers.

Different Rasch analyses lent support to the construct validity of the questionnaires used in this study. Path analysis in SEM indicated that the hypothesized model fit the data satisfactorily. An examination of the path coefficients reveals that test value had a significant positive effect on both intrinsic and extrinsic motivation, but a negative effect on CTA. Attitude towards test use had a significant positive effect on intrinsic motivation and negative effect on test anxiety. Significant positive correlations were observed between test value and attitudes towards test use. These variables in combination explained about 13% of the variance in test performance. Multi-sample SEM analysis supported the invariance of the model across male and female samples.

This study is one of the few attempts in the field aimed at exploring the complex relationships between and among test value, attitudes towards test use, test motivation, test anxiety, and test performance. The findings provide insights into how social and cognitive variables interact with each other in complex ways, which in turn shape students' performance on high-stakes language tests. The role of social factors in language learning and assessment will be discussed.

The role of feedback in the design of a testing model for social justice

Slobodanka Dimova, University of Copenhagen

2:10 p.m. to 2:40 p.m. Location: Henry Oliver

Internationalization of higher education has resulted in an increased establishment of English medium instruction (EMI) courses at universities in non-Anglophone countries. Due to the growing concerns about non-native English speaking (NNES) lecturers' ability to teach in English, universities have started implementing policies enforcing assessment of EMI lecturers' English proficiency. However, research on the effectiveness and the social consequences of these assessments remains limited, especially in terms of the power imbalances such assessments may create at the university workplace. Based on the principle that the oral English proficiency certification should provide benefits for, instead of simple exclusion of, test-takers who have less-advantaged starting position, i.e. lecturers with lower English proficiency (Davies, 2010), the paper will argue that the language assessment models for EMI certification should put emphasis on the formative feedback, and thereafter propose what may be considered relevant and effective feedback content.

The discussion will be based on the results from a mixed-methods study that examines the interpretations and the utility of the language-related feedback for test-takers at two non-Anglophone universities who took a performance-based oral English proficiency test for EMI certification. Three main data collection procedures were undertaken: reports with written feedback (n=400), a survey (n=100), and semi-structured interviews with test-takers (n=24). Results suggest that test-takers experienced difficulties comprehending the technical language in the feedback reports, especially when it included separate references to the different linguistic aspects of the performance (pronunciation, grammar, vocabulary, and fluency). However, they found the specific quotes from their own performance helpful in understanding their English language use strengths and weaknesses. Most importantly, the test-takers emphasized the utility of the feedback that grounded their English language uses in the specific EMI domain, and appreciated the opportunity to discuss their challenges and uncertainties, as well as to seek recommendations in an oral feedback session.

To conclude, test-takers may find the certification requirement meaningful and less-threatening when they are presented with the occasion to review the results and become aware of the different aspects of their English language ability. Although the feedback may focus primarily on the language characteristics of the performance, its relevance is enhanced when it is embedded in the context of the target language use domain to which it relates.

Mitigating rater bias in L2 English speaking assessment through controlled pairwise comparisons

Masato Hagiwara, Duolingo, **Burr Settles**, Duolingo, **Angela DiCostanzo**, Duolingo, **Cynthia M. Berger**, Duolingo, Georgia State University

2:10 p.m. to 2:40 p.m. Location: Decatur A

It is well-known that language performance assessments can be impacted by human rater effects (Kondo-Brown, 2002; Lynch & McNamara, 1998). Of particular concern is a form of rater bias whereby entire groups are scored differentially in systematic ways, such as gender impacting impressions of speaking ability. Lumley and Sullivan (2005) and Eckes (2005) both found that raters favored female examinees over males, while others found no evidence of rater biases based on gender (Bijani and Khabiri, 2017). Regardless, most work in this area has focused on expert raters with smaller sample sizes.

Meanwhile, there is growing interest in using crowdsourced ratings to efficiently develop non-native speaking assessments (Loukina et al., 2015; Ramanarayanan et al., 2016). While evidence suggests that crowdsourced judgments can be as accurate as experts (Evanini et al., 2010), there is reason to believe that crowdsourcing could also amplify rater bias, particularly with respect to gender (Parson et al., 2013). Unfortunately, there has been little investigation into rater bias in crowdsourcing for language assessment.

We describe a method of "controlled pairwise comparisons" to help mitigate bias in crowdsourced ratings. Following Baumann (2017), we presented raters with speech samples from two examinees, and raters chose which one demonstrated higher speaking proficiency. Pairwise comparisons were then aggregated into an overall ranking using a Bayesian skill rating system (Herbrich et al., 2007). We compare "random" vs. "controlled" pairing — the latter using stratified sampling so that paired examinees were as similar as possible (e.g., same gender). Subjects were 1,045 examinees who took an online computer-adaptive test of English, resulting in 5,225 random and controlled pairs.

In the first experiment, we used Amazon Mechanical Turk to collect human pairwise ratings for speaking responses. Results indicated a trend favoring females for the random case (+2%, $p = 0.09$), but not for the controlled, suggesting that same-gender comparisons helped mitigate gender differences. In the second experiment, we simulated pairwise judgments based on test scores with and without gender bias. In the "no bias" condition, higher-scoring examinees in each pair were rated more proficient (subject to random noise). Results showed no difference between male and female scores for either the random or controlled case, and aggregated rankings were significantly correlated with underlying scores ($\rho = 0.98$). In the "bias" condition, test scores for female examinees were artificially inflated before simulating judgements. Results yielded higher scores for females in both cases, with greater effects for random (+18%, $p < 0.001$) than for controlled (+5%, $p = 0.002$). Correlation with underlying scores also dropped for the random case ($\rho = 0.94$) but remained high for the controlled ($\rho = 0.97$).

Our crowdsourced rating results echo previous findings in favor of females in L2 speaking assessment. Furthermore, both real and simulated experiments indicated that controlled pairwise preferences may help mitigate rater effects with respect to gender bias, even when such effects are extreme. Future plans to validate aggregate rankings based on pairwise comparisons and extend inquiry to other potential sources of bias will also be discussed.

Examining the Structure, Scale, and Instructor Perceptions of the ACTFL Can-Do Statements for Spoken Proficiency

Sonia Magdalena Tigchelaar, Western Michigan University

2:10 p.m. to 2:40 p.m. Location: Decatur B

The NCSSFL-ACTFL (2015) Can-Do Statements describe what learners can do in the target language at the ACTFL Proficiency sublevels. Unlike the extensive research undertaken to scale and refine the CEFR descriptors (Council of Europe, 2001; North, 2000; North & Schneider, 1998), the NCSSFL-ACTFL statements have yet to be empirically validated. The ACTFL scale and performance indicators were constructed using language teachers' beliefs and experiences (Shin, 2013). While this is a logical starting point, concerns include whether the difficulty levels of the skills described in the statements match their assigned proficiency levels, and whether each statement accurately measures ACTFL's (2012) unitary and hierarchical model of language proficiency.

This study addresses these concerns by analyzing a self-assessment of 50 NCSSFL-ACTFL (2015) Can-Do Statements for speaking, ranging from Novice-Low to Superior. Spanish language learners (N = 886) of

varying proficiency levels rated the statements as: 1 (I cannot do this), 2 (I can do this with much help), 3 (I can do this with some help), 4 (I can do this). Item responses were analyzed using an exploratory factor analysis (EFA) and a Rasch model. The EFA revealed two possible models of spoken proficiency represented by the statements: a one-factor model in line with ACTFL's unidimensional model of proficiency, and a two-factor model that separated items into basic interpersonal communication skills (BICS) and cognitive academic language proficiency (CALP; Cummins, 2000).

The mean item difficulties at the major thresholds estimated by the Rasch model ascended in the anticipated order: Novice (M = -4.27, SD = 1.35), Intermediate (M = -0.56, SD = 1.38), Advanced (M = 2.08, SD = 1.66), Superior (M = 3.31, SD = 0.11), and the differences were statistically significant. The mean item difficulties also ascended according to the ACTFL sublevels (i.e., Low, Mid, High): There were significant differences between items from the Novice and Intermediate sublevels, but the instrument did not discriminate well between statements at the Advanced sublevels. In a follow-up focus group interview, language instructors (N = 9) suggested how the Advanced-level Can-Do Statements could be used to self-assess spoken proficiency and modified to increase psychometric value. They explained that the more advanced statements were "very abstract," required "specific background knowledge and experience," and lacked "specific requirements…context, audience." They also highlighted that it may not be possible to assign the statements to specific proficiency sublevels, given the expectation of emerging and sustaining proficiency at the Low and High sublevels, respectively.

This study has implications for the measurement of spoken language proficiency and for language learners' use of can-do statements. First, if academic speaking and general spoken proficiency represent two different factors, it may be more appropriate to design proficiency tests that measure these skills separately. This is particularly important for young language learners, who may not yet have the developmental capacity to speak about abstract academic topics. Second, writing can-do statements that include the types of detail suggested by the language instructors may enable language learners to use the statements to more effectively self-assess their spoken language proficiency.

Establishing appropriate cut-scores of standardized tests for a local placement context

Gary J. Ockey, Iowa State University, **Sonca Vo**, Iowa State University, **Shireen Baghestani**, Iowa State University

2:45 p.m. to 3:15 p.m. Location: Mary Gay C

Social justice is at the heart of placement testing in university settings. Not only are test takers themselves impacted by placement decisions but also classmates and instructors. When test takers are misplaced into classes they do not need, they must spend time and often their own money on unnecessary coursework. Such misplacements can also result in a waste of university resources. On the other hand, when students are not placed into needed language support classes, they may unduly struggle to complete a program, their classmates may have to work with a partner who cannot collaborate effectively, and their instructors may have to work with students who are not prepared for their courses.

After an L2 English user is admitted to a university, various approaches are used to help determine the need for language support courses. Some programs use standardized test scores, which are available from the admissions process, while others use locally developed placement tests, which are typically designed to be aligned with the local language needs and ESL course curricula. The latter option may be more defensible, but sufficient resources to develop, administer, and score valid assessments are often not available. A third approach is to use standardized test scores to screen out students who are very

likely to pass the local placement test, followed by locally developed placements tests for those students who need further evaluation. The latter approach has potential for limiting the use of resources, but little research has been conducted to determine appropriate standardized test scores for a local context.

Researchers in a United States Midwestern university designed a study with the aim of setting appropriate cut-scores for a standardized test (TOEFL iBT), which could be used to exempt high ability students from taking their locally developed oral communication placement test. Their locally developed test was designed to determine the need for taking an oral communication course. To achieve this aim, 664 scores for both TOEFL iBT (the standardized assessment) and EPT OC (the local placement assessment) were used to determine an optimal standardized test score, which could be used to identify students who did not need to take the local placement test. Logistic regression indicated that TOEFL iBT speaking scores were moderate predictors of the local placement test decisions. A TOEFL iBT speaking score of 22 or a combination of 22 from the speaking score and 90 from the TOEFL iBT overall score could be used as acceptable cut-scores for EPT OC exemption. In addition, teacher judgements were obtained from seven sections of classes into which students had been placed based on the local placement test. Teachers used a 7-point scale to indicate the extent to which students were properly placed in their classes. Teacher evaluations suggested that students were generally placed into appropriate classes with the local placement test. Overall, this study suggests that the use of standardized speaking section scores can be used in conjunction with local placement test oral communication scores to determine appropriate post-entry university language support needs.

Towards social justice for item writers: Empowering item writers through language assessment literacy training

Olena Rossi, Lancaster University, **Tineke Brunfaut**, Lancaster University

2:45 p.m. to 3:15 p.m. Location: Henry Oliver

Item writers play a key role in the language test cycle, as they essentially need to operationalise the construct into actual tasks. Often, however, these assessment professionals receive a rather narrowly-focused training in writing items to a particular set of specifications. Usually, this training is limited to 'item writing guidelines' or instruction in mechanical aspects of item writing.

In this presentation, we argue that mechanical training is not sufficient for item writers to consistently produce high-quality items. Item writers need to be empowered through more comprehensive assessment literacy training so that they acquire a deeper understanding of not only 'how' but also 'why' item specifications contain specific requirements and how their work contributes to test validity. Our viewpoint is based on a study in which we explored the effectiveness of a three-month online item writer course that included instruction in writing specific item types as well as in broader language assessment principles. In our talk, we will provide an overview of the course's content, and explain how we evaluated the training through a 'pretest-posttest' design. More specifically, the 25 novice item writers participating in the study completed three item writing tasks prior to the course as well as after having finished it. The quality of the items they produced were evaluated by expert item reviewers. In addition, pre- and post-course stimulated recall interviews were conducted with the item writers, course feedback was collected at various points during the training, and online group discussions were analysed. In our presentation, we will primarily focus on the interviews, which aimed to gain information on the approach, procedures, and techniques the trainees used while writing items, the difficulties they encountered, and their post-course reflections on the training's effect on their item writing skills.

Analysis of the interview data revealed that improvements in item quality pre- to post-course were associated with an increased awareness post-course – on behalf of the item writer – of fundamental assessment principles such as authenticity and fairness, a deeper understanding of the test construct, and conscious attempts to improve the validity of items by avoiding bias, construct under-representation, and construct-irrelevant variance.

The findings suggest that bringing item writers to stage 4 of Pill & Harding's (2013) language assessment literacy continuum, i.e. procedural and conceptual literacy, empowers item writers in their work by enabling them to develop higher-quality items, and by making them more interested and more confident in writing language test items. Our study has practical implications for item writer training and, more fundamentally, for the empowerment of item writers and the development of good-quality tests.

Use of automated scoring technology to predict difficult-to-score speaking responses

Larry Davis, Educational Testing Service, **Edward Wolfe**, Educational Testing Service

2:45 p.m. to 3:15 p.m. Location: Decatur A

It would seem self-evident that making a scoring decision in a language performance test is more challenging for some responses than others, and it is reasonable to hypothesize that difficulty in reaching a decision might be reflected in lower inter-rater agreement. Nonetheless, the impact of language performance characteristics on rater agreement has received relatively little attention. One reason may be that this issue was previously of little practical importance: there was no simple way to identify in advance which performances would be "difficult to score" (and might be given additional scrutiny), and if necessary, scoring quality could be controlled through other means such as double rating. However, automated scoring technology may now make it possible to automatically identify responses that are likely to be difficult to score, which can then be flagged for additional attention from a human rater. This approach could prove especially beneficial in high-volume, high-stakes tests, where double scoring is costly but it is important to ensure that scores meet the highest standards for reliability and accuracy. Recently, Wolfe, Song, and Jiao (2016) used data from an automated scoring engine to identify features associated with difficult-to-score responses in a writing assessment, but to our knowledge no study of this sort has been done for speaking. In this presentation we report on an investigation in which we used automated scoring technology to identify speaking responses where raters were more likely to disagree.

Responses to TOEFL iBT independent speaking items were scored by 31 raters and then the level of rater disagreement for each response was quantified in terms of deviation from the average score. This deviation was used as a metric of scoring difficulty. Responses were also analyzed with the ETS SpeechRater^{SM} automated scoring engine, which produced measures of various aspects of performance that were then used to produce a regression model to predict scoring difficulty. In addition, think-aloud protocols were collected for a subset of responses, and raters' reported perceptions of the difficulty in reaching a decision were compared to scoring difficulty as measured by the disagreement metric. In addition, we evaluated the extent to which raters' comments regarding specific points of difficulty in decision-making compared with the SpeechRater features used in the prediction model. We conclude the presentation with a discussion of the feasibility of predicting difficult-to-score spoken responses in operational contexts, and the extent to which automated scoring technology provided insight into sources of difficulty for human raters when evaluating such responses.

Building a Partial Validity Argument for the Global Test of English Communication

Payman Vafaee, Teachers College, Columbia University; Second Language Testing Inc, **Yuko Kashimada**, Benesse Corporation

2:45 p.m. to 3:15 p.m Location: Decatur B

In 2016, the Japanese government announced that privately-run English tests of the four skills will be used for university entrance examinations. In March 2018, the government recognized eight standardized English tests for this purpose, including TOEFL, IELTS, and the Global Test of English Communication (GTEC). GTEC is a test used to assess Japanese junior and senior high school students' English proficiency for diagnostic and university-admission purposes. GTEC has been designed based on the principles of task-based assessment, and its scores are aligned with A1 to B2 CEFR levels.

Currently, GTEC has been adopted by more than 1500 Japanese schools and administered to more than one million students annually. The widespread use of GTEC's scores for making high-skates decisions necessitates a continual examination of its validity. The current study adopted the argument-based validation framework proposed by Kane (2006) to examine the item quality, internal consistency, score generalizability and dependability, as well as the factor structure of the reading and listening sections of the test. A separate study on the speaking and writing sections is being completed.

The data were collected from 69,302 students who took a single form of GTEC in 2017. Rasch analysis showed that items in both sections had acceptable fit indices ranging between .7 to 1.3, and the internal consistency of the listening and reading sections were .79 and .83, respectively. G-theory analysis revealed that 82.38% of the reading and 75.44% of listening scores' total variance was explained by the learner-ability facet of measurement, and the scores were dependable for classifying learners into the A1 to B2 CEFR levels. Finally, after testing several nested-models, the confirmatory factor analysis showed that a two-factor model fit the data best, which supports the hypothesized structure of these two sections of GTEC. These results will be discussed within the Kane's validity framework.

Automated essay scoring: An objective support to human raters

Matthew Kyle Martin, Brigham Young University, **Matthew Wilcox**, Brigham Young University

8:00 a.m. - 8:30 a.m. Location: Mary Gay C

It is estimated that non-native English speakers outnumber native english speakers by a factor of 3 to 1 (Crystal, 2003). With a growing number of English learners worldwide, the value of certification has increased. The demand for language certification is larger than ever and tends to outweigh the available supply of time, personnel, and training. An automated essay scoring approach adds efficiency and objectivity to guide essay scoring. While machine learning is not in a position to reliably replace human rating completely, there are things that it can do better than human raters. Thus machine scoring can potentially reduce load on human raters, and significantly reduce the cost of certification for those who currently may not have the means to pay for high-stakes assessments.

Researchers may sometimes be tempted to train and retrain a predictive model on the same data such that their algorithms produce amazing results, but only on their data. To avoid overfit, this study will create a model using iLexIR data. The iLexIR data (n > 1200) is a subset of the Cambridge Learner Corpus compiled by Yannakoudakis et al. (2011). The human-scored essays in this dataset range from 120 to 180 words and contain metadata about the speaker including L1 and age. A model will identify over 40 features, including demographics, in each of the responses from the iLexIR database. The iLexIR data will then be applied to foreign data, namely a large set (n > 1,000) of essays eliciting intermediate and advanced writing samples from non-native English speakers at the English Language Center (ELC) at a major Western University. The ELC data is rated by a minimum of two trained and calibrated human raters on an 8 point scale (0-7) which align with ACTFL levels Novice-Low to Advanced-Mid. The double-ratings are then analyzed using the MFRM, and students are awarded a single “fair-average” score.

This study will examine the extent to which the machine generated scores predict the fair average scores obtained from human raters using essays from the ELC. We will use regression analyses to examine the fit of univariate and multivariate models. Further, we will examine the level of confidence in assigning a score using the automated essay scorer, and any variations in confidence based on other variables such as ability level.

Machine Learning and AI are not the end-all answer to essay scoring, but this work seeks to add what machine learning does well in order to support what human raters do well in scoring writing. Ultimately, we expect to be able to electronically grade essays with greater and greater confidence, allowing for quicker scoring and less variation due to rater bias, all at a lower cost for programs and individuals seeking to improve English-language ability.

Crystal, David. *English as a Global Language* (2nd ed.). Cambridge University Press. p. 69. ISBN 978-0-521-53032-3.

Yannakoudakis, Helen, Ted Briscoe & Ben Medlock. *A New Dataset and Method for Automatically Grading ESOL Texts*. 10.

Talk2Me Jr: A Digital Language and Literacy Assessment Tool

Samantha Dawn McCormick, University of Toronto, **Hyunah Kim**, University of Toronto, **Jeanne Sinclair**, University of Toronto, **Clarissa Lau**, University of Toronto, **Megan Vincett**, University of Toronto, **Chris Barron**, University of Toronto, **Eunice Eunhee Jang**, University of Toronto at the Ontario Institute for Studies in Education

8:00 a.m. - 8:30 a.m. Location: Henry Oliver

Developing innovative and accessible digital assessments that use machine learning and natural language processing to measure cognitive abilities, language, and literacy development among children has been a growing interest in the digital age (Chaudhri, Gunning, Lane, & Roschelle, 2013). Educators and school psychologists are moving from traditional pencil and paper assessment and are developing more of an interest in technological assessments that are accessible, are easy to administer, and reflect technology familiar to children raised as "digital natives" (Jellins, 2015; Thompson, 2013). The goal of this demo presentation is to introduce Talk2Me Jr. (TTMJ) (Jang et al., 2017, 2018), which is a computerized assessment instrument designed to screen cognitive abilities relevant to language and literacy development. It is currently in development building on Talk2Me, the work of computer scientists and computational linguists at the University of Toronto that collects the speech and cognition data of healthy adults and adults with Alzheimer's disease (Noorian, Pou-Prom, & Rudzicz, 2017). In expansion of this project, TTMJ was designed for school-aged children to complete independently with minor supervision in educational settings and will be administered by educators and school psychologists as an instrument to identify cognitive and linguistic features consistent in children and to differentiate among the distinct learning profiles of children with specific learning disorders, children who are learning English as an additional language, and children who are typically developing learners.

