ILTA
INTERNATIONAL LANGUAGE TESTING ASSOCIATION

Universidad de los Andes
Colombia

LANGUAGE ASSESSMENT LITERACY ACROSS STAKEHOLDER BOUNDARIES

LTRC2017

39TH LANGUAGE TESTING RESEARCH COLLOQUIUM

JULY 17–21, 2017 · BOGOTÁ, COLOMBIA

IELTS™

IELTS Research Reports Online Series

IELTS Partnership Research Papers

available now at

ielts.org/research

BRITISH COUNCIL

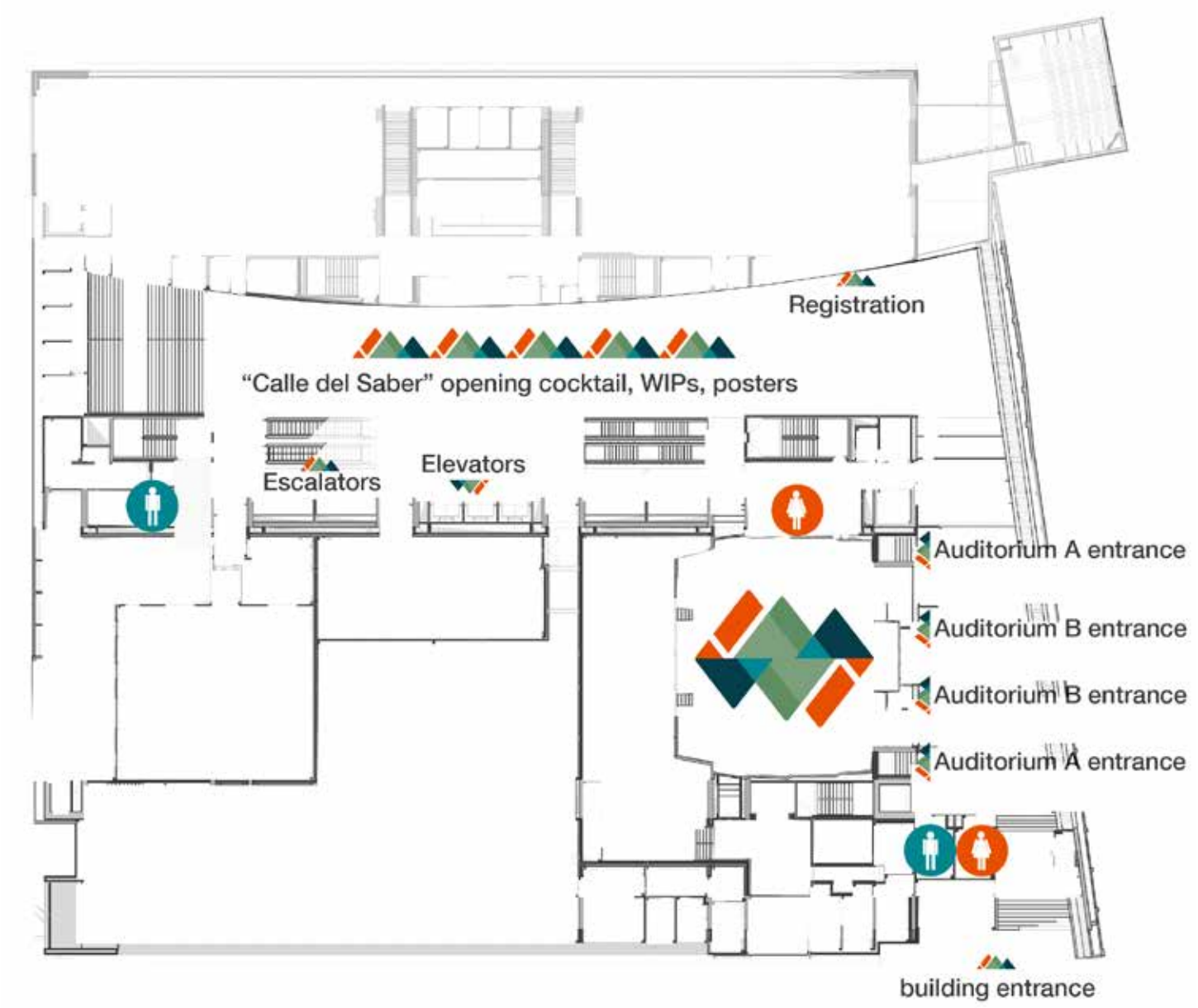
idp IELTS AUSTRALIA

CAMBRIDGE ENGLISH Language Assessment
Part of the University of Cambridge

CONTENT

CAMPUS MAPS

2nd floor · Mario Laserna (ML) / registration, opening reception, auditoriums, WIPs, posters



5th floor · Mario Laserna (ML) / pre-conference workshops, exhibitors, coffee breaks, lunch, terraces

6th floor Mario Laserna (ML) / parallel sessions

WELCOME MESSAGE FROM THE ILTA PRESIDENT

¡BIENVENIDOS TODOS Y TODAS A COLOMBIA! I am delighted to welcome you all to the 39th Language Testing Research Colloquium (LTRC) held for the first time ever in Latin America. The theme of the conference 'Language Assessment Literacy Across Stakeholder Boundaries' is particularly appropriate for a conference held in this new location, reminding us of the International Language Testing Association (ILTA)'s mission not only to share understandings and stimulate professional growth amongst its members but also to cooperate with and learn from others in contexts very different from our own, including those where English is not the primary language. I am particularly happy to recognize the Department of Languages and Culture, here at Universidad de los Andes, for the impressive investment they have made in ILTA's mission: they have organized six pre-conference workshops about the basics of language assessment supported by both CaMLA and Goethe Institut. We also very much appreciate the opportunity LTRC 2017 will afford us to hear about local assessment and evaluation issues from two invited speakers from Colombian universities, Ana María Velásquez and María Lucía Casas Pardo. And we are particularly pleased to welcome the strong contingent of Colombian and Latin American participants at this conference.

The conference co-chairs Beverly Baker, Jee Wha Dakin, Gerriet Janssen, Heike Neumann, and Isabel Tejada Sánchez have worked tirelessly over the last year to assemble a stimulating program of workshops, symposia, papers, posters and work-in-progress sessions, which tackle language assessment literacy from multiple perspectives. ILTA is extremely grateful to them for their efforts. Many presentations address the conference theme directly, attempting to define the scope of language assessment literacy, to consider the extent to which it varies across contexts, to report on conceptions of assessment and assessment practices among different groups and to document or evaluate different approaches to sharing understandings or delivering assessment information to users. Others contribute more broadly to our professional knowledge by reporting on the methodologies and findings of recent research on language tests designed for a range of purposes and users. We particularly look forward to the Messick Award lecture by Stephen Sireci and the Davies lecture by Ofra Inbar-Lourie and thank our sponsors (ETS and the British Council) for making these two events possible.

Note that if you are a first-time participant you will find it helpful to attend Tuesday afternoon's Newcomer Session (sponsored by the Language Learning and Testing Foundation) where you will

Plaza de Bolívar
Bogotá, Colombia.

meet other newcomers, receive some background information about ILTA as well as advice from old hands on how to get the most out of the LTRC.

ILTA provides many opportunities for professional growth in addition to our annual conference. Our new ILTA website links you to a range of language assessment resources including a Code of Ethics (translated into multiple languages) and the annually updated bibliography of testing publications and doctoral dissertations (coordinated by Tineke Brunfaut). Some of these resources are available to members only, so if you have not joined our organization already, now is the time! (http://www.iltaonline.com/page/MembershipPlans)

ILTA also funds assessment workshops and awards, with international testing agencies as co-sponsors in some cases. You will hear more about these initiatives and meet some of the winners at the LTRC Banquet on Thursday evening. And you can find out more about ILTA activities, including plans for future conferences, at our Annual Business Meeting from 12.20-2pm the same day, which I urge all members to attend. With a new management company ILTA is well placed to develop a more active voice in language assessment issues world-wide and it is important that you play your part in discussing how our organization might develop.

I look forward to greeting old friends and meeting new colleagues here in Bogotá.