The demo presentation will walk through the overall development process and seven tasks that compose TTMJ: 1) oral reading, 2) picture description, 3) story retell, 4) phoneme elision (Plaza & Cohen, 2003), 5) word mapping, 6) word recall, 7) a matrix reasoning (Condon, & Revelle, 2014). The TTMJ battery takes each student approximately 30-40 minutes to complete and records the children's speech throughout assessment. Following assessment, over 2,000 phonetic, syntactic, lexical, and acoustic features can be digitally extracted from the speech data. At a broader level, the tasks are designed to provide measures of phonological awareness, decoding ability, short-term memory, working memory, vocabulary, non-verbal reasoning, and comprehension. Together, these abilities are foundational to overall literacy development. For example, difficulties in short-term memory, phonological awareness, and non-verbal reasoning are associated with learning exceptionalities in children. With such data, it is our goal to be able to create distinct learning profiles for children with specific learning disorders, children learning English as an additional language, and children typically developing as learners. As each of these learning profiles is associated with specific learning needs, early identification will help children have access to early intervention and differentiated instructions necessary for academic success.

Upon completion of TTMJ and its validation, the tool will be provided to educators, school psychologists, and other interested parties free of charge for classroom assessment purposes. We are aiming for teachers to be able to automatically generate a report at the student- and class-level of assessment scores. This demonstration will provide an overview of TTMJ, its design, current evidence for validity, and recommended future applications in educational settings.

Personalized language learning as language assessment: A case study of two large learner corpora

Burr Settles, Duolingo**, Masato Hagiwara**, Duolingo**, Erin Gustafson**, Duolingo**, Chris Brust,** Duolingo

8:00 a.m. - 8:30 a.m. Location: Decatur A

Assessing and tracking learning gains over time is a generally under-researched area in language testing. While the majority of studies focus on score variability in single test administrations, much less attention has been paid to longitudinal variation (Barkaoui, 2018). At the same time, there is both evidence that low-stakes formative assessments can contribute to language skill development (Graham et al., 2015), and a growing trend in viewing personalized learning technology as a kind of formative assessment (Spector et al., 2016; William & Thompson, 2017). In other words, ongoing computer-adaptive testing is an integral part of any effective computer-assisted instruction. Yet little data exist for researchers interested in issues related to longitudinal language learning or testing, which is essential not only for understanding the sources of variability in learner performance, but also their development over time.

In this demo presentation, we will introduce two large, publicly available data sets for measuring and tracking language learners over time collected through Duolingo, an online language-learning platform. We describe the data format and features of these two complementary data sets. We also summarize key findings of more than 30 additional third-party research papers, which have re-analyzed Duolingo data to draw new insights into the cognitive processes involved in longitudinal second language development.

The Half-Life Regression (HLR) data set (Settles & Meeder, 2016) was collected to model vocabulary development over time, including not only lexical acquisition but also rates of forgetting. It contains more than 13M learning traces along with prior student-word exposure and accuracy statistics, morpho-lexical metadata, and lag times between word exposures for learners of several languages over multiple months. The Second Language Acquisition Modeling (SLAM) data set (Settles et al., 2018) was initially created for a “shared task” competition (sharedtask.duolingo.com), in which 15 research groups from nine countries participated. It contains more than 7M learning traces annotated with word-level recall accuracy for more than 6,000 students during their first 30 days of learning English, Spanish, or French. In addition to morpho-syntactic metadata, this data set also includes demographic variables about the learners themselves, as well as item response times and contextual information. While both data sets have been previously analyzed to evaluate theories and algorithms by researchers in cognitive and computer sciences, they have yet to be utilized by language learning experts or assessment specialists.

Through this demo presentation, we aim to increase interest in the use of large computer-assisted language learner corpora in language testing research. We will also highlight some of the limitations of the current data sets, and invite discussion of what additional learner data could be useful from a language assessment perspective and benefit the larger research community.

Integrating Social Justice and Student Performance Online

Noah McLaughlin, Kennesaw State University

8:00 a.m. - 8:30 a.m. Location: Mary Gay C

This demonstration will:

- Define the salient characteristics of Integrated Performance Assessment (IPA)
- Identify how IPA can engage and develop social justice
- Demonstrate an effective online IPA in situ
- Address practical steps to create online activities

Best articulated in the 2013 manual, *Implementing Integrated Performance Assessment* (Adair-Hauck et al, ACTFL), IPAs are thematic modules that include all three modes of communication: interpretive, interpersonal, and presentational. Student performances are based upon culturally authentic and engaging texts and tasks which are intercalated with dialogic preparation activities and feedback.

This structure itself has a significant simpatico with the ethical imperative of social justice, which calls for collaborative and critical reflections about cultural institutions. IPA can answer this call with activities that lead students to make informed, critical and sustained intercultural comparisons.

In a traditional course, IPA includes preparatory and feedback activities conducted in real-time, allowing for a prompt and discursive approach to scaffolded learning. This is challenging in an online modality, where students complete most activities asynchronously. I will use my own online Intermediate French Language and Culture course to demonstrate how it is possible to rise to this challenge with a thematic assessment comparing the French and US American Social Security systems.

My online IPAs begin with a video quiz that provides an overview of the assessment's content and the standards by which I evaluate student performance; I check comprehension with embedded multiple-choice questions. Students then use VoiceThread to collaboratively identify elements of language and culture useful for the given tasks. They also complete preparatory worksheets before beginning both the interpersonal and presentational assignments.

Creating an effective interpretive activity requires devising discrete comprehension questions that still cover a range of strategies; software can then provide essential immediate feedback to the student. Video chat technology makes real-time interpersonal assignments feasible. Presentational activities can be blog-style essays or videos created with Adobe Spark. Effective feedback includes encouraging students to interact with their peers' presentational performances, using rubrics to shape written remarks for individual performances, and an instructor video for the class as a whole.

With the final portion of the session, I will address audience questions about creating these kinds of activities: selecting appropriate texts and tasks, effective strategies for VoiceThread, creating a video quiz, etc.

Adapting IPA to online courses is not only possible but beneficial. IPA develops all three modes of language, and its focus on preparation and feedback encourages metacognition. Because this method engages students with culturally authentic content and tasks, it can increase awareness of issues connected to social justice, encourage collaboration and interaction, and lead students to reflect on their own values and those of others.

Development and Use of an Automatic Scoring Model for Spoken Response Test

Judson Hart, Brigham Young University**, Troy Cox**, Brigham Young University**, Matthew Wilcox**, Brigham Young University

8:00 a.m. - 8:30 a.m. Location: Henry Oliver

Evaluating a student's spoken language proficiency is a critical part of administering an intensive language learning program. As a student is enrolled in the program, the program must first determine their language ability in major skill areas including speaking before the student can be placed in classes and begin instruction. This process is further constrained by limitations in the amount of time between when a student arrives on site and the beginning of an instructional period.

Because of these limitations of time, programs will often restrict their assessments to only include dichotomous items that can be scored quickly. While listening and reading may be evaluated this way, it is tempting to bypass evaluating productive skills including speaking and writing. Such assessments are then limited in the natural language production they illicit and as a result may: 1) be perceived by students as ineffective, 2) not measure core competencies that would best fit course outcomes 3) mistakenly signal to students that the program doesn't value authentic language production. A solution to this is to supplement dichotomous item assessments with open response spoken language production tasks.

However rating these open response production tasks does create barriers of inefficiency. Open response production tasks are most commonly scored using trained human raters. Executing such rating creates complicating factors including rubric development, rater maintenance including recruitment and training and then the time and expense to execute the rating protocol. Most programs lack the overhead to execute reliable and scalable human rating. An alternative to human rating would be the use of automatic scoring models.

Brigham Young University's Center for Language Studies recently completed a development project that developed such a scoring model for open response spoken language tests. Using audio recordings for over 3,000 previously administered speaking assessments that had previously been scored by human raters on a 0-7 scale the model is calibrated to predict likely scores. The model includes the following features: articulation rate; rate of pauses; mean length of utterance, mean length of pauses; mean length of words, perplexity and content features.

The demonstration will present how this scoring model is implemented to automatically rate speaking samples collected from new students. In addition to demonstrating the scoring model in action, the presenters will discuss the elements included in the model and show how the model was refined as it was developed. The presenters will show tables and graphics that contrast inputs of time and resources between processing placement test results using human rating and processing placement test results using the automatic speech model. The presenters will also discuss plans to further improve the model and apply it to the benefit of their students and organization.

BEST Plus 3.0: Assessing Speaking Using a Multi-Stage Adaptive Test

Megan Montee, Center for Applied Linguistics, **Daniel Lee**, Center for Applied Linguistics

8:00 a.m. - 8:30 a.m. Location: Decatur A

We will demonstrate BEST Plus 3.0, a computer-based speaking assessment used to test the oral proficiency of adult English learners in the United States. The BEST Plus 3.0 is an individually administered, face-to-face adaptive oral interview designed to assesses the ability to understand and use unrehearsed, conversational, everyday language within topic areas generally covered in adult English language courses. During the assessment, test administrators read scripted interview questions and enter task scores into the computer. A computer application selects the next test question based on the examinee's performance on previous tasks.

BEST Plus 3.0 is built on an innovative multi-stage adaptive test (MST) design. In a multi-stage adaptive test, examinees proceed through the item pool in "stages," in which each stage has a fixed set of questions. The MST approach allows for greater measurement precision with fewer items than one fixed form that is administered to all examinees. It is particularly suited to cases where several items need to be administered together (in contrast to "item-level" computer adaptive testing).

In this demonstration, we will present a sample test. We will also discuss the underlying test design and show a custom-built computer application that was developed in order to support the MST design. The computer application includes a web portal that allows test users to manage test administration and score information, and also allows test developers to access test data and easily embed field test items for ongoing test refreshment. The demonstration will thus include features of the test user interface while also highlighting the behind-the-scenes components of the program that are of interest to language testers.

1. “That's a waste of time, going back and reading the text again!” – Cognitive processes in an integrated summary writing task

Sonja Zimmermann, TestDaF-Institut

1:30 p.m. to 3:00 p.m. Location: Decatur A

Academic writing typically requires students to process information from different sources and integrate this information in their own texts. Hence, language tests for university admission purposes make increasing use of integrated writing tasks – tasks that provide students with language-rich source material. Test takers are asked to write their own texts based on the information presented, transforming the language of the input material (Knoch & Sitajalabhorn, 2013). Yet, the underlying construct of this kind of tasks is still an open issue, especially when looking at integrated writing tasks from an evaluation perspective: It is unclear which factors account for the performance, i.e. to what extent writing ability, reading skills, etc. contribute to the test results. (Cumming, 2013; Weir, 2005).

This Work-in-Progress reports on a validation study that sheds light on the construct underlying the use of a summary writing task in the context of a newly developed university entrance language test in Germany. Following recent approaches (e.g., Cumming, 2014; Yu, 2013), the present study builds on a mixed-methods design to look at integrated writing tasks from process-oriented, product-oriented, and evaluation/scoring perspectives.

Nineteen international university applicants participated in a study that combined eye-tracking and stimulated recall techniques to investigate the cognitive processes test-takers engage in when writing a summary from written and graphical input. The retrospective interviews were transcribed and coded, using the framework of discourse synthesis (Spivey & King, 1989). The quantitative eye-tracking data focused on the time participants spent in different Areas of Interest (AOIs) on the screen, e.g. the source text, the instructions, or the text box when writing, as well as the transitions between the AOIs.

The aim of the presentation is twofold: First, findings from the study are used to define and specify the cognitive processes involved in summarizing from two different sources, as well as to link them to existing writing and/or text processing models. Second, the findings provide a basis for discussing to what extent the identified reading-writing relations can inform further research, in particular, product-oriented studies that aim to analyze the written performances linguistically and with respect to content, and studies that look at how these performances can be evaluated and scored.

2. Studying item difficulty. Insights from a multilingual foreign language assessment

Katharina Karges, University of Fribourg

1:30 p.m. to 3:00 p.m. Location: Decatur A

Principled item development and adequate scale interpretation rely on sufficient knowledge of the interplay between text, task and test taker characteristics. An important product of this interplay, which testers and researchers alike want to understand, is item difficulty.

One way of investigating the difficulty of specific items is to observe and/or interview the test takers, preferably during pre-operational testing. Another way, more suitable to large-scale assessment, is to identify and predict sources of difficulty by relating task and/or item characteristics to test results through statistical modelling (e.g. Buck & Tatsuoka, 1998; Carr, 2006; Embretson & Wetzel, 1987; Freedle & Kostin, 1993, 1999). These studies reveal a number of potential sources of difficulty, which are often difficult to interpret due to complex interactions between the variables (e.g. a relevant

information in the middle of the input text makes an item more difficult but this may be attenuated by lexical overlap between text and item). Also, most studies on item difficulty focus on tests of English (L1 or L2) for adults, whereas there is very little evidence for other language learning contexts (although cf. e.g. Böhme, Robitzsch, & Busè, 2010).

In my PhD project, I have the rare opportunity to study the difficulty of test items in three test versions which are translations of each other. The items were taken from a large-scale assessment that studied reading and listening competences of sixth-graders in three different foreign languages (English, French and German at A1/A2 of the CEFR (Council of Europe, 2001)). The assessment was conducted in three regions with different languages of schooling (German, French and Italian) and includes data from around 25,000 students. Potentially, the data can be complemented with data from a reading test in the language of schooling and a background questionnaire.

A comparison of the relationship between certain task and/or item features and the difficulty of an item across languages may add evidence to the general discussion surrounding item difficulty in large-scale assessments, not only for English, but also for German and French, where item difficulty has been researched considerably less. The different languages of schooling included in the study will also allow more detailed analyses of the impact of individual linguistic repertoires in foreign language assessment.

For this work-in-progress session I am mainly hoping for suggestions and comments on a) the task and item characteristics to take into account and b) the most suitable statistical modeling procedures. I am also curious to hear what people from other contexts think about the rather unusual practice of translating a foreign language test.

3. Development of Scales for the Assessment of Young Learners Functional Writing Proficiency

Gustaf Bernhard Uno Skar, Norwegian University of Science and Technology, **Lennart Joelle**, Norwegian University of Science and Technology

1:30 p.m. to 3:00 p.m. Location: Decatur A

The Norwegian project Functional Writing in Primary School (the 'FUS project') is a large-scale writing intervention project with an experiment like design aiming to increase the quality of teaching and learning writing in the first years of schooling. The intervention program will run between fall of 2019 and spring of 2021.

The intervention program targets teachers' learning and will offer a range of activities and materials for writing instruction, reflective practices, formative assessment and development of a shared meta language among colleagues. The underlying hypothesis is that students in project schools will outperform students from comparison schools.

There is a need of assessment scales to measure students' functional writing proficiency, and to derive estimates of intervention effect. However, there are currently no such scales available in Norway. A major task to complete before the intervention start is to produce and validate scales that will fit the needs of the project. The purpose of this work-in-progress paper is to present the process of scale development and validation.

The development of the FUS Scales ('FS') is divided into four distinct stages. The first two stages are completed by the time of the submission of this abstract. In stage 1 relevant areas of assessment (i.e. the scales) were identified by having a panel (N = 19) of linguists, assessment specialists and teachers reading and analyzing texts from writers in the start of 1st grade to 3rd grade. This resulted in the

following scales: (i) letter formation; (ii) usage of page; (iii) spelling; (iv) punctuation; (v) text structure; (vi) writer-reader interaction; (vii) content. In stage 2 the same panel rank-ordered texts (N = 400) using comparative judgment (CJ; Pollitt, 2012) as integrated in the software No More Marking (No More Marking, n.d.), answering the question: which of the two texts is the better one. This in turn resulted in texts ranked on a scale of 0–100. This was used to divide the texts into five piles – ranging from low rank (0–20 points) to high rank (80–100 points).

The next two steps, which will be completed or in end phase at the LTRC 2019, is as follow: Stage 3 is to first describe characteristics of scripts in each pile for each scale. This will result in five step scales that can be used when measuring the effect of the project. Stage 4 includes piloting the scales in order to generate data for the validation process. The abovementioned panel and additional teachers (N = 10) will use the scale to rate 300–400 new student texts. As a major part of the validation of the scales (cf. Knoch, 2009), the ratings will undergo many-facet Rasch measurement analysis (Linacre, 2017) to ensure that the scales are able to produce ratings that differentiate between levels of performance, and have scale steps that functions properly.

Knoch, U. (2009). *Diagnostic Writing Assessment*

Linacre, J. M. (2017). *A user's guide to FACETS*.

No More Marking (n.d.). www.nomoremarking.com

Pollitt, A. (2012). Comparative judgement for assessment. *International Journal of Technology and Design Education*, *22*(2)

4. Investigating the Interactiveness of IELTS Academic Writing Tasks and Their Washback on EFL Teachers' Test Preparation Practices

Parisa Safaei, Université Laval, **Shahrzad Saif**, Université Laval

1:30 p.m. to 3:00 p.m. Location: Decatur A

Research on high-stakes testing has shown that when test scores are used for selection purposes, candidates exert the necessary efforts to gain the language abilities required by the test (Green, 2007). Some of these skills are desired, as they prepare the applicant for functioning in the Target Language Use (TLU) domain. However, certain skills required for passing a test are not necessarily reflective of the TLU context (Bachman & Palmer, 2010). This overlap between test success and the TLU requirements lies in test validity, and positive washback is thus constituted in the extent to which success on the test reflects preparedness for the TLU context (Saif, 2006). However, washback research has confirmed that there are factors other than the test design, i.e. contextual factors, which determine the intensity and direction of washback (Alderson, 2004; Cheng, 2004). Among these, the teacher factor merits further attention, as it is the teachers' perceptions of and attitudes towards the test that shape students' awareness and directly affect test preparation (Bailey, 1996; Watanabe, 2004). Teachers make pedagogical and ethical decisions about how to teach their high-stakes test preparation classes to facilitate learning and promote optimal test results simultaneously (Spratt, 2005). With two million administrations per year, IELTS is a high-stakes test used by 8000 organizations worldwide as a gate-keeping tool for immigration, academic, and vocational purposes. Preparation for IELTS in most EFL contexts is the test-takers' main opportunity to understand language requirements of the target context. Previous studies have raised concerns about the complexity of the washback effect of the IELTS score, but their results are limited in scope due to the dependence of washback on contextual factors (Ingram & Bayliss, 2007).

Adopting a mixed-methods research design, this project investigates interactiveness, an under-researched aspect of test validity, and the effects of the test tasks on the teachers' test preparation practices in IELTS academic writing courses in an EFL context. The study looks into the reality of academic writing in real-life contexts and examines the degree of consistency between the test content/tasks, target requirements, and test preparation content. It further examines the teacher factor to study IELTS' washback effect on teachers' choice of materials, class activities, and teaching methodology in their test preparation courses. To this aim, 20 teachers and 200 students will complete questionnaires and the participating teachers will be observed and interviewed. A rigorous qualitative task analysis will also be conducted in order to evaluate the extent of interactiveness in IELTS test tasks. Through data convergence, the mediating role of the teachers in the degree and direction of IELTS writing washback will be examined. The findings are expected to shed light on teachers' classroom performance and their beliefs about the test's writing tasks and the extent to which they represent students' needs. This project is in the data collection phase. During the work in progress session, we will present the results of the initial phase (task analysis) and seek feedback on the subsequent phases of the project.

5. The methods dealing with dependent effect sizes in a meta-analysis: a review in reading research area

Jingxuan Liu, Georgia State University, **Xiaoyun Zhang**, Georgia State University, **Hongli Li**, Georgia State University, **Xinyuan Yang**, Texas A&M University

1:30 p.m. to 3:00 p.m. Location: Decatur A

Meta-analysis is a statistical method to combine quantitative research results across studies. One challenge in conducting a meta-analysis is how to deal with multiple effect sizes from a single study because they are not independent. Ignoring such dependency biases meta-analysis results. Meta-analysis has been widely used in reading research to synthesize the effectiveness of reading strategy, intervention, testing mode and so on (e.g., Swanson, 1999; Therrien, 2004; Wang, Jiao, Young, Brooks, & Olson, 2008). Understanding how to appropriately deal with dependent effect sizes is important for the validity of meta-analysis in reading research. The present study will review meta-analysis in reading research during the past five years and demonstrate how to appropriately model dependent effect sizes.

Traditionally, meta-analysts may select a single outcome over other outcomes or aggregate multiple outcomes from a study (Cooper, 1998). As a result, the number of effect sizes included in the meta-analysis could be greatly reduced. In addition, one may include multiple outcomes from a single study as if they were independent. However, ignoring such dependency may bias the meta-analysis results (Scammacca, Roberts, & Stuebing, 2014). Methodologists have proposed more advanced methods to treat dependent effect sizes via modeling the correlations among outcomes. If the effect sizes can be clearly identified as nested structure, multilevel meta-analysis can be used to model the dependent effect sizes (Van Den Noortgate & Onghena, 2003). This method, however, requires sufficient sample sizes on each level. In addition, Kalaian and Raudenbush (1996) described the use of multivariate modeling to address multiple effect sizes from a single study. However, this method requires a covariance matrix among the multiple outcomes, which is usually not available. More recently, Hedges, Tipton, and Johnson (2010) proposed robust variance estimation (RVE) method to account for the dependent outcomes by adjusting the standard errors of the effect sizes. This approach does not require the input of a covariance matrix among the multiple outcomes and can be easily implemented with an R package. The advantage of the RVE approach is apparent (Scammacca et al., 2014), and more efforts are needed to promote its use in practice.

There are two purposes for the present study. First, we will investigate how researchers in reading research area dealt with dependent effect sizes in their meta-analysis studies in the past five years (2013-2017). Using keywords “Reading” and “Meta-Analysis,” we will search PsycINFO, ERIC, and ProQuest Dissertations & Theses A & I database for eligible studies. We will present the frequency of using each of the six methods to deal with dependent effect sizes. We will also examine if there are any patterns related to researchers’ choice of a particular method to deal with dependent effect sizes, such as time frame, types of effect size, regions of researchers. Second, we will offer a tutorial on how to use RVE method with a real dataset using Robumeta package in R (Fisher & Tipton, 2014). This study will promote the application of RVE method and help researchers deal with dependent effect sizes in meta-analysis appropriately.

6. Towards the Democratisation of the Assessment of English as a Lingua Franca

Sheryl Cooke, University of Jyväskylä / British Council

1:30 p.m. to 3:00 p.m. Location: Decatur A

Automated assessment of spoken language presents an opportunity to make language testing more just and fair by removing potential bias associated with human raters. However, if automated systems use standard varieties of language or learn from big data that is not free of bias, technological advances threaten to reverse the tolerance of pronunciation variance between the range of accents that make up English as a Lingua Franca (ELF). Rather than moving towards a more inclusive paradigm for measuring the pronunciation facet of communicative competence, accentedness and nativeness could take precedence over intelligibility. The use of standard or prestige varieties of English as pronunciation yardsticks would have far-reaching social consequences.

This work-in-progress outlines a research plan driven by the questions: Which pronunciation features should be prioritised or deprioritised in L2 pronunciation assessment in an ELF context? and How can technology be leveraged to reflect the intelligibility construct of ELF? The research comprises three main phases and intends to use technology as a tool to investigate whether the assessment of ELF pronunciation can be more democratic. Phase one starts with a study of accent bias amongst raters; phase two proposes an alternative methodology for generating a core phonological inventory (Jenkins, 2000) based on intelligibility of and by a range of ELF speakers using crowd-sourcing to gather a sample that is both adequate and representative; phase three will focus on a construct-driven, technology-assisted analysis of the core features of intelligibility of ELF. The focus of the intended research project is on producing a worked example of a theorised approach to generating a more representative model for the assessment of ELF pronunciation.

McNamara (2014, p. 227) takes the position that the current convergence of technological and sociolinguistic factors place language assessment “at a moment of crisis”. If language testers are to heed Shohamy’s call for a stronger democratic and ethical approach to language testing (1998), now is the time to act to prevent the encoding of bias or a narrow range of prestige Englishes to the detriment of other varieties and their speakers.

7. Understanding Young Learners' Spoken Academic Language Development through Analyzing Oral Proficiency Test Responses

Megan Montee, Center for Applied Linguistics, **Mark Chapman**, WIDA

1:30 p.m. to 3:00 p.m. Location: Decatur A

In this session, we present the design of a research project to analyze the academic spoken language of English Learners (ELs) in the context of U.S. public schools. The project explores the linguistic features of speaking test responses from ELs in grades 1-12 at a variety of English language proficiency levels and across task types. Data from the research study will come from students' audio recorded responses to a large-scale, computer-based speaking test designed to measure academic English language development. The research will address the following questions:

How do the linguistic features of young learners' spoken academic English differ across grade levels? English language proficiency levels?

How do the linguistic features of young learners' spoken academic English differ based on task features?

Speaking test responses will be transcribed and analyzed for a variety of linguistic features, including discourse complexity, vocabulary, grammar, and oral fluency measures. The study will feature multiple task types, including recount, explanation, and argumentation tasks that address the language of different academic content areas (language arts, social studies, science, and math). The study builds on previous work analyzing the spoke discourse of young learners (Hsieh & Wang, 2017) and expands limited research about this population. In particular, there is limited data comparing learners spoken language features across age groups. The results of the study promise to provide exploratory descriptive information about the language features of spoken academic language of young learners. The findings of the study will have implications for both task development, to assist in producing tasks that present students equivalent opportunities to demonstrate their academic spoken language proficiency, and how students develop academic spoken language proficiency in different content areas.

The project is in the initial design and conceptualization phase. First, we will present the data collection and analysis plan, including a description of the distribution of language samples across grade, proficiency level, and task type. Next, we will present example data including transcriptions of student speech for a variety of grade levels and tasks in order to provide concrete examples for discussion and feedback. Finally, we will present pilot coding and analysis data in order to gather feedback from participants and generate discussion using examples. During the session we will elicit feedback from participants about all phases of the research design and analyses in order to improve the project.

8. Validating the use of a web-based rating system for oral proficiency interviews

Jing Xu, Cambridge Assessment English, **Anne Clarke**, Cambridge Assessment English, **Andrew Mullooly**, Cambridge Assessment English, **Claire McCauley**, Cambridge Assessment English

1:30 p.m. to 3:00 p.m. Location: Decatur A

The Speaking exams of Cambridge English Qualifications are paired oral proficiency interviews in which two candidates are rated by two examiners present in the test room using mark sheets. The interlocutor global score and the assessor analytical scores are weighted and combined into a Cambridge English Scale score interpretable in the Common European Framework of Reference for Languages (CEFR). In this fast-paced digital age, not only is paper-based rating logistically complex and inefficient but it also rules out the option for a test centre to utilise examiner resources in geographically remote areas. Although replacing paper-based rating with online digital rating seems a simple and intuitive solution,

validity issues associated with the utilisation of technologies for remote assessment of speaking remain relatively underresearched (Nakatsuhara, Inoue, Berry, & Galaczi, 2017; Yang, 2016).

This paper reports a study that evaluated the assumptions underlying the 'domain definition', 'evaluation', and 'generalisation' inferences in an interpretation/use argument (Kane, 2013) for online digital rating of Cambridge English Qualifications Speaking exams. A web-based rating system was developed for interlocutors to enter global scores locally and for remote assessors to perform analytical ratings on video recordings of the tests asynchronously. The study addressed four research questions: 1) Overall, do the standard paper-based rating mode and the new digital rating mode result in different test scores? 2) Are there interaction effects between the rating mode and the country where a test is administered or between the rating mode and a candidate's level of oral proficiency? 3) Does rating on a tablet have an impact on examiner behaviours? 4) Does the presence of a recording device and absence of a local assessor affect candidate behaviours?

Based on a counterbalanced repeated measures design, 101 adolescent English learners from four different countries (Portugal, Italy, Switzerland and Argentina) and of five proficiency levels (A2 to C2) sat both the standard Cambridge English Qualifications Speaking exams and the counterparts supported by the web-based rating system. A small group of examiners and candidates were interviewed immediately after testing. Remote assessors were observed during rating and interviewed on completion of the job. To investigate the effect of the rating mode on test results and possible interaction effects, a 3-way repeated measures analysis of variance (ANOVA) was performed. The qualitative data are being transcribed and thematically coded in the MAXQDA computer program.

The study found no significant effect of the rating mode on the test results but did find a mode-proficiency interaction. Specifically, high-performing C2 candidates tended to receive a lower Cambridge English Scale score in the new digital rating mode than in the standard paper-based rating mode. Initial analysis on the interview data suggested that the rating system was intuitive to examiners and that losing visual input (due to technical glitches) had a negative impact on remote analytical rating. Slight differences in interlocutor behaviours were also observed between the two rating modes. The implications for revising the interlocutor script for the digital rating mode will be discussed.

9. Test preparation materials for the Test of Workplace Essential Skills (TOWES): Validating materials for adult literacy and numeracy

Claire Elizabeth Reynolds, Carleton University

1:30 p.m. to 3:00 p.m. Location: Decatur A

Essential skills are the foundational elements required to learn the technical skills needed for most occupations (Essential Skills Research Unit, 2015). The workplace essential skills: reading, document use, and numeracy, can be assessed using the Test for Workplace Essential Skills (TOWES). The test generates three sets of results in the form of competency levels, one for each workplace essential skill. The competency levels, determined by the Government of Canada, range from level 1 (the lowest) to level 5 (the highest) (Essential Skills Research Unit, 2015). Test-takers can use the TOWES results as an indicator of their employability. Level 3 is the point at which technical skills can usually be acquired to gain stable employment (Statistics Canada & OECD, 2003). As a result, there are a plethora of training opportunities for workers who have attained level 3. However, for those who score level 1, there are few options for improving their WES (Murray, 2010). To address this lack of training, I teamed up with an online workplace assessment and planning company to create TOWES preparation materials. Using a blended classroom setting (Gruba, Carndenas-Claros, Suvorov & Rick, 2016), I have designed and developed

online tasks and face-to-face, hands-on mini-lessons using the variables Kirsch & Mosenthal (1990) identified as factors in adult literacy and the workplace essential skills criteria provided by the Essential Skills Research Unit (2015), a section of the Government of Canada.

These tasks and mini-lessons are overseen by a trained facilitator who tracks the workers progress assisting them when necessary. The tasks and activities are scheduled to be piloted with a group of workers in mid-2019. The objective of the pilot is to validate the tasks and activities through a convergent mixed methods study (Creswell, 2009) using activity theory (Engeström, 1987). The quantitative phase will be a quasi-experimental design (Lynch, 1996) with the tasks and mini-lessons as the treatment and the TOWES as a criterion reference test (Brown, 1989). The qualitative phase will have an ethnographic element (Hammersley & Atkinson, 2007) as I participate as a facilitator and subsequently write a reflective journal of events that pertain to the research (Kovach, 2009). Narrative inquiry (Polkinghorne, 1997; Costantino & Greene, 2003) will also be used by having, 5-8 participants share their experiences with the TOWES preparation materials. Narrative analysis (Polkinghorne, 1995) will be applied to their stories and to the journal entries. The two phases will then be merged to find patterns among the themes from the narrative analysis and the results of the quasi-experimental study. The findings of this study will contribute to the literature on the essential skills by providing insight into the TOWES preparation process and preparation materials. As this study design is a work in progress, I would welcome and appreciate the opportunity to share this research and receive feedback from my peers.

10. Students' and Teachers' Perception and Use of Diagnostic Feedback

Hang Sun, Shanghai Jiaotong University

1:30 p.m. to 3:00 p.m. Location: Decatur A

In recent years, due to the increasing influence that language tests play in various aspects of society, test fairness and social justice has become a primary concern in the realm of language assessment. However, studies on fairness and justice have been exclusively focusing on high-stakes standardized assessments which are conceived to have larger impact on society. Research which investigates what test fairness and justice entails in terms of low-stakes assessments is still a scarcity. At the same time, with the increased recognition about the washback of language tests, the movement of formative assessment as well as the development of cognitive psychometric models, diagnostic assessment has attracted more and more attention because of its potential to provide fine-grained feedback of test-takers' strengths and weaknesses. Diagnostic assessment has pointed out a fruitful and even heartening future to promote fair use of tests due to its potential to shape assessments "for learning", rather than "of learning". Also, the power of diagnostic tests is shared with teachers and students, who are the decision-makers and action-takers instead of passive recipients of test consequences.

As diagnostic feedback is a defining characteristic of diagnostic assessment (Alderson, 2005) and the ultimate users of diagnostic feedback are students and teachers, this study aims to investigate how students and teachers perceive and use diagnostic feedback of an online tertiary-level diagnostic reading test and the consequence of the test. Drawing on an argument-based approach (Chapelle et al., 2008; Knoch & Chapelle, 2017), this study focuses on two inferences of test interpretation and use – decision and consequence. Taken from a larger project that builds a validity argument for the test system, the present study sets out to explore whether and in what ways diagnostic assessment can promote test fairness by investigating students' and teachers' perception and uptake of diagnostic feedback and the corresponding consequences.

The study first articulates the inferences, warrants, backing and rebuttals needed to justify the respective arguments. Then it involves qualitative analysis of interview data collected from students (n=8) who took the diagnostic reading test and their teachers (n=3). The interviews were conducted twice – right after students and teachers received the feedback report and three and a half month later – to track their perceptions and actions. The interview data were then coded and categorized to form the backings or rebuttals to justify or refute the assumptions of the inferences. The findings illustrate whether and in what ways the ultimate goal of diagnostic feedback to promote tailored learning and teaching has been met. The current research serves as an explorative investigation to address the theoretical and practical concerns about diagnostic feedback as well as adds to the meager literature on test fairness and justice of low-stakes learning-oriented assessment.

11. Effects of Reader and Task Variables in L2 Reading Comprehension and Speed

Toshihiko Shiotsu, Kurume University

1:30 p.m. to 3:00 p.m. Location: Decatur A

A number of studies have compared the effects of reader variables such as decoding skills in L1 reading (Garcia and Cain, 2014) and lexico-grammatical knowledge in L2 reading (Jeon and Yamashita, 2014) to identify factors that can explain the individual ability differences. Research on reading comprehension tests has shown results pointing to the effects of text characteristics and response formats (e.g., Kobayashi, 2009), but it is still far from clear how they relate to both reading comprehension and speed, the latter of which is in particular need of empirical data compared to the former. With the increased awareness of the necessity for fluency and automaticity development in language performance (e.g., Grabe, 2010), more investigative efforts on the relationship between test tasks and reading comprehension speed can lead to significant implications.

In addition to test-taker characteristics that explain the individual differences in reading comprehension and/or speed, this study in progress is designed to address task characteristics which account for any within-group differences in comprehension and/or speed.

A total of 125 EFL learners, ranging in CEFR level A to C, participated in the study and their reading comprehension and speed were assessed on a computer-based reading test, which had two passages (A and B) followed by comprehension questions, either in multiple choice (M) or short answer (S) format. The participants were randomly assigned to one of eight groups differing in passage order and question format. Their reading speed was recorded by utilising the word-by-word text presentation on the moving-window paradigm, and no regressive reading was allowed. The participants' CEFR level was estimated based on their performance on a test reported to sample widely from the construct of reading (Brunfaut and McCray, 2015).

The analyses so far have focused on the effects of test-taker attributes on their reading test performances, and the results have indicated that, in addition to the expected significance of lexico-grammatical knowledge in explaining the reading level differences, a non-negligible relationship between L2 reading strategy use and the estimated reading ability.

The study is at the stage of processing the reading time data, and further results should be available at the presentation.

12. The Effect of Genre on Linguistic Features of Source-Based Essays by Tertiary Learners: Implications for Construct Validity

Sukran Saygi, Middle East Technical University, **Zeynep Aksit**, Middle East Technical University

1:30 p.m. to 3:00 p.m. Location: Decatur A

When L2 learners' written ability is assessed, how their performance is assessed depends on a complicated interplay of factors. The nature of the tasks and raters' behaviors are among such factors. When they need to show proof of English language ability to be admitted to programs offering English-medium education, examinees are given only one or two tasks to show their writing abilities due to practical reasons such as limited time and high costs of rating. Literature on L2 writing performance reveal that effects of task type or topic have been explored; however, not much is known about the effects of genre on students' writing performance (Kim, 2016). In an attempt to address this issue, this study aims to investigate (1) the potential effect of two genres (i.e. compare/contrast and problem solution) on examinee performance, (2) to what extent the linguistic features of these essays are affected by these genres.

The data for this study were collected in a university where the medium of instruction is English. Therefore, the students need to show proof of proficiency in English language. The university administers an in-house proficiency test four times a year (in December, June, August and September). The Performance Task, with a weighting of 20 points, requires the test takers to collocate information from two sources – a lecture and a reading passage. The source texts may be on advantages and disadvantages of something, problems and solutions to them, or principles of something and examples of these in practice. The readability scores and grade levels are checked to make tests comparable in difficulty. The test takers are required to pinpoint the relationship correctly, and collate the relevant information from the source texts in around 300 words.

For this study, essays written by 50 examinees in the exam given in August 2018 and in September 2018 were analyzed. In the first exam, the source texts had opposing views. In the second exam, source texts were on problems and solutions to these problems. The essays were scored by two raters, and a third one if need be. The scores that the examinees received for their responses have been collected from the administration. A paired-sample t-test will be conducted in order to check whether there is a significant difference between the examinees' performances in two tasks. Second, the examinees' responses will be transcribed and analyzed using Coh-Metrix computational tool assess the quality of language use. Finally, using Weigle and Parker's (2012) framework, textual borrowing strategies such as quoting, referencing, copying, or reformulating will be specified and counted per t-unit in each text. The findings of the study will guide the design processes of this integrated task.

13. Investigation of Social Justice Violation in an English Proficiency Test for PhD Candidates in Iran

Masood Siyyari, Islamic Azad University, Science and Research Branch, **Negar Siyari**, University of Arizona

1:30 p.m. to 3:00 p.m. Location: Decatur A

This paper reports on an on-going study on the uses and misuses along with the consequences of using an English language proficiency test at the PhD program of the largest Iranian private university (i.e. Islamic Azad University aka IAU) with many branches in Iran and a few foreign countries. This test, called EPT (English Proficiency Test), is probably the most worrisome name to any PhD candidate at IAU, which

is supposed to be passed with the cut-score 50% by PhD candidates of most majors as a requirement for the comprehensive exam. Although started with good intentions in the beginning (i.e. making the students improve their English proficiency), this test apparently never meets the needs and expectations of most parties involved including students and professors, and it has become the most controversial issue among PhD candidates raising a lot of disagreement. Given these points, this study attempted to investigate to what extent social justice is met/violated by means of EPT. To answer this general question, some detailed questions related to different dimensions and facets of tests which may have relevance to social justice and fairness were devised based on (a) the available literature on this issue, (b) expert opinion, and (c) test users and takers' opinion (administrators, authorities, professors & students). These questions were in general concerned with issues such as (a) accounts of personal experiences of EPT users and takers giving details of its pros and cons, (b) solutions suggested by EPT users and takers, (c) justification for EPT, and (d) EPT technicalities such as validation and standardization, standard setting etc. A major part of the data on the above questions have so far been collected through questionnaires and interviews with EPT users and takers (155 test takers, 5 university officials, 21 professors, 10 EPT preparation course instructors, 12 professors in applied linguistics with testing specialty), and observations of EPT preparation classes (10 classes). Preliminary descriptive and inferential statistics on the questionnaire data and content analysis of the interview and observation data so far have shown that this test, albeit the base for critical decisions, violates social justice specifically due to its weak validation and administration procedures. Moreover, the questionnaire and interview data analyses indicate that university-major bias (a possible source of social justice violation) seem to exist in this test; that is to say, students from some specific majors seem to have higher chances of giving correct answers esp. to reading comprehension questions. Therefore, currently real EPT data is being collected to have differential item functioning (DIF) analysis to statistically check whether any DIF really exists in terms of university major or any other factor.

14. Developing a Digital Simulation to Measure L2 Intercultural, Pragmatic and Interactional Competence: Initial Pilot Results

Linda Forrest, Center for Applied Second Language Studies, University of Oregon, **Ayşenur Sağdıç**, Georgetown University, **Julie Sykes**, Center for Applied Second Language Studies, University of Oregon, **Margaret Malone**, Georgetown University

1:30 p.m. to 3:00 p.m. Location: Mary Gay C

Comprehensive instruments to measure language learners' intercultural, pragmatic, and interactional skills are virtually non-existent, even though the ability to use language effectively in highly nuanced social encounters is essential to successful communication (see Roever, 2013 for a related exception). Complexities, including language variety, individual preferences, and the difficulty of creating efficient assessments, contribute to the void, yet they do not eliminate the need to measure students' abilities in this area. This presentation reports on early piloting results from a formative assessment tool that provides comprehensive feedback to learners about their ability to navigate the complexities of everyday interactions in the L2. Utilizing the assessment can enable learners and their instructors to focus on and understand meaning in variety of social, professional and academic contexts, thus empowering them to engage successfully in authentic, multilingual, intercultural interactions.

Previous research shows that pragmatic competency is especially difficult to measure due to learner subjectivity, lack of theoretical support, and immense variation in dialect, sociolect, and idiolect (Bardovi-Harlig, 2001; Félix-Brasdefer, 2007; LoCastro, 2003). Based on related work (e.g., Sykes, 2010, 2013, 2016; Thorne, Black, Sykes, 2009), this presentation reports pilot data for the first implementation

of an innovative digital simulation designed to assess L2 pragmatic knowledge, analysis skills, learner subjectivity, and emotional awareness. Drawing on three critical design principle 1) accommodating individualized experience validated across learners, 2) examining a set of macro-level skills to interface with other linguistic abilities, and 3) facilitating language variation – this simulation delivers authentic scenarios in an immersive digital environment to enable the measurement of L2 intercultural, pragmatic, and interactional abilities. This method also approaches the interaction from a lens of co-construction rather than one of fixed, rigid pragmatic norms and expectations. By using lifelike scenarios with varying degrees of social and individual factors (e.g., gender, social distance, power), learners' abilities to interact in culturally appropriate ways can be observed in simulated, but realistic and replicable situations.