Cathie Elder
ILTA President 2017

LETTER FROM LTRC CO-CHAIRS

On behalf of the co-chairs, welcome to the 39th LTRC, to Colombia, to the city of Bogotá, and to the Universidad de los Andes campus! We are very excited about the 2017 colloquium program, and especially the focus of extending the conversation about language assessment literacy across typical stakeholder boundaries into new and different contexts. This discussion starts during the opening symposium *Towards an ILTA policy of assessment literacy*, and concludes with Ofra Inbar-Lourie's Davies Lecture, *Language assessment literacies and the language testing community: A mid-life identity crisis?* We are also extremely pleased to welcome Stephen G. Sireci from the University of Massachusetts Amherst, USA, as this year's Messick Lecture Award speaker with his talk *How would Messick validate 21st-century language assessments?*

We are happy that LTRC 2017 comes to Colombia at a very important time in its history with the current peace process. It is our belief that a wide variety of stakeholders here in Colombia (politicians, educators, ex-combatants, and citizens in both rural and urban areas) are now particularly ready to consider new conversations about important themes such as education, additional languages, and the role that assessment can play in the educational process and political process of peace-building. We also believe that it is vital that this conversation be bi- or multi-directional, such that educational initiatives consider local conditions and empower local actors to seek meaningful and realizable outcomes. It is our hope that you get a glimpse of these developments here in Colombia with the invited plenary talks on Wednesday morning given by Ana María Velásquez from the Universidad de los Andes on *Educational challenges in Colombia: The role of language learning in peace building* and by María Lucía Casas, Rector of the Institución Universitaria Colombo Americana ÚNICA, Bogotá, *Using assessment to add value in higher education*.

We also hope that you will be able to enjoy the city of Bogotá while here. As with most major urban areas in Latin America today, it is important to always move about the city with a friend or colleague, take good care of any personal items and valuables. With this in mind, we invite you to take advantage of the great attractions the city has to offer. Near the conference site, be sure to check out the Gold Museum (el Museo de Oro), the Botero Museum (la Donación Botero), and Monserrate (Bogotá's proud symbol with the white church on top), or stroll through the historic district, la Candelaria. In the recommended hotel district, be sure to explore the incredible variety of restaurants that has made the city of Bogotá one of the emerging culinary capitals of Latin America. Whenever you move about the city, remember to always bring a sweater and water bottle as it can get really cold quickly and we are at a high altitude (Bogotá is one of the highest capitals in the world).

Finally, we hope you will fall in love with Universidad de los Andes for its historical campus, but also for its innovation and current social relevance. We are deeply grateful for the continued and fantastic support that the Universidad de los Andes has provided

Carrera Séptima.
Bogotá, Colombia.

towards the realization of LTRC in terms of the physical, human, and financial resources we have been afforded. Most touching has been the Department of Languages and Culture's commitment to LTRC: Along with CaMLA and the Goethe Institut, this department sponsored six assessment workshops so that instructors of Brazilian Portuguese, English, French, German, Japanese, and Spanish as a foreign language could meet each other and then consider assessment themes that are relevant for their teaching community. Be sure to walk around our campus, take a ton of pictures, and learn more about what los Andes is doing (https://uniandes.edu.co/en).

Questions about anything? Ask any one of the professor-volunteers, student assistants, or conference co-chairs wearing an orange name badge, and we will be glad to help.

¡Bienvenidos a LTRC 2017! Es todo un gusto tenerlos en casa y deseamos que disfruten este congreso y que pasen una buena estancia en la ciudad de Bogotá.

Yours sincerely,

The LTRC 2017 co-chairs

Beverly Baker, Jee Wha Dakin, Gerriet Janssen, Heike Neumann, and Isabel Tejada Sánchez

INVITED SPEAKERS

Special Invited Plenary: Education, Language, and Assessment in the Colombian Context

Wednesday, July 19th, 8:45–9:45

Ana María Velásquez, Universidad de los Andes, Colombia

1. Educational challenges in Colombia: The role of language learning in peace building

Colombia is currently facing a postconflict context that demands us to reconsider the purposes of education in terms of the citizens our society needs to rebuild peace. Among the various means we can consider to achieve this goal, one of them is to widen the perspective of our citizens to understand how we can live in dialog with our worldwide cultural differences, as well as learning from the experiences of other cultures who have faced similar postconflict challenges. Language learning and testing may play an important role in this process, and this talk will present some ideas on how second-language researchers and practitioners may contribute to this goal.

Ana María Velásquez is an associate professor and chair of the undergraduate program in the School of Education at Universidad de los Andes, Colombia. She holds a B.A. in Psychology, an M.A. in Education from Universidad de los Andes, and a Ph.D. degree in Developmental Psychology from Concordia University, Canada. Her areas of expertise are related to socio-emotional development, classroom social climate, and citizenship education programs. Her research is focused on examining the relationship between the classroom climate and socio-emotional development. Dr. Velasquez has participated in the design and evaluation of the violence prevention program Classrooms in Peace (Aulas en Paz). Also, she has collaborated with the Ministry of Education of Colombia and the Secretary of Education of Bogota, in the development of citizenship competencies curricula and evaluation. She has been the author of several peer-reviewed and invited publications on these topics and she participated as a co-editor of the book "Citizenship Competencies: From the standards to the classroom."

María Lucía Casas Pardo, Institución Universitaria Colombo Americana ÚNICA, Colombia

2. Using assessment to add value in higher education

ÚNICA's fundamental mission has been to become the best teacher's college in Colombia. From the day it began, we tackled this mission by providing students from the most marginalized backgrounds—from all corners of the country—with the opportunity to excel academically. After 13 years, ÚNICA's pedagogical model and continuous assessment of both teachers and students has allowed us to become the top bilingual school and teacher college in the country. ÚNICA implements assessment tools that have given high-quality results; this year our students ranked number one

La Candelaria.
Bogotá, Colombia.

in English on the Saber-Pro exams. Additionally, ÚNICA was deemed the second best, after the National University, in Education. These results prove that a thorough pedagogical model, top-quality teachers, and rigorous assessment and evaluation processes have the power to turn underprivileged youth into the best bilingual teachers in Colombia.

María Lucía Casas Pardo is the Rector of ÚNICA, the first bilingual teachers college in Colombia. She earned an undergraduate degree in Modern Languages and Education and a Master's degree from Universidad Javeriana, Bogotá. Ms. Casas has worked as a teacher at the school level, and as a Professor of English and Pedagogy at Universidad Javeriana. She was Academic Director of the Colegio Gimnasio Campestre and IB Coordinator in Colegio Gimnasio del Norte. She has also been the Dean of the Business School and Director of the International Relations Office in CESA, Director of Education in Compensar and the coordinator of the National Bilingualism Committee for over two years. She has acted as advisor for the National Ministry of Education in Colombia, for Universidad Católica de Chile and for various Mexican organizations. Ms. Casas has been a lecturer in national and international events mainly in the topics of bilingual education, teacher education and educational policies.

The Samuel J. Messick Memorial Lecture

Sponsored by Educational Testing Service

How Would Messick Validate 21st-Century Language Assessments?

Stephen G. Sireci, University of Massachusetts Amherst, USA

Thursday, July 20th, 4:00–5:00

Through his writings, Samuel Messick was clear that validation involved a comprehensive evaluation of the empirical evidence and theory that supports the use of test scores for their intended purposes. His perspectives persevere today as embodied in the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014), which stipulate five sources of validity evidence "that might be used in evaluating the validity of a proposed interpretation of test scores for a particular use" (p. 13). In this presentation, I illustrate the link between Messick's writings and the AERA et al. (2014) Standards, and illustrate how the validation framework proposed by the Standards can be used to support the validity of the intended uses of contemporary language assessments test scores. Contemporary language assessments are used for many different purposes in education, industry, and everyday life. These purposes range from very high stakes, such as placement into instructional programs in elementary, middle, and secondary school (e.g., English language proficiency assessments for public school students in the United States); to very low stakes such as self-assessment of second language proficiency. A 21st-century framework for validating language assessments (a) clearly articulates the intended purposes of a test, (b) specifies appropriate and inappropriate uses of the test, (c) identifies the sources of validity evidence needed to support intended purposes and evaluate unintended consequences, and (d) synthesizes the validity evidence in a comprehensive and comprehensible validity argument.