This presentation reviews the scenarios that have been developed and discusses initial findings from pilot testing in English and Spanish (n=30), indicating the challenges and opportunities for using digital simulations to assess pragmatic abilities with a diverse group of test takers. Data collected include examinee interactions with a digital avatar and their responses to multiple-choice and open-ended reflection questions related to the scenario. Follow-up verbal protocols gathered in-depth information on participant's cognitive processes while taking the simulation, as well as overall feedback and suggestions for future development. As such, the measure not only looks at learner's knowledge, but ways to evaluate a comprehensive pragmatic repertoire, including the choices they make. The qualitative data will be used to validate the scenario and the assessment tasks from the simulation; it will also help to elicit potential scenarios from the participants. These results will help revise the instrument in terms of its scenarios, questions, user interface, and overall organization.

15. Using unscripted spoken texts in college-level L2 Mandarin assessment

Xian Li, Georgia State University

1:30 p.m. to 3:00 p.m. Location: Mary Gay C

Recent literature has studied unscripted text use in English as a Second Langauge (ESL) and Foreign Language (FL) classrooms, but few focused on unscripted spoken text in FL listening assessment. Wagner (2014) investigated the use of unscripted listening texts in ESL classrooms and suggested that using natural, unscripted texts could improve learners' communicative competence. Additionally, Wagner and Toth (2014) conducted a study of Spanish language learners' test performance using scripted and unscripted listening texts. Their results showed that test takers scored lower with unscripted listening input, suggesting that more instructional support with naturally spoken texts in FL classrooms might benefit students' listening comprehension.

This work-in-progress study aims to examine the role of unscripted text in FL Mandarin listening assessments and FL discourse knowledge through unscripted listening tests. Similar to spoken English, spoken Mandarin contains discourse markers such as "en" or "hao" that do not contain any substantial meanings (Liu, 2009). For example, the word "hao" in Chinese means "good" and is one of first words Mandarin learners acquired. The teacher might say "hao" in class to signify the beginning of a new activity, but students might confuse it with the meaning "good" and mistake it for positive feedback.

To fill the gap in unscripted text use in FL assessments, the researcher proposes two research questions: 1) How do unscripted spoken texts influence students' test performance in Mandarin courses? 2) How do unscripted spoken texts influence students' FL Mandarin discourse knowledge? The procedure of this work-in-progress study will be similar to that of Wagner and Toth's (2014) study. Students (N= 25 -35) from two intermediate Mandarin classes in a public university will participate, and they will take three different listening tests. Two different sets of audios will be produced by native Mandarin speakers with

near-identical content, length and speed. Class A will only listen to scripted recordings that are similar to the textbook listening exercise, and Class B will only listen to unscripted recordings that contain various Mandarin discourse markers. Students in both classes will complete five multiple-choice questions on the comprehension of each text. Students in Class B will answer two additional questions on the meaning of some discourse markers from the unscripted listening tests.

The researcher will conduct ANOVA tests of both classes' test scores as well as a qualitative analysis of different factors that affect students' test performance. The results of initial findings will be discussed at the presentation. The researcher will also share some implications that are helpful to FL listening assessment, FL listening instruction and FL pragmatics assessment.

References

Liu, B. (2009). "Chinese discourse markers in oral speech of mainland Mandarin speakers", in *Proceedings of the 21st North American conference on Chinese linguistics* (NACCL-21) (Vol. 2, pp. 358-374).

Wagner, E. (2014). Using unscripted spoken texts to prepare L2 learners for real world listening. *TESOL Journal, 5*, 288 –311.

Wagner, E., & Toth, P. (2014). Teaching and testing L2 Spanish listening using scripted versus unscripted texts. *Foreign Language Annals*,*47*, 404–422.

16. LAL: what is it to teachers in their classrooms?

Sonia Patricia Hernandez-Ocampo, Universidad de los Andes/Pontificia Universidad Javeriana

1:30 p.m. to 3:00 p.m. Location: Mary Gay C

Language assessment literacy (Inbar-Lourie, 2008, 2017; Malone, 2013) has been in the spotlight for the last decade. A discussion has arisen regarding who—and to what extent—should be LA literate, whether it should be the testing professionals only or if language instructors should be included, or whether it concerns school administrators and language policy makers (Jeong, 2013; Pill & Harding, 2013; Scarino, 2013; Taylor, 2009, 2013). Furthermore, the need to set a knowledge base of LAL has been stated (Inbar-Lourie, 2013). My doctoral dissertation aims at describing this knowledge base of LAL that should be part of the curricula in language teacher education programs in Colombia, and then compares this description to what actually happens in teacher programs.

This WIP seeks to have feedback from attendees to the LTRC Colloquium regarding the completeness of my theoretical framework as well as the appropriateness of the instruments I have designed to collect information about LAL in teacher programs. Theory comprehends the foundations of language testing (AERA, APA, & NCME, 2014) and how they can be adapted to the classroom context (Brookhart, 2003, 2011; Chapelle, Enright, & Jamieson, 2010; Moss, 2003); classroom-based assessment and its domains (Black, Harrison, Lee, & Marshal, 2003; Black & Wiliam, 1998; DeLuca, Valiquette, Coombs, LaPointe-McEwan, & Luhanga, 2018; Hill & McNamara, 2011; Kingston & Nash, 2011; T. McNamara, 2001); and the nature of the different English language skills. The instruments I have designed are interviews to language and content (course on evaluation) teachers, think-aloud protocols and observations to language teachers, focus group and questionnaire to pre-service teachers, and content analysis to the courses (both evaluation and language) syllabi.

17. A case study of evidence-centered exam design and listening: Washback from placement test development to program tests, KSAs, and pedagogy

Gerriet Janssen, Universidad de los Andes, **Olga Inés Gómez**, Universidad de los Andes

1:30 p.m. to 3:00 p.m. Location: Mary Gay C

One of the most important applications arising from argument-based validation is the concept of evidence-centered design, or ECD (Mislevy, Almond, & Lukas, 2003; Riconscente, Mislevy, & Corrigan, 2016), which places validation at the center of test development. As a broad summary of ECD, test developers define the domains to be assessed (domain analysis), describe both the claims to be made about test-takers in these domains (domain modeling) and the uses these claims will have. Then, content specifications are made (conceptual assessment framework), and test items are developed using relevant formats (implementation; delivery). Though the test development process has been described explicitly by Mislevy and colleagues—especially in terms of how the extant knowledge found within a language program feeds test development (e.g., theories of knowledge and performance)—, and while the effects of testing washback on language programs and pedagogy has been a concern since the 1990s, the following intersection has been less explored: in ECD, how can test developers best take advantage of forces of washback from test development to foment cycles of program development/evaluation and professional development?

In this project, test developers position washback as an exam use, using the newly introduced listening construct (to our exam and language program, that is) to build a case study describing washback on language program development/evaluation and professional development. This WIP will briefly present what we have carried out to date: the principal use the exam has with test takers ("to determine the readiness of a University student to understand the spoken and written input of an undergraduate-level content course taught in English"; N.B. in our non-English predominant university, productive skills are seen as secondary); a revision of the current placement exam and its shortcomings based on program information; descriptions of new test domains and pilot results gathered over four test administrations (n = 3000). Then, program change is documented qualitatively by comparing pre- and post-versions of lists of KSAs; changes in pedagogical practices are documented by comparing pre- and post- versions of course level exams in light of discussions about the placement exam findings; changes in teacher knowledge are documented based on questionnaire answers; changes to the placement exam based on these feedback cycles is also documented. This WIP asks participants to explore how to best continue developing washback that brings program elements—the placement exam, the exams at different course levels, the KSAs, and the pedagogies being used—into a more coherent alignment, as drawing strong connections between these different program elements is an important opportunity to democratize language program development, with all levels of instructors coming together to consider and provide feedback and direction about program goals, test domains, and item development (i.e., language pedagogy).

18. Development of an ITA Assessment Instrument based on English as a Lingua Franca

Heesun Chang, University of Georgia

1:30 p.m. to 3:00 p.m. Location: Mary Gay C

This study presents a research plan to develop an improved assessment instrument for international teaching assistants (ITAs) using an English as a Lingua Franca (ELF) theoretical framework. Data are drawn from videorecorded 10-min mock-teaching presentations, each followed by a brief Q&A session, that were made for the purposes of ITA assessment at one large southern U.S. university. ITA assessment at this institution currently uses the ITA Test (Smith et al., 1992/2007), a measure developed based on how closely the speech of ITAs matches that of a native English speaker. A previous study (Chang, 2018) indicated that pronunciation and grammar skills are the most difficult domains in this instrument, and other studies have suggested that adult learners may never develop native-like intuitions for these skills (e.g., Abrahamsson, 2012).

Accordingly, a new instrument is proposed that will better align with ELF and World Englishes scholarship which argues that forcing non-native speakers to conform to a monolingual native speaker model is not valid (e.g., Canagarajah, 2006; Jenkins & Leung, 2017), particularly in a globalized world where much of the English communication occurs in multilingual contexts. Findings from the proposed project will shine a more empirical light on validity issues in the ITA Test and the development of an ELF-oriented assessment.

Following Hall's (2014) work, the new proposed instrument will focus more on what ITA candidates can do with English, rather than on how they use English. Similar to the ITA Test, it will be designed to assess examinees' videorecorded mock-teaching. However, the domains and rubrics of the proposed instrument will assess the ability to teach subject knowledge (or concept) based on ELF, and the potential domains to be assessed involve skills to clearly convey meanings and to sufficiently explain concepts. The instrument, including the rating scales, will be developed and validated under the Rasch measurement theory (i.e., Many-facets Rasch model) (Engelhard, 2013; Engelhard & Wind, 2017). Specifically, the domain hierarchy and rating scales will be designed in alignment with the continuum of the construct (i.e., Wright map), and they will be empirically validated in terms of various psychometric qualities, such as the psychometric dimension of construct, domain difficulty, potential biases across different subgroups of examinees through differential item functioning analysis, and rating scale functioning.

The domains and rating scales will be validated using diverse raters, who will be recruited to test the instrument by grading video-recorded mock-teaching presentations of ITA candidates. Undergraduate students, one of the most important stakeholders of ITA assessment because they are frequently instructed by ITAs, will be invited to participate in the validation process as raters. Potential rater biases (e.g., domestic students vs. international students) will be explored across different domains and subgroups of examinees. The presentation will conclude with implications and expected challenges in the process of developing the ITA assessment instrument based on ELF.

19. Applying a summative assessment speaking test for formative assessment gains: The case of a computerized speaking test in Israel

Tziona Levi, Ministry of education, Language Education, Israel, **Ofra Inbar-Lourie**, Tel-Aviv University

1:30 p.m. to 3:00 p.m. Location: Mary Gay C

Assessment of speaking has been part of proficiency tests for over a century, evolving from controlled assessments to a direct oral proficiency interviews (OPI) imitating real-life settings and different language uses. In the last two decades computer technology is increasingly utilized to enhance oral language proficiency (OLP) and to address test qualities of reliability, validity and practicality. This development has fundamental implications for how the speaking construct is understood, applied and assessed.

The current study is carried out in Israel where large-scale tests are dominant, and 70,000 12th grade students are annually tested by teacher-testers to indicate their OLP. The assessment formats have been revised a number of times since the breakthrough research by Shohamy, Reves & Bejerano (1985), due to dissatisfaction with reliability measures and practicality, which also instigated limited learning of speaking in the EFL classroom.

With the introduction of a national agenda to focus on OLP as an EFL essential skill in and the advent of formative assessment that links teaching and learning with assessment, a new computer-based version of an OLP test was introduced. Schools were invited to choose this test version which is innovative in two ways. First, it is designed on a national large-scale basis to explicitly promote washback in teaching and assessing speaking on a regular basis, thus espousing similar test items familiar to both teachers and students. Second, the test can be considered user-friendly constructed in an approachable manner using an avatar to guide the students through the test. In 2017 12,000 students took the OLP computerized test in a pilot mode.

Questions arise as to the effectiveness and potential washback of the test in terms of teachers' future use, i.e., if and to what extent will teachers embrace and implement the teaching of speaking in EFL classrooms and what will be the impact of this test as a model for classroom practice.

Hence, it is hoped that research will shed light on the following questions: To what extent will a new computer-based version of an OLP exam serve to drive classroom practice enhancing the teaching and learning of EFL speaking on a national basis? Specifically, can a large-scale test agenda motivate schools across sectors and cultures including diverse populations to raise teacher accountability for the teaching of speaking? Finally, will EFL high-school teachers learn to use the test applying formative assessment principles for their teaching and assessment of speaking?

The study aims to triangulate information from student and teacher surveys and from qualitative semi-structured interviews intended to tap views of various school stakeholders. Yet, some unresolved questions, mainly methodological, ensue from discussing precise, functional means to enhance students' oral proficiency: How can speaking progress in large-scale oral test setting be monitored? How can diverse populations of teachers and students be depicted in the surveys? Finally, what specific questions will stimulate responses to display the impact of a computerized-speaking test that are different from the OPI oral test? We are hoping to elicit answers in this session to promote our research.

20. Creating a Socially Responsible Language Diagnostic Tool to Support At-Risk Students at a Canadian Technical College

Nathan J Devos, British Columbia Institute of Technology

1:30 p.m. to 3:00 p.m. Location: Mary Gay C

In 2017, international students contributed $12 billion US to the Canadian economy, making it the fourth largest export sector in the country (ICEF Monitor, 2018). The competition between postsecondary institutions worldwide for these international student dollars has sometimes preceded pedagogical models to accommodate for the growing linguistic diversity that has accompanied it. Since studies have suggested that some students are inadequately prepared to take on the challenge of postsecondary education through the medium of English, the question arises: What moral obligation do institutes have to students once they've been accepted into a program? And how can post-admissions language assessment help or hinder students from having a positive learning experience in postsecondary schools in Canada?

This session introduces a research project that began as an online post-admissions language proficiency test of first-term students—both EAL and non-EAL—at the British Columbia Institute of Technology (BCIT). The test measured students' reading comprehension, using a tailored cloze test; vocational writing skills, based on functional adequacy; a vocabulary size test, and a timed Grammaticality Judgement Test. We also asked respondents about their language and educational backgrounds.

About 440 students from four different technical programs (i.e., Financial Management, Architectural and Building Technology, Electrical and Computer Engineering Technology, and Computer Systems Technology) participated in the study. Survey results confirm the linguistic diversity at Canadian postsecondary institutes. For instance, respondents were most confident in 22 different languages (including English), and 56% of the respondents said they could read and write in more than one language. In addition, 40% of the respondents said neither of their parents or guardians had spoken English with them since they were infants. These findings raise questions about how institutes should assess and support students if they happen to struggle with English as the medium of instruction.

Currently, the institute does not have a centralize model for English support after program admission. Therefore, this project is considered work in progress because while developing the test and analyzing the data, the researchers recognized that a language diagnostic and support model that identifies at-risk students' language issues, followed by specific language support, might better suit the institute's current needs. Furthermore, post-admissions testing that does not discriminate between linguistic or educational background coupled with individualized online and face-to-face support may begin to provide an answer to the question of moral obligation.

This session hopes to discuss and get advice on socially responsible methods to diagnose language issues in a linguistically diverse community as well as provide faculty and peer support that is particular to the requirements of different technologies at the college level.

21. Examining values and potential consequences in argument-based approaches to TOPIK-Speaking validation process

Soohyeon Park, Chung-Ang University, **Gwan-Hyeok Im**, Queen's University, **Dongil Shin**, Chung-Ang University

1:30 p.m. to 3:00 p.m. Location: Mary Gay C

This study attempts to integrate Messick's (1989) value implications and (potential) consequences into the early stage of test validation process. Value implications are still reluctant to be accepted in the stages of test validation (Fulcher, 2015), and it is often argued that they can be considered in the late stage, for example, in policy implementation (Mehren, 1997; Popham, 1997). However, McNamara (2005) argues that all language testing is political value-laden, and that language tests are products of a given policy used to achieve certain political ends (Fulcher, 2009; Shohamy, 2001). Messick (1989) also pointed out that value implications cannot be separated from validity discussions.

In this study, we attempt to include value implications and (potential) consequences into a phase of test validation process, by using a widely used argument-based approach to validation (e.g., Bachman & Palmer, 2010; Chapelle et al., 2008; Kane, 2013). It will be discussed that the argument-based approach lacks the (a) inference to discuss value implications and (potential) consequences ad hoc; (b) inference to validate multiple purposes of a test, implicitly or explicitly, set by a testing policy (Koch & DeLuca, 2012); and (c) consideration of a substantive analysis of the target language use in setting standards or scoring criteria (Fulcher, 2015). To address part of the issues, our revised validation framework consists of initial planning to examine implicit or explicit test purposes, to listen to various stakeholders' voices (e.g., values), and to conduct policy document analysis on politicians' values and possible consequences.

This study will present how to examine the discursive strategies employed by different stakeholders in the early stage of TOPIK- (Test of Proficiency in Korean) Speaking development and validation. TOPIK is the Korean government-led profiency test of Korean, offered six times annually to foreigners in Korea and twice annually to test-takers outside Korea. It only assesses listening and reading skills and now plans to include a speaking component. It is expected that discursive conflects over whether or not TOPIK-Speaking would be legitimized will flourish. Different values and (potential) consequences need to be examined in the initial stage of test development, and in the argument-based approach to the TOPIK-Speaking validation.

With all the other stages of validation, such as domain analysis to set standards and scoring criteria, and reliability check, the validation framework which put much emphasis on dimensions of value implication and consequences can address the aforementioned limitations of an argument-based approach in the context where high-stakes tests can be misused to achieve political purposes.

22. Aviation English Proficiency Test Design for a Group of Brazilian Military Pilots: A Case Study

Ana Lígia Barbosa de Carvalho e Silva, University of Campinas - Universidade Estadual de Campinas

1:30 p.m. to 3:00 p.m. Location: Mary Gay C

In Brazil, two tests have been designed according to the International Civil Aviation Organization (ICAO) prescriptions to assess proficiency in Aviation English: i) for civil pilots – Santos Dumont English Assessment (SDEA), developed by the Brazilian National Civil Aviation Agency (ANAC); ii) for Air Traffic Controllers (ATCOs) – Exame de Proficiência em Inglês Aeronáutico – SISCEAB (EPLIS), developed by the Brazilian Air Force (FAB). Nevertheless, up to the present time, an Aviation English proficiency test has

not yet been specifically designed for FAB pilots, more precisely for those belonging to the Air Demonstration Squadron (EDA), a high-performance team of military pilots whose mission is to represent FAB as a diplomatic instrument. This work-in-progress presentation introduces an ongoing Ph.D. research project that aims to design a proficiency test for EDA's pilots, whose needs are not restricted to communications between pilot/ATCOs. The study follows-up my M.A. dissertation, which outlined a language needs analysis (HUTCHINSON; WATERS, 1987; DUDLEY-EVANS ST. JOHN, 1998) which showed that the use of English by EDA's pilots encompasses several components: i) phraseology, a very specific register, or a coded language, used in radiotelephony communications between pilots and air traffic controllers; ii) plain English – "common language" used through radiotelephony, when phraseology is not sufficient. Note that, for safety reasons, both phraseology and plain English must be clear, direct and concise; iii) specific vocabulary for use in aviation contexts, even on the ground; iv) General English, necessary in social events, when acting as diplomatic representatives. Within FAB, the ability to use phraseology in English has been assessed by means of the so-called "International Air Traffic Test (TAI)", not considered a language test, properly speaking, and General English has been assessed separately. Meanwhile, Plain English in radiotelephony, as defined above, has not been directly assessed by any external proficiency exam, in the same scenario. To address such gap, this study outlines a language test for the group in a case study, through a qualitative research, which will include the analysis of data that stem from interviews, documents, observations and a workshop about SDEA, followed by a questionnaire. Data collection is to be held by Fall 2019. The purpose and the construct of the test will then be defined, as well as the profile of the examinees, the target language situation, the competences to be measured and the way to test them. This project focuses on the first three stages of the "exam preparation cycle" proposed by Fulcher (2010, p. 94): the definition of the purpose, criterion and construct of the exam. Thus, it goes beyond the scope of this study the planning of specific phases within the operationalization or application of the test itself, such as the preparation and pre-testing of tasks, or the design of a rating scale. This study seeks to contribute to the field's understanding of language test design for specific purposes. This study is being financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

1. Do students' motivation and locus of control impact writing performance through their perceived writing competency?

Clarissa Lau, University of Toronto, **Chris Barron**, University of Toronto

1:30 p.m. to 3:00 p.m. Location: Lobby

Motivation variables such as self-efficacy, goal orientation, and locus of control have been widely recognized as crucial predictors of overall writing performance in young students (Meier et al., 1984; Pajares, 2003). Specifically, goal orientation (GO) and locus of control (LOC) are found to influence a student's academic performance (e.g., Dweck & Leggett, 1988; Findley & Cooper, 1983; Pintrich, 2000, Ross & Broh, 2000). Yet, minimal research has focused on analyzing the relationship between students' motivational characteristics and specific writing features. Drawing from earlier research that discovered writing beliefs directly and indirectly impact writing performance (e.g., Pajares et al., 1999; Bruning & Horn, 2000), this study includes perceived writing competency to explore the structural relationships among GO, LOC, and writing performance.

We recruited 178 Grade 4-6 students to participate in a battery of questionnaires targeting literacy and motivational traits. Each student wrote an expository text regarding the effects of social media, and rated their competency in specific writing skills such as syntax, vocabulary use, and organization. The writing texts were analyzed using TextEvaluator (Sheehan, Kostin, Futagi, & Flor, 2010) to derive writing text feature scores. A total of ten writing features were extracted, which includes syntactical complexity, academic vocabulary, word unfamiliarity, concreteness, lexical cohesion, interactive conversational style, level of argumentation, degree of narrativity, and total text complexity score. Another feature, lexical variation, or type-token ratio was included to elucidate the number of unique words expressed in a given text.

Structural Equation Modeling (SEM) integrates factor analysis (i.e., measurement models) with regression and path analyses (i.e., structural models) to measure the relationships between both observed (manifest) and unobserved (latent) variables while accounting for measurement error (Hox & Bechger, 1998; Narayanan, 2012). Confirmatory factor analysis was utilized to ensure the factorial validity of the latent GO and LOC variables. Next, we applied a full SEM to examine the structural relationships among GO, LOC, students' perception of writing ability, and students' writing feature scores. We hypothesized that students' GO and LOC would influence writing performance through students' perceived writing competency.

The proposed SEM model adequately fitted the data. Key findings identified significant predictive relationship between mastery GO with structural features of text such as syntactical complexity, concreteness, and lexical variation. Results also indicated that students expressing external LOC negatively predicted inferential features of text, specifically degree of narrativity. This illustrated that individuals who perceive their control as externally-driven tended to write more narrative texts, which would be inappropriate as students were asked to write an expository text. Findings supported a direct relationship for students' perceived writing competency and writing.

Implications of this study provide evidence to support previous research, re-iterating a predictive relationship of motivation traits with writing outcomes. They also illustrate the need to investigate the effects of motivational traits at the writing skill level rather than overall writing outcomes. Practically, students would benefit from receiving writing support informed by their motivational profiles, and teachers would be more vigilant to recognize that individuals' writing complexity is influenced by their motivational profiles.

2. Deconstructing writing and a writing scale: How a decision tree guides raters through a holistic, profile-based rating scale

Hyunji Park, University of Illinois at Urbana-Champaign, **Xun Yan**, University of Illinois at Urbana-Champaign

1:30 p.m. to 3:00 p.m. Location: Lobby

Direct, performance-based writing assessment relies heavily on raters and their use of a rating scale. While rater training is necessary to ensure rater reliability, the choice of a rating scale also affects how raters rate. While a holistic scale requires a single score, raters need to weigh multiple rating criteria. In contrast, an analytic scale prompts raters to focus on one criterion at a time; however, raters can only focus on a limited number of criteria. In addition to these two traditional scale types, empirically derived, binary-choice, boundary-definition scales and decision trees are more recent innovations (Upshur & Turner, 1995; Fulcher, Davidson & Kemp, 2011) that represent the middle-ground combining the advantages of both holistic and analytic scales. Despite the theoretical soundness and methodological rigor behind these scales, little is known about how they affect rater behavior and performance. As such, this study employed a sequential mixed-methods design to investigate how a decision-tree scoring guide complemented an existing holistic, profile-based rating scale and affected rater performance on a college-level English placement test.

Nine certified raters participated in this two-phase study. In the first phase, the raters rated 152 essays using the holistic rating scale. Rater performance and use of the scale were examined both quantitatively through Rasch modeling and qualitatively via think-aloud protocols and semi-structured interviews. Based on rater behavior analysis, the decision-tree scoring guide was developed. The scoring guide consisted of two sequences of four binary questions, decomposing two major criteria in the holistic scale: argument development and lexico-grammar. Each sequence of questions add up to an analytic component score, which aimed to assist raters to assign a final holistic score. In the second phase, the raters rated 200 benchmark essays only with the scoring guide and 100 operational test essays using both the scoring guide and the holistic scale. Rater performance and use of the scale were analyzed in the same way as the first phase, with the addition of a post-rating questionnaire inquiring their perception of the scoring guide.

The many-facet Rasch measurement modeling of the ratings revealed that the scoring guide led to a much higher exact agreement than the holistic scale. The majority of raters' performance improved in terms of the exact agreement, severity measure range and internal consistency, while others did not exhibit much change. The results of qualitative analysis suggest that the rating guide had more positive impact on the raters who struggle with the holistic scale by decomposing it into more manageable sub-components. However, for those with higher reliability on the holistic scale, the scoring guide did not lead to much improved rating performance. Considering that the addition of the decision-tree scoring guide did not negatively affect any raters, using it to complement the holistic scale is justified. More importantly, the scoring guide helped many raters to deconstruct and better operationalize writing ability. This study represents a creative use of decision trees to guide raters through a holistic rating scale. Implications regarding how it affects rater cognition will also be discussed.

3. Assessing EFL college students' speaking performance through Google Hangouts

Yu-Ting Kao, National Kaohsiung Normal University

1:30 p.m. to 3:00 p.m. Location: Lobby

Oral communication in a foreign language often does not come easily, especially to beginning- and intermediate-level students because they need to speak by focusing attention on many things including developing an idea, mapping that idea onto appropriate structures, keeping conversational turns ongoing, and worrying about their interlocutors' response (Kern, 1995). Studies reported that communication activities implemented in synchronous computer-mediated communication (SCMC) settings have beneficial effects on learners' oral proficiency development. SCMC tools provide opportunities for L2 learners to interact with other language learners or teachers outside the classroom, and facilitate collaborative and comprehensible interaction by offering learner-centered interaction occasions. Additionally, L2 learning processes can be recorded through data collection tools and methods such as screen-capture software—recording chat scripts, interlocutor images, or learner on-screen behaviors. These features are attractive to both L2 learners and instructors, given that it is not always possible to provide students with plentiful opportunities for productive performance due to time constraints, and the high level of anxiety usually felt by students in face-to-face conversation. Thus, the SCMC environment has been regarded as an ideal alternative to develop L2 learners' speaking ability.

However, very few researchers have examined the effects of using SCMC tools to assess language learners' oral proficiency. This study aims to investigate the effects of using online interactive communication system to provide interactive supports to promote EFL college students' speaking performance. In order to assist students' speaking development, the teacher/researcher provided interactive supports based on the premise of Dynamic Assessment (DA). Grounded in Vygotsky's theory of mind, specifically, mediation, DA is defined as a procedure for simultaneously assessing and promoting development that takes account of the individual's zone of proximal development. Through the collaborative construction of knowledge, "the process of production changes" and learners are able to achieve a performance that they are unable to accomplish by themselves.

Twenty freshmen college students participated in the study and underwent six speaking practice throughout an eighteen-week of semester. Through individually joining the online chat room with the teacher/researcher via Google Handouts, students responded to various integrated speaking tasks while receiving the interactive supports from the researcher. The tasks required students to generate oral responses by integrating textual and/or aural information provided ahead of time. Comparing students' speaking performance from the pretest and posttest, the results indicated that their content was enriched and they were able to applied more speaking strategies in their responses, such as paraphrasing and showing varieties in word choice. It demonstrated the effects of using interactive supports in diagnosing and supporting students' development in a computer-mediated context. Additionally, compared with face-to-face speaking assessment, students also reported from the open-ended questionnaires that they experienced less anxiety when engaging in the online interactive setting. They were more willing to clarify tasks, ask questions, and express needs and concerns to the researcher. Future implications on the use of computer-mediated tools to assess speaking performance were discussed.

4. Is it fair to use scores from a test of grammar and vocabulary to refine grade boundary decisions in other skill areas?

Karen Dunn, British Council, **Gareth McCray**, Keele University

1:30 p.m. to 3:00 p.m. Location: Lobby

All tests contain a certain amount of measurement error, a circumstance that is crucial to the candidates whose performances fall just shy of a grade boundary. The decision to award one grade over another may have immediate and significant impact on the life of the candidate. To increase the fairness and accuracy of grade allocation in borderline cases, one international computerised test of English language proficiency uses score information from the grammar and vocabulary sub-section (the 'core' component) to refine CEFR decisions for listening, reading, speaking and writing components. This information determines whether a given candidate should remain at the lower level or be upgraded, a procedure justified by the understanding of grammar and vocabulary as key sub-processes in models of L2 language ability (e.g. Khalifa and Weir, 2009; Field, 2013) and empirical models showing them to be good predictors of language proficiency and to correlate highly with other skills (e.g. Shiotsu, 2010; Joyce, 2011).

The study reported here is an empirical investigation into the role of this grammar and vocabulary test and its relationship with the other four skill areas for a large testing population (n= 66,847, from over 250 test centres). The separability of grammar and vocabulary constructs are examined (RQ1) as well as the statistical basis for considering the grammar and vocabulary component as 'core' to the skills-based test (RQ2). Insights are also given into the numbers of candidates whose final grades are affected by this scoring framework (RQ3). A range of analytic approaches were employed in statistical software R to explore these issues, including MIRT (Multidimensional IRT) and GAMLSS (Generalised Additive Models for Location, Scale and Shape) analyses. As well as considering findings under these research questions, the paper also addresses some pros and cons of moving towards the use of big data in a language testing context; not least enabling use some of the more complex statistical methods employed here, albeit with attendant cautions.

Findings showed that the combined grammar and vocabulary items did not violate the unidimensionality assumption (RQ1), with a bifactor model accounting best for the factor structure across all components (RQ2). Most interestingly it was found that in this bifactor model of the full test, the loading of the core component on its specific factor was very small, compared to considerably higher specific values for the four skill components. This indicates that the core component carries virtually no additional information to the general factor, thus supporting the interpretation of its role as 'core' to the overall construct of L2 English. Additionally, results of the GAMLSS modelling exercise indicated that there is a strong association between scores on the core and skills components across the full score range. Having established this, it is noted that the process currently in place affects the CEFR grading of only the higher-level candidates (RQ3). Recommendations are made for a more detailed investigation into the value of applying such adjustments across the full range of CEFR levels, including drawing upon some more qualitative insights into comparative candidate performances.

5. Test taker characteristics as predictors of holistic score on independent and integrated-skills writing tasks

Analynn Bustamante, Georgia State University, **Scott Crossley**, Georgia State University

1:30 p.m. to 3:00 p.m. Location: Lobby

Because high-stakes standardized test scores are used to make real-world decisions regarding test-takers' academic careers, it is important to explore a test's fairness through investigating how the individual differences of test-takers may contribute to test performance. This study specifically focuses on the writing section of the TOEFL public use data set. While there are many studies that explore the relationship between text features and writing scores, more research is needed on the relationship between individual differences and writing scores. Therefore, the present study examines the predictive relationship between the demographic and English language experience characteristics of test-takers and their writing performance on a high-stakes standardized test.

This study addresses the follow research question:

Which selected examinee characteristics are predictive of holistic score on both the TOEFL independent and integrated writing tasks?

The corpus analyzed for this study consists of 480 essays produced by 240 test-takers wherein each test-taker responded to one integrated-skills and one independent writing task. Linear mixed effects models of test-taker characteristics were developed to predict holistic writing scores for both tasks. The models included fixed factors of age, gender, socio-economic status (SES), the number of years the test-taker has studied English, the number of years the test-taker had subject matter courses in English, the number of years the test-taker lived in an English-speaking country, and the linguistic distance (LD) between the test-taker's L1 and English. The data for these variables was collected through the TOEFL test-taker survey. Test-takers were asked about their annual monetary contribution to their university studies, which serves as a proxy for SES in this analysis. LD is measured on a scale developed by Chiswick and Miller (2005), where they defined LD as how difficult it was to learn a target language; test-takers' L1s were hand-coded by this scale. Country of origin was added as a random effect. The results indicated that for both tasks, SES and LD were significant predictors of writing scores. The integrated task model predicted 39% of the variance and independent task model predicted 35% of the variance. Age, gender and variables related to test-takers experiences learning English were not significant predictors. The findings suggest that the quantity of English exposure and traditional demographic variables may not be as predictive of writing proficiency as SES and LD. The present findings may help test developers better understand the relationship between an examinee's demographic and background characteristics and writing test performance, the implications of which can be directly applicable to the goal of social justice in language testing. The present study provides evidence that more research is needed on how and why SES and linguistic factors may contribute to test-takers' writing performance.

6. An Investigation of the Validity of a New Speaking Assessment for Adolescent EFL Learners

Becky Huang, University of Texas San Antonio, **Alison Bailey**, University of California, Los Angeles, **Shawn Chang**, National Taipei University of Technology, **Yangting Wang**, University of Texas San Antonio

1:30 p.m. to 3:00 p.m. Location: Lobby

This study presents the validity evidence for a newly developed speaking assessment for adolescent English as a foreign language (EFL) learners. The assessment was created in response to the growing emphasis on developing young EFL learners' oral communicative competence (McKay, 2006). It included two semi-structured interactive tasks that were administered on an individual basis. The rating rubric was adapted from the Communicative Language Competence Framework (CEFR), and was revised to better fit the content and topic of the new assessment.

The study utilized a cross-sectional, developmental design to investigate criterion, predictive, and construct validity. Participants included 257 EFL learners (Mean age = 14) who were enrolled in middle schools (grades 7-9) in Taiwan. All participants completed the new speaking assessment, an existing standardized English language test that has been previously validated (Author & Author, 2014), and a background survey. For the new assessment, participants completed two interactive role-play tasks, one of which focusing on the topic about birthday gifts while the other involving descriptions of a flower plant. The existing English test included two monologue subtests. One of them focused on a daily routine activity whereas the other a basic arithmetical task. For the survey, participants answered questions about their demographic background and history of English instruction from kindergarten to middle school, and self-evaluated their English language skills. All participants' final grades for the EFL content-area and teachers' ratings (n = 6) of participants' English language proficiency were also collected. The end-of-semester final grades were an average of their performances on quizzes and exams over the semester.

We double-scored 40% of the data to have a better estimate of the inter-rater reliability for the new test. The two-way mixed intra-class correlation (ICC) for the two raters was .87 for task 1 and .86 for task 2. Because scores on the two tasks were highly correlated ($r = .819$; $p < .001$), we created an average score and used the average score for the rest of the analyses.

We investigated criterion validity via correlating participants' scores on the new assessment with their scores on the existing test, their self-ratings of English proficiency, and their teachers' ratings of participants' English proficiency. For predictive validity, we examined the associations between their scores on the new assessment and their end-of-semester final grades. For construct validation, we conducted multiple regression analyses to determine whether demographic and instructional variables predict participants' test performances.

Results from the study demonstrated robust evidence for the new test's criterion and construct validity. The correlations between performances on the new assessment and ratings of participants' English proficiency were positive and strong, and English instruction variables also significantly predicted test performances. However, the association between participants' performances and final grades was relatively weak ($r = .259$; $p < .001$), suggesting potential discrepancies between the target domains of the new assessment and the English content-area school assessments. The new speaking assessment can contribute unique new information to teachers' knowledge of their EFL students. Implications for EFL assessments will be discussed in the presentation.

7. Instructors as Agents of Change: A Systematic Approach to Developing Proficiency-Oriented Assessments in Less Commonly Taught Languages

Shinhye Lee, University of Chicago Language Center, **Ahmet Dursun**, University of Chicago Language Center, **Nicholas Swinehart**, University of Chicago Language Center

1:30 p.m. to 3:00 p.m. Location: Lobby

Despite substantial growth in interest in Less Commonly Taught Languages (LCTLs) in recent years, what has been continuously identified is a severe lack of systematicity to sustain high-quality programs for those languages in higher education (Wang, 2009). More precisely, the biggest challenge that collegiate-level LCTL programs encounter is the lack of appropriate resources and reliable assessment tools enabling effective documentation and diagnosis of student learning (Nier, Donovan, & Malone, 2009). Due to the dearth of assessment-related training of LCTL instructors, LCTL assessments often take on an achievement-based testing format that renders limited insight as to students' progress and functional language use. Given that poor assessment practices may result in ineffective pedagogic practices and overall program evaluation (Alderson, 2005), systematic advocation for LCTL instructors is crucial to help them create assessment tools that feed into both enhanced curriculum in LCTLs and appropriate student evaluation.

In this study, we report on the Language Pedagogy Innovation Initiative (LPII), which was developed along with an existing Mellon Foundation Grant awarded to the University of Chicago Language Center (CLC) supporting collegiate-level LCTL programs. The highlight of LPII is its implementation of a systematic approach that empowers LCTL instructors to become the agents of change in developing proficiency-oriented assessment tools. During the first half of 2018, a total of 12 instructors across the Midwestern U.S. opted into the project; they represented eight different LCTLs (Bangla, Bosnian/Croatian/Serbian, Bulgarian, Catalan, Hebrew, Polish, Portuguese, and Tibetan). As part of the LPII, the instructors received funding to participate in two major phases of the project. First, they went through CLC-led language assessment workshops to build the assessment literacy necessary to design and develop proficiency assessments in LCTLs. Then they proceeded onto the Assessment Design and Development phase, in which they embarked on creating their own assessment tools. In this presentation, we particularly focus on describing how instructors followed a three-step assessment development procedure devised by CLC specialists, which included: (1) a test design stage in which the instructors defined target language functions and created test specifications for their tests; (2) a test development stage in which the instructors developed all test materials (listening and reading test contents) and administration protocols; and (3) an evaluation stage in which the instructors took an exit survey to share their thoughts on the overall project.

To explicitly demonstrate the outcomes of the LPII, we first showcase the progression of test materials (task types, the wording of prompts, and input materials) for eight assessment projects from their inception to completion. We then present a series of comparative analyses and examples of the instructors' previous and new versions of assessments, highlighting how the newly developed assessments are designed to better meet the specified target language functions. In light of the instructors' survey responses, we conclude the presentation by addressing the need to establish a systematic model for developing LCTL instructors' agency in making assessment-informed decisions and documenting students' proficiency development (Riestenberg, Di Silvio, Donovan, & Malone, 2010).

8. Multimodality, Social Semiotics, and Literacy: How LESLLA Learners from Refugee Backgrounds Make Meaning in Official U.S. Naturalization Test Study Materials

Jenna Ann Altherr Flores, University of Arizona

1:30 p.m. to 3:00 p.m. Location: Lobby

The purpose of this study is to gain insight into how adults from refugee backgrounds and non-Western cultures construct meaning from multimodal texts associated with assessment, specifically those related to the high-stakes U.S. naturalization test. The research focuses on refugee-background LESLLA (Literacy Education and Second Language Learning for Adults) learners.

Making meaning from multimodal texts requires understanding headings, directions, images, graphic devices, top/down and left/right organization, and the relationships between such elements. Because LESLLA learners are becoming literate (while simultaneously learning the language their literacy is manifesting in) due to limited/interrupted schooling, their visual literacy is also emergent. Many materials designed for beginning English language learners rely heavily on visual cues, which this population may miss. Knowledge of how diverse populations make meaning from multimodal texts is thus crucial for designing tests and study materials, which claim to support other kinds of learning.

There are no official U.S. naturalization test study materials that have been created specifically for LESLLA learners. Pilot research on official U.S. Citizenship and Immigration Service (USCIS) study materials has shown that materials display ideologies of (multi)literacy and assumed (universal) referential background knowledge, positing that such study texts are likely ineffective for refugee-background LESLLA learners. Furthermore, Kunnan (2009) also concludes that the naturalization test itself is nothing more than a redesigned test of English literacy skills.

Following these critiques, the current ethnographic study aims to provide insight into how LESLLA learners engage with official USCIS naturalization test study cards, and how U.S. naturalization assessment practices may be inadvertently biased against individuals with limited or extremely different literacy experiences. The research questions are: 1) How do non-Western LESLLA learners from refugee backgrounds make meaning from the multimodal test study aids, i.e. how do they understand the materials?, and 2) What kinds of implications or understandings do they perceive from the wider aspect of these materials?

Data come from semi-structured interviews with six refugee-background LESLLA learners who originate from African and Near Eastern countries. Interviews occurred in the participants' first languages with the assistance of refugee-background interpreters. The interviews elucidated how the participants understand and make meaning from the multimodal official USCIS flash cards. Data analysis utilized a critical multimodal social semiotic approach (Kress, 2010; Kress & van Leeuwen, 2006; Pennycook, 2001). Data was coded according to themes of multimodal (textual) composition, interpersonal relationship, and ideational content.

Results from the study are twofold. They show: 1) a disconnect between the multimodal content and composition of the study cards and how the participants understand them, and 2) a disconnect between the intent and use of the U.S. naturalization test as noted from the perspective of refugee-background participants.

From these results, a set of recommendations for multimodal text design and multimodal assessment practices will be developed, from which USCIS could create a set of U.S. naturalization test study materials specifically for refugee-background LESLLA learners. The results also yield implications of gatekeeping in regards to citizenship for this population; such results highlight sociopolitical issues in the high-stakes assessment of this vulnerable population.