In this presentation I will give examples of how validation frameworks can be developed for language assessments that are consistent with the AERA et al. (2014) Standards, and with Dr. Messick's validity theory. By following Dr. Messick's guidance, and the guidance provided by the 60-year history of the Standards, we will not only develop strong validity arguments to support the use of language tests in specific situations, we will also develop better language tests.

Stephen G. Sireci is Professor and Director of the Center for Educational Assessment in the College of Education at the University of Massachusetts Amherst. He earned his Ph.D. in psychometrics from Fordham University and his master and bachelor degrees in psychology from Loyola College in Maryland. Before UMASS, he was Senior Psychometrician at the GED Testing Service, Psychometrician for the Uniform CPA Exam and Research Supervisor of Testing for the Newark NJ Board of Education. He is known for his research in evaluating test fairness, particularly issues related to content

validity, test bias, cross-lingual assessment, standard setting, and sensitivity review. He is the author of over 100 publications and conference papers, and is the primary architect of the Massachusetts Adult Proficiency Tests, which is the primary assessment of reading and math skills used in adult education programs in Massachusetts. He currently serves or has served on several advisory boards including the National Board of Professional Teaching Standards Assessment Certification Advisory Panel, the Texas Technical Advisory Committee, and the Puerto Rico Technical Advisory Committee. He is a Fellow of the American Educational Research Association and a Fellow of Division 5 of the American Psychological Association. Formerly, he was President of the Northeastern Educational Research Association (NERA), Co-Editor of the International Journal of Testing, a Senior Scientist for the Gallup Organization and a member of the Board of Directors for the National Council on Measurement in Education. In 2003 he received the School of Education's Outstanding Teacher Award, in 2007 he received the Chancellor's Medal, which is the highest faculty honor at UMASS, and in 2012 he received the Conti Faculty Fellowship from UMass. He has also received the Thomas Donlon Award for Distinguished Mentoring and the Leo Doherty Award for Outstanding Service from NERA. Professor Sireci reviews articles for over a dozen professional journals and he is on the editorial boards of Applied Measurement in Education, Educational Assessment, Educational and Psychological Measurement, the European Journal of Psychological Assessment, and Psicothema.

The Davies Lecture

Sponsored by the British Council

Language Assessment Literacies and the language testing community: A mid-life identity crisis?

Ofra Inbar-Lourie, Tel-Aviv University

Friday, July 21st, 3:40–4:40

The conceptual framework and dilemmas accompanying Language Assessment Literacy (LAL) study and research can be traced to a large extent to the scholarly legacy that Alan Davies offered the language testing community as a leading theorist and researcher. One of his notable contributions in this sense is the attempt to define, along with other scholars, the fundamentals of the field in the Dictionary of Language Testing (Davies, Brown, Elder, Hill, Lumley & McNamara, 1999), and later on in a review of language testing textbooks (Davies, 2008). In this more recent publication, Davies first described the components of language testing knowledge emerging from the books. He then identified a move towards professionalism amongst language testers, wondering however whether this trend might lead to utilizing only in-house resources, resulting in the insulation of language testing from other "potentially rewarding disciplines".

In this presentation I will further probe the nature of language assessment literacy from this latter perspective, by making a case against insulation in defining the knowledge base required for conducting, analyzing and

making decisions based on assessment data by diverse protagonist groups. I will argue for adopting a pluralistic loose descriptive language assessment literacies framework (rather than taking a prescriptive monolithic literacy approach), where existing paradigms are integrated with other areas of expertise and embedded in particular settings thus allowing for the creation of local scripts. Additionally, I would like to argue that the current focus on LAL as the conference theme, allows for a reflection on identity issues within the language testing community at this point in time, as it heads closer towards its 40th anniversary.

Ofra Inbar-Lourie trained at Tel Aviv University, where she was awarded a BA in American and English Literature, an MA in Education and a PhD in Language Education. She taught at Beit Berl Academic College of Education in the area of EFL teacher education and language assessment, and served as the head of the English department. She was also an EFL inspector for the Ministry of Education, and coordinated the MA TESOL program for overseas students at Tel-Aviv University. Currently she lectures in the Multilingual Education program in the School of Education at Tel-Aviv University on assessment issues, language policy and curriculum design, and since 2011 chairs the Teacher Education Unit in the School of Education.

Her professional experience includes officiating in local and international committees dealing with language policy, language planning, and language testing and teacher education. She has researched, published and presented locally and internationally on a range of language education issues, including language assessment and language assessment literacy, language policy, language teachers, especially with regard to native and non-native speaking background, young language learners and recently on English Medium Instruction.

The Cambridge/ILTA Distinguished Achievement Award

Sponsored by Cambridge ESOL/ILTA

Sauli Jaakko Takala (1941-2017)

This year's Distinguished Achievement award was granted posthumously to Sauli Jaakko Takala and it is with great sadness that we announce his passing.

Sauli Takala was emeritus professor of the University of Jyväskylä, Doctor of Philosophy honoris causa and Doctor of Education honoris causa. Sauli was awarded an MA in Philosophy at the University of Jyväskylä in 1970, obtained a 'Qualified Language Proficiency Degree (English)' in 1974 and went on to obtain his Ph.D. at the University of Illinois at Urbana-Champaign, USA in 1984. From 1964, he worked as a teacher of Swedish and English at secondary schools in Finland and filled research positions at educational research institutes of the University of Jyväskylä. In 1994, he was appointed Professor of methodology at the Institute of Educational Sciences of the University of Jyväskylä and in 1998 obtained the position of Research Professor at the Department of Applied Language Studies at the same university, a position which he held until his retirement in 2002.

For 15 years Sauli was a member of the Finnish Matriculation Examination Board and was the key figure in the introduction of modern quality assurance mechanisms (e.g., item analysis, standard setting) in an examination that had previously been very traditional in its approach. From the 1970s onward he was involved in numerous international projects including IEA's six subject study, the IEA International Study of Writing, the EU DIALANG project and the Council of Europe's Language Project.

Sauli's research activity focused on assessing language proficiency as well as on language planning, policy and curricula. Over the course of his career Sauli published more than 150 research reports and articles in Finnish, Scandinavian and international journals (see: http://kiesplang.fi/S_Takala/Publications.html). He supervised a number of doctoral dissertations and examined dozens of doctoral theses in Finland and other Nordic countries. He also served as editor for the Finnish Journal of Educational Research, as co-editor of the Scandinavian Journal of Educational Research, and as member of the Language Testing Editorial Board.

Sauli was founding member of EALTA (European Association for Language Testing and Assessment) and President of EALTA from 2007 to 2010. He was heavily involved in Council of Europe developments in language education over several decades, participating in developing the widely-used Manual for Relating Language Examinations to the CEFR, and, most recently, serving as member of the expert group for the development of additional CEFR descriptors. In addition, he was consultant to the European Centre for Modern Languages (ECML).

Sauli was a great friend, mentor and colleague, and those who had the pleasure of working with him, enjoyed his scholarship, his gentle attitude and genuine love for the fields of language teaching and assessment. He made a tremendous contribution to the assessment community in his unique considerate way, drawing on his wealth of knowledge and experience and always willing to share his wisdom and his vast collection of books, materials and articles. He had an open mind, a warm heart and a winning personality.

This tribute is drawn from the Award citation prepared by John de Jong, Chair, 2017 Cambridge/ILTA Distinguished Achievement Award.