9. Raters' decision-making processes in an integrated writing test: An eye-tracking study

Phuong Nguyen, Iowa State University

1:30 p.m. to 3:00 p.m. Location: Lobby

Integrated writing tasks have been adopted in both large-scaled, high-stakes assessments and local, placement assessments in many U.S. universities. Although researchers have suggested an expansion of the construct of integrated writing (Gebril & Plakans, 2013), more evidence is needed to examine the validity of this argument. A few studies investigating raters' decision-making processes in integrated writing (Cumming et al., 2001; Gebril & Plakans, 2014) have adopted the think-aloud method which has been criticized for its inability to give a complete picture of raters' cognition (Barkaoui, 2010; Lumley, 2005). Therefore, this study investigated raters' decision-making processes when rating integrated writing task responses using stimulated recalls and eye-tracking data. It aimed to address the following research questions:

1. What decision-making processes did raters underwent when rating integrated writing task responses?

2. How did raters attend to source use as revealed by eye-tracking data?

Nine raters rated 50 essays on paper before rating another 10 on the computer with a Gazepoint eye-tracker which recorded their eye movement trajectories. After rating each essay, raters verbalized their thoughts using the eye-tracking recordings as stimuli. Multifaceted Rasch analysis of the test scores (n = 540) indicated raters' consistency in their rating and comparability in their severity and the effect of rating medium (paper-based versus computer-based) on raters' decisions. To address RQ1, the verbal reports (n = 90) were transcribed, segmented, coded by two coders, and grouped into six broad strategy categories, namely Interpretation, Management, Organization-Argument, Grammar-Lexis, Convention, and Others before percentages of decision-making strategy use were calculated for each strategy and the broad strategy group. To answer RQ2, total fix duration (i.e., total time in seconds a rater spends fixating on a feature while rating) for the rubric categories, were calculated from the eye-tracking recordings. Also, raters' source text viewing patterns were analyzed to examine how raters attended to phrases or ideas borrowed from the source texts when reading the essays.

The results indicated that raters employed a combination of decision-making processes and attended to organization and argument the most frequently. Additionally, all raters considered source-based information accuracy, a criterion extraneous to the rating scale, in their rating. Eye-tracking data, combined with verbal reports, showed that raters referred to the source texts when they located verbatim text borrowing or source-based ideas in test-takers' essays, as thus, attended to source use in terms of citation mechanics, source use quality, and source use accuracy. The findings supported the argument that reading ability should be considered part of the construct of integrated writing and proposed a change in the rating rubric used to assess responses to integrated writing tasks.

This study was the first test validation study to employ eye-tracking data to investigate the test construct from raters' perspectives. The findings could provide insights into the raters' minds when rating responses to a source-based writing task, contributing to the knowledge about rater variability and the construct of source-based writing and implications for the use of eye-tracking as a method for cognitive data collection to investigate test constructs.

10. Story of an education system accountable for exam-success but not for learning: A washback study

Nasreen Sultana, Queen's University, Kingston

1:30 p.m. to 3:00 p.m. Location: Lobby

This study investigated the washback effects (influence of testing on teaching and learning) of an English public examination at the secondary level in Bangladesh on classroom teaching and learning in Bangladesh. Uniquely, the research examined washback from the alignment perspective of the examination with the national English curriculum and its associated textbook, and classroom instruction. The aim was to investigate how the alignment relationship among curriculum, classroom instruction, and examination produced washback effects on teaching and learning. This examination under discussion is known as the Secondary School Certificate (SSC) examination, which is the most important and biggest school leaving examination in Bangladesh- both socially and culturally (Author, 2018). Pellegrino, DiBello, and Goldman's (2016) argument and evidence-based validity framework was operationalized to guide this qualitative multimethod research design. Unlike earlier washback models (Bailey, 1996; Cheng, 2002, Green, 2007; Shih, 2007), which do not offer guidance to explore the alignment of tests with curriculum and instruction, Pellegrino et al.'s validity framework creates opportunities to explore test washback from an alignment viewpoint by incorporating evidence from various sources in the educational system. Of the three components of the framework, cognitive, instructional and inferential validity, I chose to employ Instructional validity, because it seeks evidence about the alignment of the tests with the knowledge and skills as defined in the curriculum and the usefulness of the testing for teachers and students to guide their classroom teaching and learning.

According to the adapted framework, data were collected from four sources: 1) document analysis of the SSC English National Curriculum (2012), textbooks, and a set of question papers of the English 2018 SSC examination; 2) Interviews with 12 English teachers at the SSC level at two schools and one curriculum, textbook and test developer; 3) English classrooms observation at the SSC level; 4) eight focus group discussions with the SSC test candidates from two schools.

Among many other interesting findings, the major results of the study were: a) the vision of the curriculum, textbook, and test developers was not translated into the classroom instruction because in developing the curriculum, textbook, and test, the teachers were never involved, and the socio-financial factors in implementing the test were not considered. Thus, the English curriculum is not democratic. B) The SSC examination is so powerful in the society of Bangladesh that teachers are socially trained to teach to the test superseding the curriculum objectives, which is the general norm in the classrooms. C) Students do not study English at the secondary level to develop English proficiency- they target to get the highest score in the examination to fulfill the expectations. Most of them are sure of their success in the examination but not sure about their ability in using the target language in real-life. This study resonates Shohamy's (1998) concerns about critical language testing, that, when tests are dominant, they override the curriculum by becoming the de facto curriculum and creates a negative washback on teaching and learning.

11. Using multimodal tasks to promote more equitable assessment of English learners in the content areas

Scott Grapin, New York University, **Lorena Llosa**, New York University

1:30 p.m. to 3:00 p.m. Location: Lobby

Across the US, students classified as English learners (ELs) are learning and being assessed upon academic content in English. Traditionally, assessments of content learning have been carried out through written language, with oral language and nonlinguistic modalities (e.g., visuals) being deprioritized or excluded altogether (Fernandes, Kahn, & Civil, 2017). However, new content standards in U.S. K-12 education expect students to demonstrate their learning using multiple modalities. Assessments focused exclusively on written language could miss ways that all students, especially ELs, communicate competently in the content areas using their full range of meaning-making resources. With new content standards emphasizing the multimodal nature of content learning, it is critical that assessments move toward embracing a broader spectrum of linguistic and nonlinguistic modalities.

The purpose of the study was to examine whether multimodal assessment tasks aligned to the new content standards provide more equitable opportunities for all students, and ELs in particular, to demonstrate their content learning. Specifically, the study addressed the following question: How does eliciting student responses to science tasks in visual, oral, and written modalities provide different information about students' science learning?

As part of a larger study, 30 fifth-grade students, including current ELs, former ELs, and non-ELs, responded to a science modeling task in visual and written modalities. These students were also asked to respond to the same prompt in the oral modality using an interview protocol similar to those described in Jewitt et al. (2001) and Kress et al. (2014). Interviews were transcribed using multimodal conventions for transcription to reflect students' use of gestures in addition to oral language (Bezemer & Mavers, 2011).

The analysis was carried out in two phases. In the first phase, we analyzed student responses in visual, written, and oral modalities to identify "critical events" (Powell, Francisco, & Maher, 2003), defined as those instances when students communicated different disciplinary meaning between or across modalities. In the second phase, we focused specifically on differences between oral and written modalities and the different ways that students of varying EL classifications communicated (or failed to communicate) precise disciplinary meaning in each modality. Preliminary findings indicate that different modalities expose different aspects of students' science understanding. This was the case for all students but even more so for current ELs, who used a combination of visuals (e.g., drawings and symbols), gestures, and oral language to communicate aspects of their science understanding that were not reflected in their written language responses.

This study challenges a deficit view of ELs associated with traditional written language assessments and suggests the potential of multimodal assessment tasks for providing more complete information about what all students, especially ELs, know and can do in content areas such as science.

12. Familiarizing standard-setting panelists with the CEFR: A three-step approach to attaining a shared understanding of just-qualified candidates

Sharon Pearce, Michigan Language Assessment, **Patrick McLain**, Michigan Language Assessment, **Tony Clark**, Cambridge Assessment English

1:30 p.m. to 3:00 p.m. Location: Lobby

In standard-setting, the concept of the just-qualified candidate is viewed as a crucial one, as it gives meaning to the panelists' judgements. They must have not only a firm grasp of the concept, but also a shared group understanding of how the just-qualified candidate is defined (Council of Europe, 2009). If such requirements are not met, the panel risks setting cut scores that are too high or not reaching a consensus on their recommendations at all (Lim et al., 2013; Tannenbaum & Cho, 2014). To ensure that test takers' abilities are evaluated as accurately and fairly as possible, and to avoid compromising the validity of the cut scores, standard-setting literature stresses the importance of training panelists to envision the just-qualified candidate. However, despite this importance, there is little documentation on how such training should be carried out (Mills et al., 1991; Kaftandjieva, 2004). Much of the guidance on training panelists for studies linking examinations to the CEFR centers on distinguishing the levels from one another, which, while necessary, only helps the panelists to understand the average candidate at each level, not the just-qualified one.

In this poster, we outline our three-step approach to training panelists to attain a shared understanding of the just-qualified candidate using a recent four-skill linking study at the A1, A2, and B1 levels of the CEFR as an example. This approach was designed to 1) introduce panelists to the CEFR levels relevant to the study, 2) train panelists in understanding and distinguishing between each level, and 3) focus panelists in on the key characteristics that defined the just-qualified candidate at each level.

The first activity took place pre-study and asked panelists to individually review several CEFR scales and describe what they perceived as key characteristics of both the average and just-qualified candidate at each level. The second activity, a descriptor sorting task, asked panelists on the day of the study to individually sort through and assign CEFR levels to decontextualized descriptors from scales relevant to the study, followed by a group discussion of the results to clarify any misclassified descriptors and ensure panelists understanding of each CEFR level (Council of Europe, 2009). Finally, in the third activity, panelists worked in small groups to share their pre-study definitions of the just-qualified candidates and agree on the key characteristics. Each group then contributed to a running list of key characteristics on a white board. Then, the newly-formed group definitions were discussed among the entire panel and modified until the panelists were satisfied that they had identified the key characteristics of the just-qualified candidate at each level.

In addition to these activities, this poster will describe how we selected the CEFR scales and descriptors that were relevant to the study, and how we approached creating our own definitions of the just-qualified candidates prior to the standard setting meeting. Overall, this poster aims to help address the shortage of information on how to train panelists in understanding just-qualified candidates by sharing one approach used with a four-skill linking study.

13. Using Machine Learning Techniques in Building and Evaluating Automated Scoring Models for ITAs' Speaking Performances

Ziwei Zhou, Iowa State University

1:30 p.m. to 3:00 p.m. Location: Lobby

The development of automated of scoring systems has largely resided in commercial sectors and only centralized around a few proprietary systems owned by large testing companies. Research on automatically scoring spontaneous speech has remained unfamiliar and mysterious to many language assessment professionals (Chapelle & Chung, 2010). As technology advances, computationally intensive methods, such as supervised machine learning, have become practical to implement. However, despite its popularity in other scientific fields, machine learning methods remain largely unexplored in educational measurement and language assessment (Sinharay, 2016).

Using a variety of existing speech and natural language processing tools, this study explores the extent to which machine score can be a reasonable prediction of human score from an in-house speaking proficiency test for international teaching assistants (ITAs) at a large Midwest university. Leveraging the machine learning techniques, this study focuses on the Evaluation and Explanation inferences of argument-based validation approach in addressing the tension between human and machine scoring. Specifically, a number of popular machine learning algorithms were attempted to map the attributes of the test takers' oral responses to their speaking proficiency levels, as assigned by trained raters. Model stability issues were addressed by constructing a bootstrapping confidence interval for the 10-fold cross-validation results, as well as extrinsic evaluations by testing models on new speaking prompts in the ITA speaking test so as to gauge the degree of internal consistency of machine scores. Model evaluation was conducted based on a variety of model performance metrics informed by both machine learning and psychometrics communities (e.g. Williamson, Xi, & Breyer, 2012). Further insights were gained by explicating the scoring logic adopted by the best-performing automated scoring model and analyzing how the model made use of the available features to make predictions.

Results indicated that the highest model performance was achieved by constructing 500 full-length decision trees by the random forest algorithm (exact agreement = .84 correlation = .76, quadratically-weighted Kappa = 0.71, and standardized mean score difference = 0.12) based on a subset of fluency, pronunciation, prosodic, and audio signal features. Cross-prompt evaluation showed good consistency of the machine scores generated by the best-performing model. Feature analysis revealed that, while the automated scoring model made use of all features in making the decisions, it assigned dominant weights to a number of influential features, including fluency, pronunciation, and audio signal features.

This study is representative of the increasing efforts to incorporate state-of-the-art natural language processing, speech processing, and machine learning technologies in support of language learning and assessment. It also calls for a focus on augmenting our understanding of how technological innovations could be exploited to make potential contributions to a coherent assessment argument (Williamson & Mislevy, 2006).

14. A systematic review: Ensuring high quality ELP assessments for all

Jo-Kate Collier, University of Texas at San Antonio

1:30 p.m. to 3:00 p.m. Location: Lobby

In the U.S., beginning with No Child Left Behind (NCLB) legislation in 2002 and continuing under the Every Student Succeeds Act (ESSA) of 2015, all states were federally mandated to monitor the progress of all identified English learners (ELs) in acquiring English by administering a statewide English language proficiency assessment (ELP) annually. To fulfill this obligation test developers joined together with states to form assessment consortia. Out of these collaborations a number of new generation ELP assessments resulted. Two consortia assessments remain in use today; the ACCESS 2.0 and the ELPA 21. Some states chose not to join consortia and developed their own assessments. These included states with the largest concentrations of EL populations such as California, Texas, New York, Arizona and Florida (Migration Policy Institute, 2018). While a few of these states eventually ceded their independence and joined consortia, some have remained independent. The larger consortia have resources available on a grand scale including federal funding grants to meet validity, reliability and research responsibilities. While the numbers of ELs in the states that remained independent are high, the research dedicated to these assessments does not equal the research invested in the consortia assessments. With the purpose of ensuring that all ELs have access to instruments that are held accountable to rigorous critical research investigations providing for validity and reliability, technical quality, and engaging stakeholders in ongoing regular feedback to address any unintended negative social consequences, this study aims to conduct a systematic review of the field to examine the pool of available research on all ELP instruments currently in use by states for the purpose of fulfilling federal obligations to monitor the progress of ELs in acquiring English. The research questions the study intends to answer include:

RQ1: Are the ELP instruments used by states for federal accountability represented equally in the research?

RQ2: What are the identified areas of need that will equalize representation in the research and increase the validity and quality of ELP instruments across all states?

The inclusion criteria to be used are:

1. May be theoretical or empirical studies, technical reports, and policy analyses

2. Must be published between 2008 and 2018

3. Must focus solely on ELP assessments currently used by states

4. Must be peer reviewed article, book chapter, dissertation, or state report

Exclusion criteria include research not directly relevant to the research questions, studies published outside the stated timeframe, and studies focused on ELP instruments no longer in use or not used for federal accountability purposes.

Preliminary results suggest that independent assessments have less presence in the available body of research and that independent states have been less rigorous with regards to validity studies. This literature review contributes by highlighting gaps in the research and generating knowledge to inform a future research agenda with the goal of guaranteeing high quality assessment instruments for all ELs and holding all states accountable for making certain the assessments relied upon for accountability measures related to ELs are valid, just and fair.

15. Bridge to Seven (Language Testing and Social Justice)

Johanna Motteram, British Council

1:30 p.m. to 3:00 p.m. Location: Lobby

Transparency in high stakes assessments is a fundamental demand of Critical Language Testing (Shohamy 2001). To address this, publication of sample papers, scoring criteria, and research reports has become the norm for large scale tests. In the case of IELTS, research reports dating back nearly 20 years are in the public domain. Careful attention to these documents should lead to a clear understanding of the construct of the test. However, in contexts where the discourse of language teaching, learning and testing is oriented towards structure and vocabulary, these explanations can be misinterpreted, misunderstood and eventually miscommunicated.

This poster reports on “Bridge to Seven”, a British Council and ClarityEnglish collaboration which aims to demystify the demands of IELTS Academic writing tasks one and two for a particular candidate population. The target population comprises trained nurses in the Philippines who require an overall IELTS score of 7, with no individual band score below 7, for registration as nurses in the United Kingdom. These are repeat candidates who have achieved 7 in the speaking, writing and reading subtests but have not been successful in the writing subtest. The nurses display high motivation to achieve their scores but face significant barriers to access focussed test preparation support. Many of the nurses have heavy caring responsibilities and work and commute long hours.

In order to support flexible access and engagement, “Bridge to Seven” has been developed as an online test preparation product. It can be accessed on mobile phones, tablets or laptops, and the full version includes a feedback stage where students submit their writing for specialised feedback from writing instructors.

This poster reports on the project phases of understanding the specific needs of the candidate population through surveys, text collection and text analysis, and developing and piloting the course to meet those specific needs. In response to the needs analysis, and in contrast to the prevailing discourse of language learning in the Philippines, the course focuses on developing participants’ resources for writing cohesive and coherent texts, and in identifying and meeting the requirements of the task.

Bridge to Seven is also an example of a project where a language test owner (British Council) has maintained a commitment to language test candidates beyond the provision and administration of the test. In the Philippines the British Council works to educate test preparation trainers and test candidates about the test, and to support candidates as they pursue the scores required to achieve their goals.

16. Beyond the Test Score: Developing Listening Test Feedback & Activities to Empower Young Learners and Teachers of English

Brent Miller, Michigan Language Assessment, **Luke Slisz**, Michigan Language Assessment, **Patrick McLain**, Michigan Language Assessment, **Rachele Stucker**, Michigan Language Assessment, **Renee Saulter**, Michigan Language Assessment

1:30 p.m. to 3:00 p.m. Location: Lobby

Providing test takers with scores alone is insufficient if a test is to have an impact on learning (Alderson, 2005; Kunnan & Jang, 2009; Lee, 2015). Such a test must consist of an appropriate construct for the stage of cognitive development of the target test taking population, and tasks that cover a range of levels and which lend themselves to the provision of diagnostic feedback (Buck, 2001; Gu, 2015). In this poster, we report on the development of a multi-level listening test designed to track young learners'

development of English from the A1 to B1 levels on the Common European Framework of Reference (Council of Europe, 2001 & 2018).

In particular, we feature the listening task types developed for the test and report on their effectiveness in assessing the ability of young learners, aged approximately 11 to 15. The items were developed to assess a variety of skills appropriate for our test taking population, with different task types designed to be accessible to both lower- and higher-level test takers. Piloting showed that the task types assessed a wide range of ability levels, and that there was a progression in task difficulty based on the CEFR levels targeted by the task types. This provided a basis for giving diagnostic feedback to test takers.

We discuss the personalized feedback descriptors that were created to provide test takers with an understanding of their strengths and weaknesses, and to give them more ownership of their performance and learning. The sub-skills tagged on each item provide the basis for this feedback: for example, if a test taker performed well on items testing main idea but not on items testing details, they received the feedback "When listening, you get the main points, but try to listen for the details, too." Thirty-three percent of test takers received unique feedback blocks, where "block" refers to a combination of feedback descriptors. Test takers were able to understand the feedback they received because the language used in the descriptors was written at the level of the English ability they demonstrated on the test.

We describe the learning activities that were paired with the personalized feedback to help test takers pursue learning on their own or with their teacher's help. Qualitative research suggests that test takers prefer learning activities they can engage with in real life, rather than vague recommendations to practice language skills (Sawaki & Koizumi, 2017). The learning activities we developed help test takers focus on overcoming their challenges by engaging in interesting, authentic activities outside of the classroom.

Additionally, we discuss the development of the assignment procedures that we established to evaluate each individual test taker's performance and select the most appropriate feedback descriptors and recommended learning activities based on their performance profile.

Considering the need to provide test takers with more than just a test score, this poster aims to share the results of a multi-level listening test development project focused on empowering test takers and their teachers with informative, personalized feedback and real-life learning activities.

17. Cyberpragmatics: Assessing Interlanguage Pragmatics through Interactive Email Communication

Iftikhar Haider, University of Illinois at Urbana-Champaign

1:30 p.m. to 3:00 p.m. Location: Lobby

Assessment of second language (L2) pragmatic knowledge is still a new and understudied area of research. Some researchers (Hudson, Detmer, & Brown, 1992, 1995; Roever, 2006, 2011) have played an important role in advancing the field (Walters, 2007; Grabowski, 2009; Roever, 2011; Youn, 2015). Methodologically, past studies mainly used closed role-plays based on predetermined interactional outcomes (Youn, 2015). Kasper and Rose (2002) doubted the validity and authenticity of closed role-play tasks. In order to address these research gaps, and as a contribution to the general understanding of Second Language Pragmatic Testing (SLPT), this study combines second language pragmatics and computer-mediated communication to assess the pragmatic knowledge of second language users of English. A thorough needs analysis was conducted first through semi-structured interviews and then through an online survey by involving different ESL stakeholders at a large university in the Midwestern

U.S. The results of the needs analysis helped to determine appropriate role-play situations. A set of communicative role play tasks was developed following Davidson and Lynch's (2002) test specification theory. A group of 52 graduate ESL students completed email role-play tasks. Role-play cards were used to enhance standardization, and test takers were allowed to communicate naturally without following fixed interactional outcomes. Two native-speaking data coders evaluated the pragmatic ability of test takers and assigned scores using Brown and Levinson's (1987) politeness framework-based scoring rubric for the email threads. Qualitative analysis of interactive data revealed a lack of knowledge of politeness norms by the lower proficiency groups. Therefore, the low stakes test might have a great potential for developing assessment materials in an academic email communication context.

18. Reverse-engineering L2 reading and listening assessments for sub-score-reporting purposes

Yeonsuk Cho, Educational Testing Service, **Chris Hamill**, Educational Testing Service

1:30 p.m. to 3:00 p.m. Location: Lobby

A score report is an important dimension of a test's design because it is serves as the medium of communication between test developers and score users regarding test takers' performances. Especially as interpretations and uses of test scores have increased in significance within the current approach to validation of assessments, it is hardly defensible to evaluate the effectiveness of a score report without considering the needs and practices of score users as they use it. To that end, we conducted research in 2017 into the needs and practices of teachers and administrators in eight countries who use the score report of a English-language test for young students (primarily ages 11-16). Their feedback indicated that while they were overall satisfied with the existing score report, they also expressed a desire for the score report to provide more fine-grained information than is currently provided regarding students' specific English skills or sub-scores within each modality. Building on this finding and taking into consideration that one of the intended uses of test scores is to help understand English language learners' progress over time, we explored whether it would be feasible to summarize test takers' performances on the Listening and Reading sections in a detailed, diagnostic manner without modifying the actual assessment's design.

We employed a process of "reverse-engineering" the test's content in order to devise categories for reporting sub-scores as the test was not originally designed to provide such detailed information. In order to report sub-scores, the sub-score categories had to meet two constraints: (1) each category must have a sufficient number of items within each test form to ensure reliability; (2) the proportions of items in each category must be comparable across all test forms. The project was conducted in two phases. During the first phase, a panel of experts reviewed the Listening and Reading sections of five test forms, categorizing each item into one of two groups within each skill, according to levels of comprehension commonly reported in the literature (e.g., Alptekin, 2006; Johnson & Jacobson, 1968; Kintsch & Kintsch, 2005; Pearson, 1978): explicit and literal comprehension, and implicit and inferential comprehension. Analyses showed that the majority of test items could be reliably classified into these two categories and that the proportions of the subcategories were sufficiently consistent across the test forms. During the second phase, we conducted focus groups with EFL teachers to validate the groupings of the items from the first phase and refine the definitions of the sub-score categories using pedagogical language that is accessible to teachers in EFL contexts.

In this presentation, we will demonstrate how the levels of comprehension were defined in the study context, and share the challenges we faced in operationalizing the levels of comprehension when analyzing the test items for L2 learners. We will also present the results of the findings from both phases of our study.

19. Scenario-based tasks for a large-scale foreign language assessment: a mixed-methods exploratory study

Malgorzata Barras, University of Fribourg, **Katharina Karges**, University of Fribourg, **Peter Lenz**, University of Fribourg

1:30 p.m. to 3:00 p.m. Location: Lobby

In this contribution, we will present the development and pre-operational testing of scenario-based assessment (SBA) tasks testing 9th graders' receptive skills in French and English as foreign languages. SBA (Sabatini et al., 2014) was originally developed to implement a "broader construct for reading" which targets academic reading in the language of schooling, including online reading. As SBA focuses on purposeful reading tasks and the relevance of reading for content learning, it fits well with the action and content-oriented approaches to foreign language teaching and learning currently dominant in our country's teacher education.

Our computer-based tasks, assessing reading, listening or a combination of the two, intended to emulate relevant language use within larger scenarios, e.g. preparing a trip to another city. To complete this scenario, the students solved several tasks, such as getting advice on sights to see from peers or deciding on a hotel based on online reviews and specified criteria. To account for current reading and listening habits, we used digital text types such as smartphone chats, websites, audio messages and blogs. All scenarios were developed to be suitable for large-scale assessments, with selected-choice item formats only, an easy-to-use screen design and simple instructions in the language of schooling. The tasks were located around the CEFR levels A2/B1.

We investigated the functioning of the tasks in a mixed-methods design. Qualitative data was collected through think-aloud protocols and/or retrospective interviews of more than 80 students, 30 of which participated in 2-hour in-depth interviews. Quantitative data was gathered in a subsequent pilot study involving nearly 600 students. It includes student responses from the scenario-based tasks, from a series of enabling skills tests, and from questionnaires.

In our poster presentation, we will first introduce some task types and then focus on qualitative evidence about the students' perceptions of some central features of our scenario-based tasks: perceived relevance of the tasks; inclusion of recent, electronic texts; embedding of the tasks in scenarios. Students find tasks relevant, and therefore motivating, when they consider them a part of their (potentially future) every-day life. For the same reason, they generally liked the digital text types we employed. The scenario embedding of the various tasks usually went unnoticed and did not seem to create any particular engagement with the tasks.

20. Developing a local-made English test for Thai EFL grade 6 students: Concurrent validity and fairness issues

Jirada Wudthayagorn, Chulalongkorn University Language Institute, **Chatraporn Piamsai**, Chulalongkorn University Language Institute, **Pan-gnam Chairaksak**, Chulalongkorn University Demonstration Elementary School

1:30 p.m. to 3:00 p.m. Location: Lobby

In Thailand, the Common European Framework of Reference (CEFR) has been introduced into the Thai Basic Education policy since 2014. The policy suggests that grade 6 students reach A1 proficiency, grade 9 students A2, and grade 12 students B1. English has a foreign language status in Thailand, and it is a core subject in basic education. On the whole, Thai students' exposure to English varies greatly. The

majority study English in Thai program with Thai-nationality teachers. Some study with native English-speaking teachers in such programs as English program, British Council program, bilingual program, and international program. Others may also have opportunities to attend cram schools or study abroad during summer break. Thus, it seems that educational opportunities and privileges within the Thai society lay the foundation for students' English proficiency. One important question thus arises: What would be an appropriate test to assess students of such diverse language education background?

This poster presentation demonstrates how a local-made test for Thai EFL grade 6 students is developed. The 50-minute paper-based test includes 20 listening items, 20 reading items, and 10 writing items. Additionally, in a separate sitting, students perform one speaking task via face-to-face interview with trained teachers. The current study was conducted at one primary school in Bangkok where three types of programs (i.e., Thai, British Council, and English programs) are offered. Consent forms were sent to parents asking for permission to recruit their children into this study. The test was first piloted with 10 students, after which some items were revised. Then, all grade 6 students (n = 210) took the revised test. The reliability of the test was relatively high (KR 20 = 0.82, KR 21 = 0.78). Later, 110 students were randomly selected to take the Cambridge English: Young Learners tests. The correlation coefficients between the two tests were computed. The listening parts from the two tests yielded the strongest relationship ($r = .80$, $p < .01$), while the speaking parts the weakest ($r = .26$, $p < .01$). The writing parts showed moderate, yet significant, relationship ($r = .46$, $p < .01$). The local-made test therefore needs further improvement because it is not well concurred with an established test of the same purpose. Results also showed that over 80 percent of test takers in this study reached A1 level, with students in British Council and English programs outperforming those in Thai program and demonstrating more potential to move on to the next CEFR levels more easily and quickly. Although common core items that will be fair to all students are under investigation, it is important to note that fairness may lie beyond the test itself. The opportunity for the students to have access to different types of programs may be more crucial to their development of English. Thus, the concept of fairness also needs to be redefined for the Thai basic education context.

21. Holistic and analytic scales of a paired oral test for Japanese learners of English

Rie Koizumi, Juntendo University, **Yo In'nami**, Chuo University, **Makoto Fukazawa**, University of the Ryukyus

1:30 p.m. to 3:00 p.m. Location: Lobby

The significance of enhancing and assessing interactional competence has recently garnered special attention in the language assessment community (Lim, 2018; Plough, 2018). Considering learners' future use of their target language, spoken interaction is increasingly highlighted in the English-as-a-foreign-language context in Japan, as suggested by its explicit inclusion in the Course of Study (Japanese national curriculum of English for secondary schools), which will be implemented from 2020. However, English teachers in Japan have limited knowledge of and experience in assessing spoken interaction, particularly in classroom assessments, and there is an immediate need for raising their language assessment literacy related to tasks and rating scales of assessing oral interaction. One viable format for classroom assessment in this context is a paired oral test, where two students talk or play assigned roles based on instruction cards. This method has been used to elicit relatively natural oral interaction between two people with similar status and is believed to generate positive washback on students' learning (Galaczi & Taylor, 2018). We have previously developed paired oral tasks and a holistic rating scale for Japanese learners of English and presented positive evidence for the validity of interpretations and uses of test scores (Koizumi, In'nami, & Fukazawa, 2016). However, previous research suggests that, although a

holistic scale produces fairly reliable scores and is more efficient than an analytic scale, it lacks the diagnostic information to help improve future learning and teaching that the analytic scale offers (e.g., Brown, 2012). Therefore, in order to provide language teachers with two scale types that function adequately for our test, this study reports on the development of an analytic scale, examines its quality using a multifaceted Rasch analysis, and compares it with our holistic scale.

Students at four Japanese universities (N = 110) with novice to intermediate English proficiency levels took a paired oral test. As part of the instruction in an English class, they paired up and completed three to 10 tasks that required each pair to talk for two to three minutes. Their interactions were recorded separately for each task and marked by two trained raters from a pool of three or four, using a holistic scale and a newly developed analytic scale developed based on Nakatsuhara (2007, 2013). The analytic scale consisted of four categories: Pronunciation & intonation, Grammar & vocabulary, Fluency, and Interactive communication. Each scale was awarded 1–3 points. The ratings were analyzed using a separate multifaceted Rasch measurement for each scale. The analysis showed that the two rating scales basically functioned as expected, with tasks and raters fitting the Rasch model well, except for the Pronunciation category. The holistic scale was found to correlate strongly with the analytic scale categories. Differences in the scales as found in bias analyses are also discussed in this presentation. Moreover, possible directions for future research and practices are presented, including plans to encourage secondary and tertiary English teachers to use paired oral tasks and rating scales, along with rating training sessions, as a part of enhancing their language assessment literacy.

22. Accessibility in testing: generating research from good practice

Richard David Spiby, British Council, **Judith Fairbairn**, British Council

1:30 p.m. to 3:00 p.m. Location: Lobby

As demand for language assessment continues to grow, many governments and institutions are passing new laws and regulations to ensure that tests can be accessed by citizens with a range of needs. This means that language test developers have a greater legal responsibility to ensure that their tests are accessible as well as a moral and professional responsibility to ensure that all test takers have the chance to perform to the best of their ability. These principles are widely recognised by testing associations in the publication of guidelines for the production and delivery of ethical tests (e.g. ILTA, 2007; AERA, 2014). However, there are also research-related reasons for improving the accessibility of tests, since the amount of research conducted in this area of L2 assessment is relatively small (Taylor, 2012), even though such research is an essential source of validity evidence for test modifications.

This poster addresses the issue of accessibility in tests in terms of operational testing and assessment research. First, it shows the process of formulating good policy and practice within the context of the British Council, an organisation with a public commitment to providing educational opportunity and to mainstreaming equality, diversity and inclusion. A framework integrating the many facets of this work is presented, including: consultation and specialist expertise; development of coherent ethical guidelines, communication with internal and external stakeholders; and investigation of case studies. Within this framework, accessibility issues are addressed at different stages of the test development cycle, from design, production and delivery through to scoring and validation. As such, the framework provides a basis for investigation into the effectiveness of accommodations together with their personal and social consequences for stakeholders.

Second, the poster presents relevant examples of such research. There are many challenges in conducting research into special needs testing, especially in terms of quantitative analysis and the

generalisability of results. However, a useful method of collecting empirical data has been to compile a log of requests for accommodations. The details of candidate requests are recorded along with responses given and any issues related to feedback, practicality or theoretical challenges associated with meeting the requests. In this way, a series of observations from the field have been gathered, built on iteration and intuition using grounded theory (Stake, 2005). As data provided by test takers and test administrators globally has accumulated, an ethnographical study of L2 testing for special educational needs is starting to emerge.

Representative case studies illustrate the process of identifying and assisting test takers with special needs as well as the appropriateness of the reasonable adjustments made. Theoretical implications and practical challenges for test producers will be addressed, while discussion of test impact and concepts of social justice will be encouraged.

23. Listening to test-takers' perspective in the validation process: the case of the Aviation English Proficiency Exam for Brazilian Air Traffic Controllers

Natalia de Andrade Raymundo, University of Campinas / Brazilian Air Force

1:30 p.m. to 3:00 p.m. Location: Lobby

This study aims at presenting findings on the ongoing validation process of the Aviation English Proficiency Exam for Brazilian Air Traffic Controllers, named EPLIS, by hearing the voices of the test-takers. EPLIS is a performance test developed by language teachers and aeronautical subject matter experts in Brazil, in compliance with the International Civil Aviation Organization (ICAO) requirements, the language policy that all member states should follow. EPLIS, which is a high stakes exam because of its relevance for society, assesses the performance of Brazilian ATCOs in communicating in non-routine situations within an international community of users of the English language.

We will make a parallel between what the ICAO rating scale considers as operational in Air Traffic Control regarding English proficiency, and the language skills Brazilian Air Traffic Controllers need to deal with language complications. A 45-item questionnaire was designed and answered by a great number of Brazilian Air Traffic Controllers who got proficiency Level 3 (considered non-operational) and Level 4 (considered by ICAO as the minimum level necessary to control aircraft using the English language). The questionnaire was divided into 3 parts: test-takers' perceptions of the rating scale established by ICAO; test-takers' perceptions of their proficiency while controlling aircraft in English; and problems they faced while controlling in English. First, descriptive statistical analysis of the demographic variables was conducted. Then, exploratory factor analysis and principal component analysis were employed to identify the major constructs present in the questionnaires. Finally, correlation analysis was used to plot the significance of the relationships between variables.

Findings suggest misalignment between the test-takers' perceptions of the use and the need for English language proficiency and the rating criteria recommended by ICAO emerged in this study. Implications for changes to the aviation language policy and test development are discussed.

24. Pre-service and in-service language teachers' conceptions of LA: towards the construction of LAL knowledge base

Sonia Patricia Hernandez-Ocampo, Universidad de los Andes/Pontificia Universidad Javeriana

1:30 p.m. to 3:00 p.m. Location: Lobby

The democratization of assessment (Shohamy, 1998, 2001) exemplifies well how language testing and assessment involve social justice. This is because it entails the active agency of teachers in the development of sound assessments in their classroom, which is completely dependent on context. In Colombia, for instance, language assessment has not been given all the attention it deserves. According to my review of the literature—in six well-known journals, during the last 10 years—when scholars report their research on language teacher education, the focus is the pre-service teacher practicum, the role of research in such programs, pre-service teachers beliefs (about issues other than assessment), or competencies of teachers. The educational community does not consider LA when establishing the competencies an English language teacher should have. Instead, teachers are expected to have good pedagogical and methodological competences, good in-class research skills, full command of the target language and culture, and vast knowledge of other topics (Cortés Cárdenas, Cárdenas Beltrán, & Nieto Cruz, 2013); assessment, however, is not even mentioned. In addition, I have found that still very few of the accredited language teaching programs have a course on evaluation or assessment in their curricula.

In an attempt to encourage such assessment-oriented empowerment of teachers, this study collects self-reports concerning evaluation in order to identify pre-service and in-service teachers' perceptions of LA—as a first stage (pilot study) to determine the knowledge base of LAL that should be included in English language teacher education programs in Colombia. To do so, I have collected information from content (course on evaluation) and language teachers, pre-service teachers and the courses syllabi in a language teacher education program at one private university. The instruments used are: interviews to teachers, think-aloud protocols to language teachers, focus groups and questionnaires to students and content analysis to syllabi. The analysis of data involves the use of software for qualitative analysis that allows categorization; both emergent and a priori categories have resulted.

The pilot findings at this stage suggest that decisions about LA knowledge in the evaluation courses are made on the basis of what the course designer (content teacher) considered was necessary for pre-service teachers to know about it. Some language teachers—despite reporting to know what evaluation, assessment and testing entail—do not make any difference between evaluation and testing, and consider assessment apart from those two. Pre-service teachers barely refer to assessment; instead, they focus on testing and how difficult it is to obtain good scores.

25. Language assessment literacy in Brazil: analyses of undergraduate and graduate courses at federal universities

Gladys Quevedo-Camargo, University of Brasilia, **Matilde V. R. Scaramucci**, University of Campinas

1:30 p.m. to 3:00 p.m. Location: Lobby

Taking into consideration a) the current national and international socio-political-economic scenario; b) the growing importance of internationalization in the Brazilian educational system; c) the undeniable presence of high-stakes exams in the Brazilian elementary and higher education; and d) the crucial role (language) assessments play both as an integrating element in the (language) teaching-learning process and as a propelling mechanism of changes in this process (Scaramucci, 2016), it is of paramount importance to improve the level of language assessment literacy of the language teaching professionals in Brazil, mainly of those who work at the public sector. This poster presents data which is part of a larger study which aims at Investigating whether and how language assessment has been approached by language teacher education courses at undergraduate and graduate levels in Brazil. The curricula and the programs of 50 Brazilian federal universities from all of the five regions in the country and the federal district were analyzed. The results show that, although (language) assessment is considered important for both pre-service and in-service language teacher education, this subject has not been given proper attention in the curricula and programs analyzed. The likely causes and consequences of this situation will be presented during the poster presentation.

26. Impact of Language Background on Response Similarity Analysis

James Robert Davis, University of North Carolina at Greensboro

1:30 p.m. to 3:00 p.m. Location: Lobby

Response similarity statistics, such as the Generalized Binomial Test (GBT, van der Linden & Sotaridona, 2006) and M4 (Maynes, 2014), have been developed to detect potential examinee collusion (e.g., obtaining prior access to test content from the same illicit source; discussing item content during test administration). While similar response patterns—especially those involving substantial incorrect identical answers—provide convincing evidence that collusion has occurred, validation of collusion inferences requires consideration of other plausible explanations for pattern similarity. Such explanations may include language-related examinee characteristics, such as native language or proficiency in the test language (Mueller, Zhang, & Ferrara, 2017). For example, complex or nuanced language contained in item response options may result in examinees with similar language backgrounds selecting the same incorrect options at a rate greater than chance.

The proposed study will investigate whether examinees with similar language backgrounds are more likely to produce response patterns that are flagged for collusion. Data for this study come from a certification testing context in the area of health care. First, Maynes (2014) M4 statistic will be used to quantify response similarity between all pairs of examinees. M4 is preferable because it accounts for both incorrect and correct identical responses, leverages a generalized trinomial distribution to increase sensitivity, and allows for many standard measurement models to derive response probabilities. Second, hierarchical agglomerative cluster analysis, based on M4 as a distance measure, will be used to identify groups of examinees who have potentially colluded. Third, descriptive and inferential statistics will be used to investigate relationship of language background similarity to clustering.

Such investigations are critical to ensure examinees are not unfairly suspected of and penalized for cheating. Results may inform testing program actions or policies when confronted with cases of potential collusion involving examinees from diverse language backgrounds. The proposed study will

also provide a discussion of theoretical and empirical supports with respect to (1) how (and extent to which) language may be related to response pattern similarity; and (2) potential consequences to testing programs and examinees for ignoring such relationships when conducting forensic investigations.

All analyses will be completed by the end of November 2018. Results will be shared at the conference.

Maynes, D. (2014). Detection of non-independent test taking by similarity analysis. In N. M. Kingston & A. K. Clark (Eds.), *Test fraud: Statistical detection and methodology*. New York: Routledge.

Mueller, L., Zhang, Y., & Ferrara, S. (2017). What have we learned? In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (pp. 47–69). New York: Routledge, Taylor & Francis Group.

van der Linden, W. J., & Sotaridona, L. (2006). Detecting answer copying when the regular response process follows a known response model. *Journal of Educational and Behavioral Statistics*, 31(3), 283–304.

27. Japanese EFL Learners' Speech-in-Noise Listening Comprehension Process: Use of Context Information

Ryoko Fujita, Juntendo University

1:30 p.m. to 3:00 p.m. Location: Lobby

Background noise significantly affects language learners' listening comprehension. Notably, past studies have suggested that even bilingual speakers who acquired their target language at an early age have poorer listening comprehension than native speakers under noisy conditions (Rogers, Lister, Febo, Besing, & Abrams, 2006; Shi, 2010). Field (2008) argued that listeners need to draw heavily on context information to recognize words. Although some past studies have focused on background noise and listening comprehension, few have been conducted in the EFL context.

In a study that focused on Japanese EFL learners, Fujita (2016) found that contextual information aided the participants' listening comprehension when the noise level was moderate; however, their listening comprehensibility deteriorated as noise levels increased. The current study builds on that study, which used a quantitative approach for its experiment. It employed a qualitative method and analyzed the listening comprehension process on a smaller scale by investigating learners' use of context information under various noise conditions.

The participants of this study included seven Japanese undergraduate students whose English proficiency levels were high-intermediate. The Speech-Perception-in-Noise (SPIN) test (Kalikow, Stevens, & Elliot, 1977) was used. The SPIN test includes a list of sentences, with the last word in each sentence serving as the target word. The target word is either predicted with contextual cues or unpredicted without contextual cues. Four signal-to-noise ratio (SNR) conditions (SNR = 0, 5, 10, 15) and also without a noise condition were added to the SPIN test.

Data were collected using think-aloud protocol procedures. The participants were asked to verbally report what they were thinking during the SPIN test. After the listening session, they were individually interviewed, and they answered questions about the background noise and their use of context information in listening. The think-aloud protocol data as well as their answers to the SPIN test were carefully examined.

The current study showed several findings. First, EFL learners could make use of context information in all noise levels except for the SNR=0 condition, which the noise level was the highest. Second, the tolerance level of the noise varied among the participants. Third, the level of noise did not affect how leaners used contextual information except for the SNR=0.

28. Certifying language ability for immigration purposes in Switzerland

Peter Lenz, Institute of Multilingualism, in collaboration with the Swiss State Secretariat for Migration and the FIDE Secretariat

1:30 p.m. to 3:00 p.m. Location: Lobby

Since around 2004, language requirements in the immigration context have proliferated in Western European countries (Pochon-Berger & Lenz, 2014). In Switzerland, language knowledge was not absolutely crucial for immigration until recently. In 2018, language proficiency requirements became part of the national naturalization law. Further Language requirements are being introduced for residency. In order to facilitate administrative procedures, the authorities created the 'Language Passport', which summarizes the crucial information in the shape of Council of Europe levels, separately, for spoken and written language use.