CONFERENCE SUPPORT

LTRC Organizing Committee 2017

Co-Chairs

Beverly Baker
Jee Wha Dakin
Gerriet Janssen
Heike Neumann
Isabel Tejada Sánchez

Organizing Committee

Martha Patricia Ferreira
Edith Nohemi Flechas
Victor Gómez
Nicolás González
Juan Carlos González
Alejandra Isaza de Larrañaga
Alex Kasula
Ana Patricia Muñoz Restrepo
Pablo Emilio Patiño Delgado
Edgar Picón Jácome
Hugo Hernán Ramírez Sierra
Elsa Adriana Restrepo
Emma Rye
Andrea Sánchez Noguera
Katrina Schmidt
Kathleen Sheridan
Juan Camilo Vargas Plazas
Diana Carolina Vesga

WIP Chair

Candace Farris

Abstract Reviewers

Beverly Baker
Khaled Barkaoui
Vivien Berry
Tineke Brunfaut
Micheline Chalhoub-Deville
Ikkyu Choi
Deborah Crusan
Sara T. Cushing
Jee Wha Dakin
John de Jong
Barbara Dobson
Glenn Fulcher
Evelina Galaczi
Ardeshir Geranpayeh
April Ginther
Liz Hamp-Lyons
Luke Harding
Ching-Ni Hsieh
Talia Isaacs
Gerriet Janssen
Ahyoung Alicia Kim
Gad Lim
Lorena Llosa
Sari Luoma
Heike Neumann
Spiros Papageorgiou
Lia Plakans
India Plough
David Qian
John Read
Daniel J Reed
Jamie Schissel
Jonathan Schmidgall
Rob Schoonen
Charles Stansfield
Ruslan Suvorov
May Tan
Dina Tsagari
Carolyn Turner
Alistair Van Moere
Erik Voss
Elvis Wagner
Yoshinori Watanabe
Mikyung Kim Wolf
Xiaoming Xi
Guoxing Yu
Ying Zheng

Volunteers

Emma Rye (coordinator)
Luciana Andrade
Natalia Benavides
Nicole Bier
Lorena Bustos
Oscar Gómez
Alex Kasula
Hiromu Kondo
Jose Molina
Beatriz Peña Dix
Michael Ropicki
Katrina Schmidt
Douglas Waters

ILTA EXECUTIVE BOARD AND COMMITTEE MEMBERS 2017

ILTA Executive Board 2017

President: Catherine (Cathie) Elder (University of Melbourne, Australia)
Vice-President: Micheline Chalhoub-Deville (The University of North Carolina, Greensboro, USA)
Past President: Anthony (Tony) Green (University of Bedfordshire, UK)
Secretary: Ute Knoch (University of Melbourne, Australia)
Treasurer: Jayanti Banerjee (Paragon Testing Enterprises, Canada)

Members at Large

Beverly Baker (University of Ottawa, Canada)
Talia Isaacs (University College London, UK)
Tineke Brunfaut (University of Lancaster, UK)
Claudia Harsch (University of Bremen, Germany)

ILTA Nominating Committee 2017

Chair: Diane Schmitt (Nottingham Trent University, UK)
Atta Gebril (The American University of Cairo, Egypt)
Gerriet Janssen (Universidad de los Andes, Colombia)
Lia Plakans (University of Iowa, USA)

Award Committees

Cambridge/ILTA Distinguished Achievement Award

Chair: John de Jong (VU University of Amsterdam, Netherlands)
Micheline Chalhoub-Deville (University of North Carolina, US)
Cathie Elder (University of Melbourne, Australia)
Nick Saville (Cambridge English Language Assessment, UK)

ILTA Student Travel Awards

Chair: Cathie Elder (University of Melbourne, Australia)
Tineke Brunfaut (Lancaster University, UK)
Ute Knoch (University of Melbourne, Australia)

ILTA Workshops and Meetings Award

Round 1

Chair: Cathie Elder (University of Melbourne, Australia)
Talia Isaacs (Kings College London, UK)
Eunice Eunhee Jang (OISE, University of Toronto, Canada)

Round 2

Chair: Anthony Green (University of Bedfordshire, UK)
Micheline Chalhoub-Deville (University of North Carolina, USA)
Eunice Eunhee Jang (OISE, University of Toronto, Canada)

Lado Award (to be awarded at LTRC)

Chair: Claudia Harsch (University of Bremen, Germany)
Erik Voss (Northeastern University, USA)
Mikyung Kim Wolf (Educational Testing Service, USA)

ILTA Best Article Award 2015

Chair: Talia Isaacs (University College London, UK)
Claudia Harsch (University of Bremen, Germany)
Scott Crossley (Georgia State University, USA)
John Pill (American University of Beirut, Lebanon)
Alan Urmston (Hong Kong Polytechnic University, Hong Kong)

AWARDS

ILTA Student Travel Awards 2017

Saerhim Oh (Teachers College, Columbia University, US)

Jacqueline Ross TOEFL Dissertation Award

Dr Jing Xu. *Predicting ESL learners' oral proficiency by measuring the collocations in their spontaneous speech.* Supervisor: Dr. Carol Chapelle at Iowa State University, US
Runner up: Dr. Sarah Goodwin. *Locus of control in L2 English listening assessment.* Supervisor: Dr. Sara T. Cushing at Georgia State University, US.

Caroline Clapham IELTS Masters Award, 2016

Dai Wei (David). *Multidialectal listening assessment of young learners*. Supervisor: Carsten Roever at University of Melbourne, Australia.

ILTA Workshops and Meetings Award 2017 (Round 1)

John Pill and Kassim Shaaban. *Strengthening language assessment literacy in tertiary education in Lebanon*
Antony Kunnan. *Assessing World Languages in Macau, SAR China*
Aylin Ünaldı. *Developing language teachers' assessment literacy in classroom-based achievement assessment for young learners in Turkey.*

ILTA Workshops and Meetings Award 2017 (Round 2)

Beverly Baker, Jee Wha Dakin, Gerriet Janssen, and Heike Neumann. *Towards a working group meeting to establish a Latin American Language Assessment Association.*

Claudia Harsch and Ivonne de la Caridad Collada Peña. *Assessment Literacy for Higher Education – setting up a Cuban Network of Language Testers.*

ILTA Best Article Award 2015

Alderson, J. C., Brunfaut, T., & Harding, L. (2015). Towards a theory of diagnosis in second and foreign language assessment: Insights from professional practice across diverse fields. *Applied Linguistics, 36*(2), 236-260.

The Davies Lecture Award

Ofra Inbar-Lourie (Tel-Aviv University, Israel)
Language Assessment Literacies and the language testing community: A mid-life crisis?

Samuel J. Messick Memorial Lecture Award

Stephen G. Sireci (University of Massachusetts Amherst, USA)
How would Messick validate 21st-century language assessments?

Cambridge/ILTA Distinguished Achievement Award

Sauli Takala, University of Jyväskylä, Finland

Robert Lado Memorial Award

To be announced at the LTRC banquet

TIRF 2016 Doctoral Dissertation Grant Awardees in Language Assessment

Christopher van Booven, New York University, Dissertation title: *Assessing Interactional Affordances and Gains in the Study Abroad Homestay and the Language Classroom: A Conversation-Analytic Approach.* Supervisor: Dr. Lorena Llosa at New York University, US.

Hung-Jo Yoon, Michigan State University, Dissertation title: *Investigating the Interactions among Genre, Task Complexity, and Proficiency in L2 Writing: A Comprehensive Text Analysis and Study of Learner Perceptions.* Supervisor: Dr. Charlene Polio at Michigan State University, US.

Naoki Ikeda, University of Melbourne, Dissertation title: *Exploring Construct and Practicality of an Interactional Test for L2 Oral Pragmatic Performance.* Supervisor: Dr. Carsten Roever at University of Melbourne, Australia.

Mikako Nishikawa, University of Bristol, Dissertation title: *Test-takers' Cognitive Processes during Integrated Writing Tasks Which Use Multiple Texts and Graphs as Prompts.* Supervisor: Dr. Guoxing Yu at Bristol University, UK.

Yongfei Wu, Queen's University, Dissertation title: *Relationships among Students' Perceptions of and Reactions to Teacher Feedback, Self-regulation, and Academic Achievement in the Chinese Tertiary EFL Context.* Supervisor: Dr. Liying Cheng at Queen's University, Canada.