The national regulation provides for three possibilities to attain the desired entry in the Language Passport:

1) The fide Language Certificate (officially «Proof of Language Competency») – a language exam that was commissioned by the federal authorities to cover a wide range of uses mainly in the immigration context

2) A general-purpose language certificate that meets pre-defined quality requirements

3) A 'Validation Dossier' providing individual evidence of language proficiency at or above the minimally required level(s). The claims made in the Dossier are verified by means of a reduced-length exam

Our poster presentation focuses on the properties of the fide Language Certificate. This certificate is part of the fide system initiated and owned by the State Secretariat for Migration, aiming at promoting language learning and teaching and establishing high quality standards in the field.

The specific features of the exam include

- accessibility
- separate assessment of oral and written skills
- adaptivity
- closeness to everyday communicative tasks

The fide Certificate intends to be low-threshold, i.e. accessible for a very diverse group of young adults and adults. In principle, no specific exam preparation should be necessary. At the beginning of each exam, the participants are informed by the interlocutors about the form of the tasks and what they are expected to do.

The oral examination is completely separate from the written examination. Speaking and listening do not depend on writing or reading skills. Even illiterate candidates should be able to pass the oral exam. The written test papers (reading and writing) are administered in a group setting, with test facilitators present to keep the candidates on track if necessary.

The fide Certificate is geared towards the levels A1-B1 and is adaptive to some degree: based on candidates' performance on the first speaking tasks, the exam branches off towards A1-A2 tasks or A2-B1 tasks. Whenever possible, the choice of the difficulty range is taken in accordance with the candidate. Naturalization requires B1 in the oral skills and A2 in the written skills, illiterate persons being exempt from the written section. For other immigration statuses, the requirements will be lower.

The tasks in each version of the fide Certificate cover several everyday contexts that were identified as most crucial in a needs analysis. Experience to date shows that less familiarity with some tasks is usually counterbalanced by more familiarity with others.

Test validation is still ongoing during the introductory phase.

29. Comparing rater and score reliability under holistic and analytic rating scales in assessing speech acts in L2 Chinese

Shuai Li, Georgia State University

1:30 p.m. to 3:00 p.m. Location: Lobby

Holistic and analytic rating scales are both widely used in assessing second language (L2) performance. Researchers have reported mixed findings regarding the effects of different types of rating scales on rater reliability, score reliability, and rating processes (e.g., Barkaoui, 2010; Metruk, 2018; Zhang et al., 2015). However, this ongoing debate on the functions of the two types of rating scales has not been extended to the realm of assessing L2 pragmatic competence, which is a key component of communicative language ability (Bachman & Palmer, 2010). Meanwhile, although both holistic and analytic scales have been adopted in L2 pragmatics research since the inception of the field (Taguchi & Roever, 2017), no study has compared the two types of rating scales on rating outcomes and/or rating processes. The current study thus represents an initial effort in this direction and specifically focuses on rater and score reliability under holistic and analytic rating scales in assessing speech acts in L2 Chinese.

In this study, the examinees were 109 learners of Chinese whose Chinese proficiency fell in the range of intermediate-mid to advanced-low based on the ACTFL standards. The examinees completed an Oral Discourse Completion Task (Oral DCT) consisting of 12 scenarios, which were evenly distributed across three types of speech acts, namely, requests, refusals, and compliment responses. The raters were four native Chinese speakers trained in applied linguistics, who evaluated the examinees' oral productions based on five-category holistic and analytic rating scales. The holistic rating scale simultaneously tapped three dimensions in oral speech act production: communicative function, situational appropriateness, and grammaticality; on the other hand, the analytic rating scales separately examined the three dimensions. In assessing the oral productions, the order of rating scale use was counterbalanced, that is, two raters first evaluated the entire dataset with the holistic rating scale, and after two weeks evaluated the same dataset (with a different order of data entries) with the analytic rating scales; the other two raters evaluated the same dataset during the same period of time, but they first used the analytic rating scales before using the holistic rating scale.

Data analysis is currently underway and is planned to focus on the following two sub-questions: (1) Is there any difference between holistic and analytic scales in rater reliability? (2) Is there any difference in ratings under holistic versus analytic scales. The author expects to complete data analysis in early 2019 and share preliminary findings at the LTRC conference for suggestions on improvement.

30. Exploring raters' perceptions of Oral Proficiency Interview Tasks as "promotion" or "demotion"

Jeremy Ray Gevara, Defense Language Institute Foreign Language Center

1:30 p.m. to 3:00 p.m. Location: Lobby

This study examines the raters' perception of Oral Proficiency Interview (OPI) tasks. The OPI is used by numerous educational and occupational institutions, and therefore, the way each institution as well as various stakeholders understand and administer the test has a serious implication in socially just test administration. Particularly concerning and relevant to the current study is the role of raters in the administration of the test. Research shows that the interaction between raters and tasks can explain as much as 63% of a test's variance, meaning that raters' task selections play a majority role in test takers' performances. as they relate to the final holistic decision. To this end, this study aims to answer the following research questions: RQ1) What tasks do raters perceive as significantly contributing to level assignments? RQ2) Do raters perceive any tasks as significantly contributing to placement beyond their assigned levels? To answer these research questions, I analyzed a dataset of 18 raters' task level evaluations of 458 test takers at an occupational training institution in the western United States in 2017. For each task, raters assign a score on the same scale as their holistic decisions. The range of possible scores are Level 0 (Beginning Proficiency) to 4 (Advanced Professional Proficiency) with half step designations (Plus Levels). Using logistic regression, I analyzed 19 tasks from Levels 1 to 4 as predictors with overall holistic decisions as outcomes. For answering RQ1, the results show that raters only perceived Level 2 tasks as significant predictors of their assigned level. For answering RQ2, raters perceived significant several tasks that were beyond the tasks' assigned levels. Although all of the tasks within Level 2 were significant predictors of level assignment, raters perceived the Level 1 Role Play task as the strongest predictor of overall performance (exp(B) = 10.102). "Promotion" tasks are especially prominent for the plus level decisions (Plus levels do not have assigned tasks). Opposite but still similarly significant to "promotion" tasks, raters perceived several tasks as predictors of lower level decisions. One example of this is that three Level 3 tasks were significant predictors of a Level 1+ decision. This result is especially meaningful when knowing only one task of the next higher level (2) was a significant predictor of a Level 1+ decision. Another example of "demotion" tasks are two Level 4 tasks being significant predictors of an overall Level 2+ decision. These "demotion" tasks are expected in this study's OPI, but the results show raters perceived tasks that are two steps above the overall decision as valuable. This result is move deserving of future studies considering raters perceived no significant "promotion" tasks that were two steps above the overall decision. In group conversations, I would like to share ideas about potential future studies that could explore why raters have these perceptions about tasks. Finally, I would like to discuss with the group potential methods to examine raters' overall perceptions of each level in order to explain why some levels had no significant task predictors.

31. Mapping the Path to Advanced Second Language Literacy in Adults Using Eye-Tracking: A Look at Portuguese

Troy Cox, Brigham Young University, **Larissa Grahl**, Brigham Young University, **Logan Blackwell**, Brigham Young University

1:30 p.m. to 3:00 p.m. Location: Lobby

Researching second language (L2) reading poses particular challenges. L2 students and learning contexts vary widely. Moreover, L2 students have much wider ranges of language proficiencies, unlike most L1 readers who have considerable implicit linguistic knowledge by the time they begin to read. Many have already learned to read in their L1 more or less successfully. However, L1 reading strategies and processes can either facilitate the transfer of reading skills or become a source of interference. As a basic example, eye tracking research has found that proficient L1 English readers tend to make larger eye movements (saccades) than proficient L1 Chinese readers (Liversedge et al., 2016). This change reflects properties of the writing system: Chinese characters, for example, are much more informationally and visually dense than English words, so the eyes need to move differently for maximum efficiency and comprehension. As another example, proficient English readers typically fixate just to the left of word center. However, in languages with complex morphological structure (i.e. lots of informative suffixes on words, such as Russian), readers adjust where they look within the word to ensure that they can clearly see the suffixes (Yan et al., 2014). English speakers who fail to make this adjustment will miss important information. These examples illustrate that habits developed for successful L1 reading may not transfer well to the new language, instead interfering with efficient L2 reading.

Eye tracking can overcome these challenges better than other assessment methods. In proficient readers, the eyes move rapidly from one word to another. Because the eyes must point at an object to acquire high-quality visual information about it, what a person is looking at reliably reflects what they are thinking about. Eye tracking can measure what the reader is doing moment-to-moment, rather than simply measuring the final output of the reading process as most currently-employed reading assessments do. For this reason, eye-tracking has proven to be a superior tool for investigating reading (Rayner, 1998, 2009).

Offline measures of L2 reading proficiency (i.e., comprehension questions) cannot identify these challenges or assess success in overcoming them. However, eye tracking can, meaning that it has tremendous potential for assessing L2 reading proficiency. When reading in an L2, readers' eye movements change; adult bilinguals read less efficiently in their L2 than in their L1 (Cop, Drieghe, et al., 2015). These findings suggest that eye tracking can be used to assess L2 proficiency as well.

We will present the methodology of a study in which participants comprised of native and non-native speakers of three different non-linguistically-related languages with different writing systems—Chinese, Portuguese, Russian—were given reading proficiency tests and then used eye-tracking technology to measure the eye movements of the individuals reading in their L1 and their L2. ESL students from those same language backgrounds were administered the same instruments. For this works-in-progress session, we will present the findings of the Portuguese portion of this larger project.

Georgia State University

APPLIED LINGUISTICS & ESL

Welcomes you to... Atlanta

About our department:

We are a multifaceted applied linguistics department that focuses on post-secondary/adult language learning, teaching, and use. Our faculty specialize in a number of subdisciplines including second language (L2) acquisition, L2 writing, sociolinguistics, language assessment, corpus linguistics, educational technology, and L2 teacher education.

From our Alumni:

"The PhD program in applied linguistics, with its world-class faculty, equipped me with the cognitive, methodological, and pedagogical tools necessary to become both a teacher and scholar." – Joe Lee, Ohio University, GSU AL Ph.D. Graduate

"The Department of Applied Linguistic and ESL was fundamental in preparing me for my current position and for my wider involvement in the field of English language teaching." - Heather Hobson, Kennesaw State University, GSU AL MA Graduate

Programs Offered:

- Master of Arts in Applied Linguistics
- Doctor of Philosophy in Applied Linguistics
- B.A. in Applied Linguistics
- Graduate Certificate in Teaching English to Speakers of Other Languages
- Accredited Intensive English Program (IEP)

Ready for the next step in your career? Apply now!

CONTACT:

25 Park Place, 15th floor
Atlanta, GA 30302-4099, U.S.A.
Tel: 404-413-5200
www.gsu.edu/alesl
www.facebook.com/ALESLatGSU

SUPPORTING RESEARCH AND LEARNING IN ASSESSMENT

The British Council supports research and learning in assessment in a number of ways. The Assessment Research Group offers research grants, provides free learning materials, publishes reports and undertakes research projects. You can access free training videos, materials and publications from the website links below.

Offering awards and grants

The British Council offers a range of research awards and grants for research in language assessment.

Assessment Research Awards and Grants: These awards and grants recognise achievement and innovation within the field of language assessment. They are aimed at both research students and more experienced researchers in the area of assessment. See: **www.britishcouncil.org/exam/aptis/research/assessment-advisory-board/awards**

Reading into Research Grants: MetaMetrics and the British Council Assessment Research Group invite applications for research which will contribute to our understanding of the construct of EFL reading comprehension and reading comprehension assessment. Grants are offered to qualifying institutions and/or individuals. **www.britishcouncil.org/research-reading-grants-scheme**

Providing research publications

We publish research across a range of assessment topics and areas of expertise.

- **Technical Reports:** primarily focused on the test development and validation studies related to the Aptis test system.
- **Assessment Research Awards and Grants Reports:** projects carried out by external researchers that have been funded through our awards scheme.
- **Non-Technical Summaries of ARAG Reports:** short, non-technical overviews of the Assessment Research Awards and Grants Reports listed above.
- **British Council Validation Series:** studies in collaboration with external researchers to target areas of importance for Aptis and for language assessment generally.

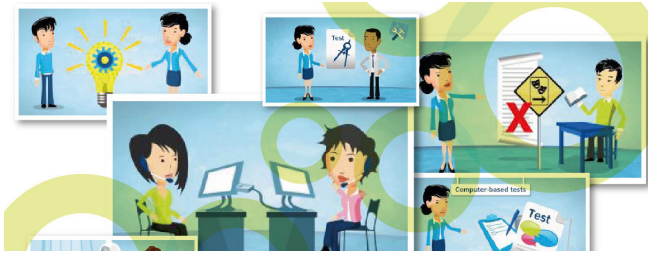
These publications can be downloaded free at: **www.britishcouncil.org/exam/aptis/research/publications**

Collaborative research with China

In 2016, the Ministries of Education of the UK and China agreed to conduct collaborative research on linking various English language tests to China's Standards of English Language Ability. The National Education Examinations Authority (NEEA), Ministry of Education, China and the British Council were appointed to implement this joint programme. The two-year project brought together experts from the British Council's Assessment Research Group, Cambridge Assessment English, NEEA and leading Chinese universities to develop a model of best practice for linking examinations to the CSE.

Training: free videos and glossary

How Language Assessment Works is a project providing free information, materials and training on language assessment. Our short, animated videos give you an insight into some of the main topics. The practical skills topics have accompanying worksheets and answer keys.



We have recently published a Glossary, consisting of hundreds of definitions of terms related to language assessment. Experienced practitioners wrote the definitions, with language teachers in mind.

Watch the videos or download a free copy of the Glossary at: **www.britishcouncil.org/exam/aptis/research/assessment-literacy**

Evaluating English language capability

The English Impact projects 2017 surveyed the English language capability of a representative sample of 15-year-old students in four countries: Sri Lanka, Bangladesh, Spain (Madrid) and Colombia (Bogota). Capability looks at both current achievement and future opportunity to succeed. The projects provided a comparable baseline showing levels of English language capability in publicly-funded schools, which is where government policy makes an impact.

To find out more or to download the reports, see: **www.britishcouncil.org/exam/aptis/research/english-impact**

WIDA™

WIDA provides proven tools and support to help educators and multilingual learners succeed. We strive to provide multilingual learners with an equitable opportunity to obtain a quality education that reinforces learners' assets.

Read the Latest WIDA Research Publications at wida.wisc.edu

Recent Research Reports

Exploring the Long-term English Learner Population Across 15 WIDA States
What are the characteristics of students who remain in EL support programs for longer than average?

Recent WIDA Focus Bulletins

Language-Focused Family Engagement
Language-focused family engagement keeps the unique needs and experiences of multilingual learners and their families at the center of home-school interaction.

More about WIDA

WIDA is based on the campus of the University of Wisconsin-Madison. WIDA offers research publications designed for various audiences, from scholars, to policy-makers, to classroom teachers. Our research spans diverse focus areas:

- Academic language and literacy across content areas
- Psychometrics
- Education policy

Wisconsin Center for Education Research
SCHOOL OF EDUCATION
UNIVERSITY OF WISCONSIN–MADISON

WIDA is housed within the Wisconsin Center for Education Research at the University of Wisconsin-Madison.

MICHIGAN LANGUAGE ASSESSMENT

LTRC 2019

A nonprofit joint venture between the University of Michigan and Cambridge Assessment English, we build on a long history of excellence in applied linguistics and language testing to deliver a range of quality English language tests.

Selected Research Publications

Linking the MET Go! and the Common European Framework of Reference

Pearce, S., McLain, P., Clark, T., & Haines, S. (2018)

Michigan Language Assessment Technical Report, Ann Arbor, Michigan

Development of the MET Go! Speaking Test

Michigan Language Assessment (2018)

Michigan Language Assessment Technical Report, Ann Arbor, Michigan

Development of the MET Go! Writing Test

Michigan Language Assessment (2018)

Michigan Language Assessment Technical Report, Ann Arbor, Michigan

Creating a Listening Test to Track Young Learners' Language Development

McLain, P., Miller, B., Saulter, R., & Stucker, J. R. (2018)

Poster Presented at 2018 MwALT, Madison, Wisconsin

Approaches to Scoring Task Completion in Speaking Performances

Piotrowski, J., Basse, R., Miller, B., & Pearce, S. (2018)

Poster Presented at 2018 MwALT, Madison, Wisconsin

Determining Seat Time For a Shortened Section of a High-Stakes Standardized Test

May, N., Mclain, P., & O'Connell, S. (2018)

Poster Presented at 2018 LARC, Ames, Iowa, and at 2018 LTRC, Auckland, New Zealand

LTRC 2019 Poster Presentations

Familiarizing Standard-Setting Panelists with the CEFR: A Three-Step Approach to Attaining a Shared Understanding of Just-Qualified Candidates

Pearce, S., McLain, P., & Clark, T. (2019)

Poster 2019 LTRC, Atlanta, Georgia

Beyond the Test Score: Developing Listening Test Feedback and Activities to Empower Young Learners and Teachers of English

Miller, B., Slisz, L., McLain, P., & Stucker, J. R. (2019)

Poster 2019 LTRC, Atlanta, Georgia

Research & Internship Opportunities

Michigan Language Assessment conducts and supports research to inform test development and revision, to provide evidence of quality and validity for its tests, and to contribute to knowledge in the language testing field. Visit our website to learn more about summer internships in language assessment and about funding opportunities under the **Spaan Research Grant Program** and **Latin America Research Grant Program**. Michiganassessment.org/about-us/research

Cambridge Assessment English

UNIVERSITY OF MICHIGAN

MichiganAssessment.org

Pearson
PTE Academic

The English test for international study

The most accurate and objective test of academic English available.

PTE Academic is a computer-based language test that offers international students the fastest, fairest and most flexible way of proving their English language proficiency for university admission.

- **Secure** testing worldwide ensures test score validity.
- **Recognised and accepted** as proof of English proficiency by thousands of institutions worldwide.
- **Testing globally** every week in over 250 test centres around the world.
- **Results** typically within five business days.

Visit **www.pearsonpte.com** to learn more

Pearson will be presenting the following sessions at LTRC 2019:

Linguistic tools in writing assessment: Their impact on test-takers' writing process and performance

Presented by: **Saerhim Oh**

What aspects of speech contribute to the perceived intelligibility of L2 speakers?

Presented by: **Bill Bonk and Saerhim Oh**

42nd Language Testing Research Colloquium

LTRC 2020 TUNISIA

البحوث في التقييم اللغوي

Assessment in multilingual contexts:

Models, practices, policies & challenges

Hammamet, Tunisia

9-13 June 2020

Join us as we take LTRC to Africa for the first time!

Venue:
Medina Convention Center
close to hotels, restaurants, bazaars, and the beach

Less than 1 hour from main airports and 1-2 hours to major cities and attractions

Everything you've come to expect from LTRC and much more!

Cambridge Assessment English

The Story of Assessment

Studies in Language Testing

www.bit.ly/storyofassessment

ETS TOEFL

Quality Beyond Measure.

More than four decades of research supports the validity of the *TOEFL*® Family of Assessments

Domain description: Validating the interpretation of *TOEFL iBT*® Speaking scores for ITA screening and certification purposes. (In press.) Cotos, E., & Chung, Y. *TOEFL iBT Research Report*. Princeton, NJ: Educational Testing Service.

Interpreting the relationships between TOEFL iBT scores and GPA: Language proficiency, policy, and profiles. (2018). Ginther, A., & Yan, X. *Language Testing*, 35(2), 271–295.

Out of many, one: Challenges in teaching multilingual Kenyan primary students in English. (2017). Hsieh, C., Ionescu, M., & Ho, T. *Language, Culture and Curriculum*, 31(2), 199–213.

Assessing syntactic sophistication in L2 writing: A usage-based approach. (2017). Kyle, K., & Crossley, S. *Language Testing*, 34(4), 513–535.

Young learners' response processes when taking computerized tasks for speaking assessment. (2018). Lee, S., & Winke, P. *Language Testing*, 35(2), 239–269.

Comparability of students' writing performance on TOEFL iBT and in required university writing courses. (In press.) Llosa, L., & Malone, M. E. *Language Testing*.

Adding value to second-language listening and reading subscores: Using a score augmentation approach. (2018). Papageorgiou, S., & Choi, I. *International Journal of Testing*, 18(3), 207–230.

Do the TOEFL iBT® section scores provide value-added information to stakeholders? (2017). Sawaki, Y., & Sinharay, S. *Language Testing*, 35(4), 529–556.

Screener tests need validation too: Weighing an argument for test use against practical concerns. (2018). Schmidgall, J., Getman, E., & Zu, J. *Language Testing*, 35(4), 583–607.

Using corpus-based register analysis to explore authenticity of high-stakes language exams: A register comparison of TOEFL iBT and disciplinary writing tasks. (2018). Staples, S., Biber, D., & Reppen, R. *The Modern Language Journal*, 102(2) 310–332.

www.ets.org/toefl/research

ETS — *Measuring the Power of Learning.*®