Simon Davidson, University of Melbourne, Dissertation title: *Investigating and Revising the Standards Set on the Occupational English Test's (OET) Writing Sub-test.* Supervisor: Dr. Ute Knoch at University of Melbourne, Australia.

CONFERENCE SITE DETAILS

Recommended Hotels

Hotel	Neighbor-hood	Address	Webpage	Contact email
HOTEL EMAUS	Zona G/Rosales	Carrera 4 No. 69A-46 (57-1) 5420091 / (57) 3003158328	www.hotelemausbogota.com	corporativo@hotelemausbogota.com
BH TEMPO	Chapinero	Carrera 7 No. 65–01 (57-1) 7447790 / (57) 321 2336107	www.bhhoteles.com	mariaclaudia.gonzalez@bhhoteles.com
BH QUINTA	Zona G/Rosales	Carrera 5 No. 74–52 (57-1) 7447790 / (57) 321 2336107	www.bhhoteles.com	mariaclaudia.gonzalez@bhhoteles.com
FOUR SEASONS-- CASA MEDINA	Zona G/Rosales	Carrera 7 No. 69–22 (57-1) 3257900	www.fourseasons.com/bogotahotels	Marcela.Ocampo@fourseasons.com
GHL STYLE MIKA	Zona G/Rosales	Calle 70A No. 4–08	www.ghlhoteles.com	carolina.barrios@ghlhoteles.com
HOLIDAY INN AND SUITES	Chapinero	Carrera 7 No. 67–39 (57-1) 6510000 (57) 3102447028	www.hiexpress.com	crojas@oxohotel.com
HOTEL ARTISAN D.C	Zona G/Rosales	Avenida Calle 72 No. 5–51	www.hiexpress.com	crojas@oxohotel.com
HOTEL EMBASSY SUITES	Zona G/Rosales	Calle 70 No. 6–22 (57-1) 3171313 (57) 3102094479	www.embassysuites.com	reservas@embassy-bogota.com.co ventas2@embassy-bogota.com.co
HOTEL CASONA DEL PATIO	Zona G/Rosales	Carrera 8 No. 69–24 (57-1) 2828640 / (57) 3133176534	www.hotelcasadeco.com	reservas@hotelcasadeco.com
HOTEL JW MARRIOTT BOGOTA	Zona G/Rosales	Calle 73 No. 8–60 (57-1) 4816000 / (57) 3214901278	www.marriott.com/bogjw	fabian.maldonado@R-HR.COM
HOTELES BS ROSALES	Zona G/Rosales	Carrera 4 No. 71–22 (57-1) 7447790 / (57) 321 2336107	www.bhhoteles.com	mariaclaudia.gonzalez@bhhoteles.com

Transportation to and from the site

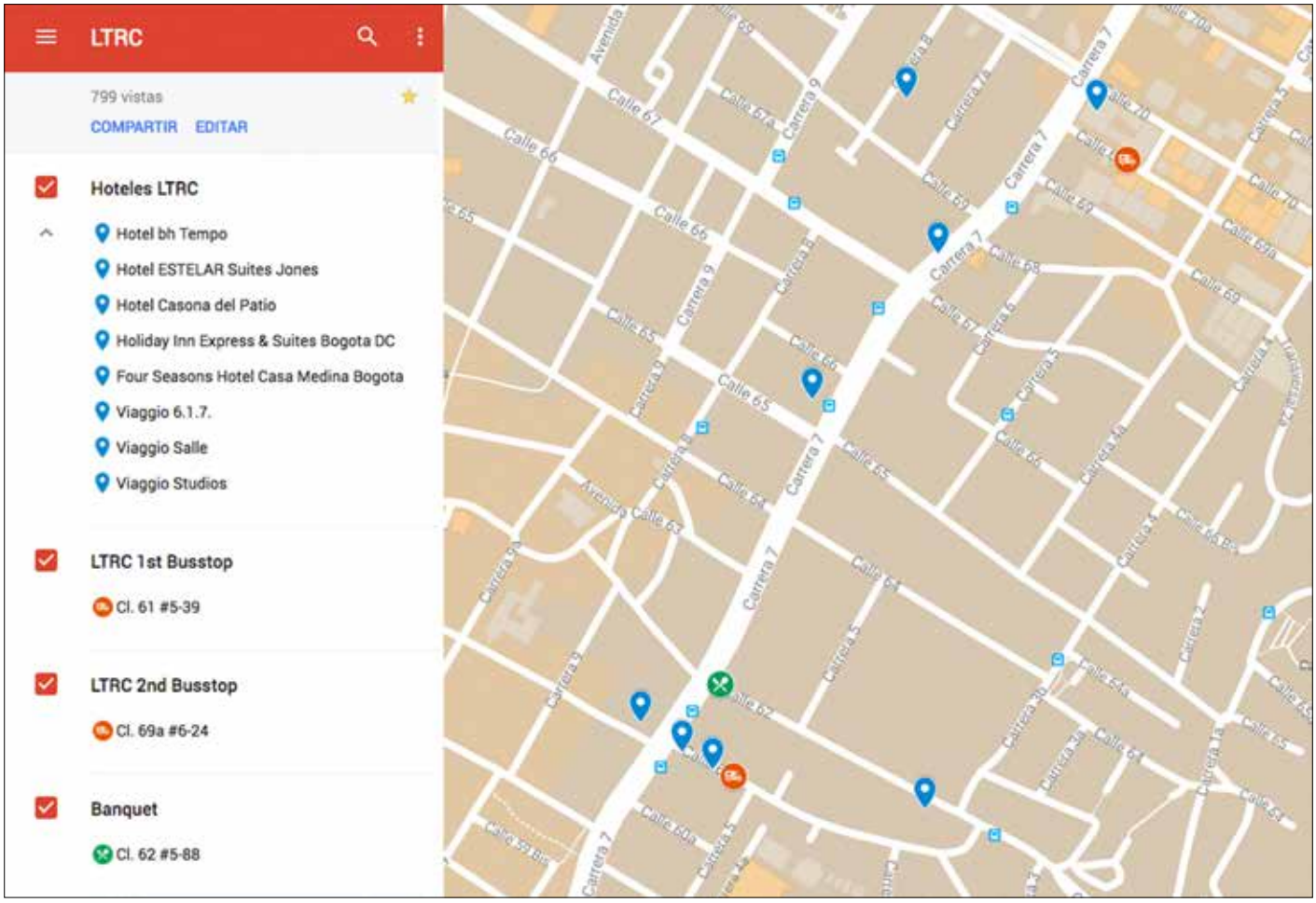
It's important in Bogotá to take a registered taxi, an Uber car, or the LTRC shuttle buses from your hotel to the conference site.

LTRC Shuttle Bus, sponsored by Paragon Testing Enterprises

Bus stop locations

STOP 1 →Hotel Estelar Suites Jones – 20 meters mountainside of Calle 61 and Carrera 7

STOP 2 →Four Seasons Hotel's restaurant, Castanyoles – 20 meters mountainside of Calle 69a and Carrera 7



LTRC Shuttle Bus to Los Andes, sponsored by Paragon Testing Enterprises

Day	Time	Pick up location
Monday (**workshop attendees only)	7:45am** (STOP 1) 8:00am** (STOP 2)	STOP 1 → Hotel Estelar Suites Jones – 20 meters mountainside of Calle 61 and Carrera 7 STOP 2 → Four Seasons Hotel's restaurant, Castanyoles – 20 meters mountainside of Calle 69a and Carrera 7
Tuesday (**workshop attendees only)	7:30am** (STOP 1) 7:45am** (STOP 2)	STOP 1 → Hotel Estelar Suites Jones – 20 meters mountainside of Calle 61 and Carrera 7 STOP 2 →Four Seasons Hotel's restaurant, Castanyoles – 20 meters mountainside of Calle 69a and Carrera 7
Tuesday (everyone)	3:00pm (STOP 1) 3:15pm (STOP 2)	STOP 1 → Hotel Estelar Suites Jones – 20 meters mountainside of Calle 61 and Carrera 7 STOP 2 →Four Seasons Hotel's restaurant, Castanyoles – 20 meters mountainside of Calle 69a and Carrera 7
Wednesday (everyone)	7:30am (STOP 1) 7:45am (STOP 2)	STOP 1 → Hotel Estelar Suites Jones – 20 meters mountainside of Calle 61 and Carrera 7 STOP 2 →Four Seasons Hotel's restaurant, Castanyoles – 20 meters mountainside of Calle 69a and Carrera 7
Thursday (everyone)	7:30am (STOP 1) 7:45am (STOP 2)	STOP 1 → Hotel Estelar Suites Jones – 20 meters mountainside of Calle 61 and Carrera 7 STOP 2 →Four Seasons Hotel's restaurant, Castanyoles – 20 meters mountainside of Calle 69a and Carrera 7
Thursday (banquet attendees)	Banquet starts at 7-7:30pm	Walk (800 meters, 10 mins.) or arrange a taxi from your hotel to the venue: Club de Comercio, Calle 62 # 5–88, on the corner of Carrera Septima (7) and Calle 62
Friday (everyone)	7:30am (STOP 1) 7:45am (STOP 2)	STOP 1 → Hotel Estelar Suites Jones – 20 meters mountainside of Calle 61 and Carrera 7 STOP 2 →Four Seasons Hotel's restaurant, Castanyoles – 20 meters mountainside of Calle 69a and Carrera 7

LTRC Shuttle Bus to Hotel Area (Rosales), sponsored by Paragon Testing Enterprises

Day	Time	Pick up location
Monday (****workshop attendees only**)	4:15pm	In front of ML building (main workshop building)
Tuesday (everyone)	7:40pm 7:45pm	In front of ML building (main workshop building)
Wednesday (everyone)	5:30pm 5:40am	In front of ML building (main workshop building) to hotel ** Montserrate Tour leaves from the front of the ML building, facing the park; returns you to the hotel district after
Thursday (everyone)	5:10pm 5:15pm	In front of ML building (main workshop building)
Thursday (**banquet attendees**)	After 9:15pm	Shuttle service to hotel area every 20 minutes starting at 9:15pm
Friday (everyone)	5:15pm 5:30pm	In front of ML building (main workshop building)

Restaurants, cafés and bars around the recommended areas

	Name	Description	Address / Phone Number	Vegan Friendly?
CHAPINERO AREA	Salvo Patria	Homemade Colombian style cuisine, ingredients from local producers.	Calle 54A N°. 4-13, Phone: +57 1 702 6367	Yes
	Mini-mal	Homemade Colombian style and creative cuisine, ingredients from local producers.	Carrera 4A N°. 57-52. Phone: +57 1 347 5464 Also at Museo Nacional's restaurant	Yes
	El Cebollero	Homemade hot dogs with ingredients from local producers.	Calle 58 No 3a-17 Phone: +57 1 8043767	Yes
	La Castaña	Homemade Latin-American cuisine	Calle 57 # 5 – 17 Phone: +57 3055792	Yes

ROSALES AREA	Arte-Sano Crepes y Waffles	Colombia's most famous restaurant for crepes and waffles. Best ice cream in the country.	Carrera 5 # 70A-08 Phone: +57 1 3463622	Yes
	El Corral Gourmet	More than 30 types of burgers. Colombian style burgers are a must!	Cl. 69a #5 09 Phone: +57 1 2171123	No
	Hamburguesa Sierra Nevada	Homemade burgers and milkshakes.	Carrera 5 # 71-12	Yes
	El Ciervo y El Oso	Homemade Colombian style and creative cuisine, ingredients from local producers.	Carrera 10a 69a – 16 +57 322 2669792	Yes
	De Raiz	Homemade Colombian style and creative vegan cuisine, ingredients from local producers.	Calle 65 # 5-70 Phone: +57 1 7568504	Yes
	WOK	South-East Asia and Japanese cuisine with a Colombian touch, one of Bogotá's all-time favorites.	Carrera 9 # 69A-63 Phone: +57 1 2120167	Yes
	Bogota Beer Company	Bogota's first and most famous local brewery, also serves pizzas and snacks.	Carrera 5 #71a-75	Yes
	Amor Perfecto	Premium Colombian coffee and pastries	Carrera 4 # 66 - 46 Phone: +57 1 2486955	Yes
	Juan Valdez Café	Premium Colombian coffee and pastries, salads and sandwiches	Calle 70 # 6-09 But also all over Colombia	Yes
"ZONA T" AREA	Restaurante Club Colombia	Typical food from Colombia, reinvented by one of Colombia's best chefs: Harry Sasson.	Carrera 9 and Avenida 82 (house in the corner) Phone: +57 1 2495681	No
	Gaira Café y Cumbia House	Typical food from the north coast of Colombia. Bar and nightclub.	Carrera 13 N°. 96 - 11. Phone: +57 1 746 2696.	No
	La Plaza de Andrés	Typical food from Colombia. Bar and nightclub.	Calle 82 #12-21 Inside "El Retiro Centro Comercial"	No

LTRC 2017 CONFERENCE OVERVIEW

LTRC 2017 CONFERENCE OVERVIEW

Pre-conference Day 1, Monday, July 17th

Time	Event	Location	
8:00–5:00	Conference Registration	Mario Laserna—2nd floor—"Calle del Saber"	
9:00–4:00	Outreach Workshops on campus with language teachers (various languages)	French: Mario Laserna—ML—513 German: Mario Laserna—ML—514	Japanese: Mario Laserna—ML—509 Portuguese: Mario Laserna—ML—515 Spanish as a Foreign Language: Mario Laserna—ML—511
9:00–4:00	Workshop 1 – Ikkyu Choi	Lleras—LL—204	
9:00–4:00	Workshop 2 – Sara Cushing	Mario Laserna—ML—512	
9:00–4:00	Workshop 3 – Erik Voss	Mario Laserna—ML—516	
10:30–11:00	Coffee break		
12:30–1:30	Lunch	Independent	

Pre-conference Day 2, Tuesday, July 18th

Time	Event	Location	
8:00–5:00	Conference Registration	Mario Laserna—2nd floor—"Calle del Saber"	
8:30–3:30	Outreach Workshops on campus with language teachers (various languages)	English: Mario Laserna—ML—510 French: Mario Laserna—ML—513 German: Mario Laserna—ML—514	Japanese: Mario Laserna—ML—509 Portuguese: Mario Laserna—ML—515 Spanish as a Foreign Language: Mario Laserna—ML—511
8:30–3:30	Workshop 1 - Ikkyu Choi	Lleras—LL—204	
8:30–3:30	Workshop 2 - Sara Cushing	Mario Laserna—ML—512	
8:30–3:30	Workshop 3 – Erik Voss	Mario Laserna—ML—516	
10:30–11:00	Coffee break		
12:00–3:30	ILTA Executive Board Meeting	Lleras—LL—001, "Hemiciclo"	
12:30–1:30	Lunch	Independent	
3:30–4:30	LTRC Newcomers' Session	Mario Laserna—ML—614	
4:40–5:00	LTRC Opening: Welcome Remarks	Mario Laserna—ML—240, (Auditorium B)	
5:00–7:00	Opening Symposium ***Towards an ILTA policy on language assessment literacy: How to define the construct and for whom?*** Elder, Kremmel, Eberharter, Harding, Tsagari, Malone, McNamara, Chalhoub-Deville	Mario Laserna—ML—240, (Auditorium B)	
7:00–8:30	Opening Reception	Mario Laserna—2nd floor—"Calle del Saber"	

Day 1, Wednesday, July 19th (a.m.)

Time	Event		
7:45–5:00	Conference Registration Calle del Saber		
8:00–5:30	Exhibits Mario Laserna—ML—5th floor		
8:30–8:45	Welcome to LTRC – Mario Laserna—ML—240, Auditorium B		
8:45–9:45	Opening Plenary: Language and Assessment in the Colombian Context Ana María Velásquez & María Lucía Casas Mario Laserna—ML—240, Auditorium B		
9:45–10:15	Coffee Break Mario Laserna—ML—5th floor		
	Paper Sessions		
	Room: ML—604	**Room: ML—606**	**Room: ML—615**
10:15–10:45	**Zhang & Yan:** Effects of Rating Method on Oral English Proficiency Assessment: An Investigation of Rater Orientations	**VanWagoner, Krauel, Chase, Cox, & Hart:** An Investigation into the Lexical-Grammatical Item Type	**Wang & Yan:** Working towards professional standards for EFL test developers in China: An investigation into stakeholder perceptions of language testing practice
10:50–11:20	**Brooks & Saleh:** Authentic Arabic speaking testing: the evolution of a test to meet policy and construct best practices	**De Jong & Buckland:** Developing a grammar scale linked to functional communication needs	**Kremmel, Eberharter, Holzknecht, & Konrad:** Enhancing language assessment literacy through teacher involvement in high-stakes test development
11:25–11:55	**Hsieh:** Impact of language requirement policy: Students' perceptions about the use of TOEIC as an exit test	**Wei, Montee, Bitterman, & Norton**: Developing a writing rubric using a hybrid approach	

Day 1, Wednesday, July 19th (p.m.)

Time	Event		
11:55–1:30	Lunch – Mario Laserna—ML—5th floor LAQ Editorial Board Meeting Mario Laserna—ML—515		
	Paper Sessions		
	Room: ML—604	**Room: ML—606**	**Room: ML—615**
1:30–2:00	**Farris:** Seeking the construct of interactional competence in aviation: A meta-synthesis	**Neumann, Padden, & McDonough:** Beyond English proficiency scores: Understanding academic performance of internationals students in the first year at an English-medium university	**Kim, Chapman, Wilmes, Cranley, & Boals:** Validation research of preschool language assessment for dual language learners: Collaboration between educators and test developers
2:05–2:35	**Arias, Baker, & Hope:** Skimming, scanning, search reading and reading comprehension: An exploration of constructs through exploratory and confirmatory factor analyses		**Hauck:** Classroom educators and the design of K-12 ELP assessments: Impact on stakeholder language assessment literacy
2:40–3:00	Group Photo Location: Plazoleta R		
3:00–3:20	Coffee Break Location: Tx Building 1st floor		
	Symposia		
	Room: R 209	**Room: R 210**	
3:20–5:20	**Jin, De Jong, Gu, Yao, Benigno, Zhu, Jin, Li, Fan, Wang, & Davis:** Human-machine teaming up for language assessment: The need for extending the scope of assessment literacy	**Cheng, Shohamy, Tsagari, Saif, & Wang:** Test preparation: A double-edged sword	

6:00–9:00 p.m. Informal Outing: Cable car up to Montserrate.
Leaves from the main entrance to Mario Laserna (ML).

Day 2, Thursday, July 20th (a.m.)

Time	Event		
8:00–5:00	Conference Registration (Mario Laserna—2nd floor—"Calle del Saber")		
8:00–5:00	Exhibits (Mario Laserna—ML—5th floor)		
	Paper Sessions		
	Room: ML—604	**Room: ML—606**	**Room: ML—617**
8:30–9:00	**Oh:** Investigating second language learners' use of linguistic tools and its effect in writing assessment	**Owen:** Comparing the cognitive processes in the reading sections of IELTS and TOEFL	**Kremmel & Harding:** Towards a comprehensive, empirical model of language assessment literacy across different contexts
9:05–9:35	**Janatifar, Marandi, & Babaei:** Defining EFL teachers' language assessment literacy and its promotion through virtual learning teams	**Baumann, Dursun, Swinehart, & McCormick:** Embarking on developing a meaningful measure of graduate-level L2 reading comprehension: The role of primary stakeholders' understanding of the construct	**Wolf, Wang, Oh, & Tsutagawa:** Investigating the types and uses of feedback for middle-school English language learners' academic writing performance
9:40–10:10	**Berry, Nakatsuhara, Inoue, Galaczi, & Patel:** Video-conferencing and face-to-face delivery of a speaking test: Implications for different assessment stakeholders	**López, Schmidgall, Blood, & Wain:** Using Assessment information: How young English learners interpret assessment information on score reports	**Froetscher:** Washback on classroom testing: assessment literacy as a mediating factor
10:10–10:40	Coffee Break (Mario Laserna—ML—5th floor)		

	Paper Sessions		
	Room: ML—604	**Room: ML—606**	**Room: ML—617**
10:40–11:10	**Cooke, Barnett, & Rossi:** An evidence-based approach to generating the Language Assessment Literacy profiles of diverse stakeholder groups	**Jang, Strachan, Sinclair, Larson, & Gallo:** Influence of contextual specificity on interaction between language ability and specific-purpose content knowledge: The case of OELPE ESP Test	**Valeo & Barkaoui:** How teachers' conceptions mediate their L2 writing assessment practices: Case studies of ESL teachers across three contexts
11:15–11:45	**Harsch, Seyferth, & Brandt:** Developing assessment literacy in a dynamic collaborative project: what teachers, assessment coordinators, and assessment researchers can learn from and with each other	**Knoch, Elder, Huisman, Flynn, Manias, Mcnamara, & Woodward-Kron**: How to move from Indigenous criteria to a usable scale? Evolving guidelines for ESP test developers	**Barkaoui & Valeo:** Designing L2 writing assessment tasks for the ESL classroom: Teachers' conceptions and practices
11:50–12:20	**DEMO: Tsuprun, Evanini, Timpe-Laughlin, & Blood:** Automated task-based feedback in a formative assessment of workplace English speaking skills using a spoken dialog system		**Wagner & Jang:** Assessment competencies for effective feedback: Psychological and environmental factors influencing language assessment
12:20–2:00	Lunch (Mario Laserna—ML—5th floor) ILTA Business Meeting (Mario Laserna—ML—608)		

Day 2, Thursday, July 20th (p.m.)

	Works–in-Progress (WIPs)	
	Mario Laserna (ML)—2nd floor—"Calle del Saber"	
2:00–3:30	**1. Llosa & Grapin:** Exploring the possibility of using integrated assessments of science and language **2. Benigno:** Assessing young learners of English: principles and challenges **3. Guerra & Rosado:** Towards a text-based framework for academic Spanish reading and writing assessment **4. González:** Writing assessment training and Mexican EFL University teachers: a study of training impact. **5. Restrepo & Jaramillo:** Preservice teachers' language assessment literacy development **6. Eberharter:** Judgment and decision-making in rating speaking: Exploring the role of cognitive attributes **7. Quevedo-Camargo & Scaramucci**: The assessment literacy of Brazilian language teachers **8. Wang & Zhang:** China's Standards of English: The uses of English speaking activities in China **9. Son:** Measuring heritage language learners' proficiency: Validating the Korean C-test for research purposes **10. Im:** Stakeholders' voices: Alignment or discrepancy between proposed and perceived inferences based on English test scores in international business contexts **11. Reed, Van Gorp, Ohlrogge, Kim, Isbell & Fox:** Towards a principled approach to the design of rater training and norming protocols	**12. Chen:** A corpus-driven receptive test of collocational knowledge **13. Ribeiro e Souza:** The washback effect of EPLIS on teachers' and learners' perceptions and actions **14. Dunaway:** Pairs in written assessments **15. Sultana**: In VIVO Vs. in VITRO: The journey of a secondary English public examination in Bangladesh **16. Hernández Ocampo:** How literate in language assessment should English teachers be? **17. Chen, Banerjee, & Li:** Reducing the number of options in multiple choice items: Does the option reduction method matter? **18. Beltrán:** A comparison of scenario-based and traditional achievement tests **19. Fernandes Yamamoto & Consolo:** Lexical competence in Japanese and the definition of a test construct for teachers of Japanese as a foreign language **20. de Andrade Raymundo:** The on-going validation process on the Brazilian English Proficiency Exam for air traffic controllers **21. Seong:** Examining the cognitive dimension of academic speaking ability through a scenario-based speaking test
3:30–4:00	Coffee Break (Mario Laserna (ML)—2nd floor—"Calle del Saber")	
4:00–5:00	Messick Award Lecture *Stephen G. Sireci (*Mario Laserna—ML—240, Auditorium B)	
7:30–10:00	Banquet and Awards Presentation (Club de Comercio, Calle 61 @ Carrera 7)	

Day 3, Friday, July 21st (a.m.)

8:00–12:00	Conference Registration (Mario Laserna—2nd floor—"Calle del Saber")	
8:00–12:00	Exhibits (Mario Laserna—ML—5th floor)	
	Symposia	
	Room: ML—240, Auditorium B	**Room: ML—346, Auditorium C, "upstairs"**
8:30–10:30	**Schissel, Chalhoub-Deville, Shohamy, Leung, Lopez-Gopar, Limerick, Saville:** The construct of multilingualism and language policies in language testing	**Read, Ginther, Dimova, Swerts Knudsen, Osorio, & Weideman:** Assessing the academic literacy of university students through post-admission assessments
10:30–10:45	Coffee Break (Mario Laserna—2nd floor—"Calle del Saber")	
	Poster Sessions	

	Location: Mario Laserna (ML)—2nd floor—"Calle del Saber"	
10:45–12:00	**1. Tsagari, Vogt, Csepes, Green, & Sifakis:** Our TALE: Developing Online Teacher Assessment Literacy Training Materials **2. Kim, Chapman, & Wilmes:** Developing materials to enhance the assessment literacy of parents and educators of K-12 English language learners **3. Castro, Gottlieb, & Chapman:** Measuring language proficiency in K-12 US bilingual programs: Contesting the monolingual narrative of bilingual learners **4. Baten, Van Maele, Moreno, Dàvila, & van Splunder:** Stakeholder views on the assessment and certification of English language proficiency in global interuniversity collaboration **5. Consolo & Aguena:** An electronic pre-test for the EPPLE examination: test design, preliminary results, technical challenges and mobile technology **6. Okabe, Hama, & Umakoshi:** Introduction of four-skilled college entrance exams: English teachers' perceptions and needs for assessment literacy among teachers **7. Villa-Larenas:** Language assessment literacy of EFL teacher trainers **8. Arias:** Re-evaluating commonly held views of residual-based fit statistics in language assessment research: Rasch analysis of the CanTEST listening test **9. Rock:** Establishing construct validity in the development of an oral placement test at an intensive English program	**10. Davis, Yoon, Loukina, & Zechner:** Automated scoring of speaking – Lessons from a real-world assessment **11. Tsagari:** Why listening to the voices of stakeholders matters in defining LAL **12. Harrison & Walczak:** The road to success: Advising English language education policy in Chile – Findings and recommendations from research on a national benchmarking test **13. Cox & Thompson:** The effects of timed and untimed testing conditions on the item difficulty parameters of a Spanish reading proficiency test **14. Cushing, Armistead, & Chiesa:** Post-enrollment English language assessment and support for international graduate students: Is it worth the effort? **15. Bustos González:** English language assessment in Colombia: A teachers' perspective **16. Suvorov, Cárdenas-Claros, & Rick:** Using an argument-based framework for language program evaluation: A case study **17. Farhady & Tavassoli:** Developing a professional knowledge of language assessment test for EFL teachers **18. Kolesova:** Assessment of ESL sociopragmatics for informing instruction in an academic context: From Australia to Canada
12:00–1:30	Lunch (Mario Laserna—ML—5th floor) Language Assessment in Latin America meeting (Lleras—LL—001, "Hemiciclo") LT Editorial Board Meeting (Mario Laserna—ML—516)	

Day 3, Friday, July 21st (p.m.)

	Paper Sessions		
	Room: ML—604	**Room: ML—606**	**Room: ML—607**
1:30–2:00	**O'Connell & Pearce:** Defining the "just qualified" speaker: Comparisons from two CEFR linking panels	**Berry, Sheehan, & Munro:** Mind the gap: Bringing teachers into the language literacy debate	**Nakamura**: National college entrance examination reform and the effort to raise assessment literacy of stakeholders in Japan
2:05–2:35	**Thompson, Cox, & Brown:** Understanding temporal fluency: The influence of the ACTFL OPIc speaking prompts on L2 speakers of Spanish	**Jin & Jie:** Do workshops really work? – Evaluating the effectiveness of training in language assessment literacy	**Cardenas-Claros, Suvorov, & Rick:** A language program evaluation using an argument-based approach: a case study in Chile
2:40–3:10	**Hill & Ducasse:** Contextual variables in written assessment feedback in a university level Spanish program	**Yan, Fan, & Zhang:** Understanding language assessment literacy profiles of different stakeholder groups in China: The importance of contextual and experiential factors	**Finardi, Amorim, & Kawachi-Furlan:** Internationalization and language testing in Brazil: Exploring the interface of TOEFL ITP and rankings at UFES
3:10–3:40	Break (ML—5th floor)		
3:40–4:40	Davies Lecture *Ofra Inbar-Lourie* (ML—240, Auditorium B)		
4:40–5:00	Closing (ML—240, Auditorium B)		

ABSTRACTS

PRE-CONFERENCE WORKSHOPS

Workshop 1: Data and Model Visualizations for Effective Communication of Statistical Results

Monday, July 17, 9:00–4:00, Lleras—LL—204; Tuesday, July 18, 8:30–3:30, Lleras—LL—204

Ikkyu Choi, Educational Testing Service, USA

Statistics is about using data to make a best guess, and to quantify how bad that guess is (Rao, 1997). Despite these intuitive goals, the communication of statistical results has been challenging, even within academic communities. This communication challenge is particularly pertinent to language assessment professionals who frequently need to communicate about statistical concepts and results with peers, students, and/or assessment stakeholders. Effective visualizations of data and statistical models can facilitate such communications by replacing technical terminology with straightforward visual representations.

The goal of this workshop is to provide participants with an understanding of principles for effective data and model visualizations, and hands-on experience in creating them. All visualizations in this workshop will be demonstrated and created in real-time using R, a free statistical program. The corresponding R code will be provided to participants in an editable format for their own projects.

Participants should bring their own laptops (with any major operating system) with the most current version of R installed. Required packages will be installed as part of the workshop. Previous experience with R will not be required.

Part 1 (Day 1, Morning): The first part of this workshop will provide theoretical and practical background for constructing effective data visualizations. We will begin with contrasting examples of data visualization that facilitate or hinder the communication of intended messages, and derive a set of principles for effective statistical graphics. We will then learn the basics of R and its graphics engine, and use them to create "facilitating" data visualization examples.

Part 2 (Day 1, Afternoon): The second part of this workshop will focus on refining data visualizations for presentations and publications. This part will be organized as a series of activities, each of which will attempt to recreate published data visualization examples. Through the activities, we will learn how to customize data representation, to add labels and legends, and to emphasize focal points. We will also cover a number of export options, and discuss preferred options depending on the operating system and the target outlet.

Part 3 (Day 2, Morning): In the third part of this workshop, we will focus on the visualization of statistical models. We will learn how to visualize many popular statistical models among

language assessment professionals, such as linear regression models, multilevel models, and factor analysis/item response theory models, which are presented as straight lines or curves on top of data. We will compare the resulting visualizations and the standard outputs from the models, and discuss how effective visualizations of models can show what the models actually do, and how well they fit the data. We will also discuss how to represent the uncertainty of the models to signify the quality of information contained in the graphics.

Part 4 (Day 2, Afternoon): The last part of this workshop will introduce interactive and dynamic visualizations. We will learn how to create interactive graphs that take user input and return visualized results, and dynamic representations of time series and/or repeated events. We will then see how popular statistical and measurement concepts such as p-values, power, and reliability, can be communicated with interactive and/or dynamic visualizations in an intuitive and straightforward manner.

Reference: Rao, C. R. (1997). Statistics and truth: Putting chance to work (2nd ed.). River Edge, NJ: World Scientific.

Ikkyu Choi is an associate research scientist in the Research and Development division at Educational Testing Service. He earned his PhD from the University of California, Los Angeles in 2013, with a specialty in language assessment. His research interests include second language development profiles, test taking processes, and quantitative research methods for language testing data.

Workshop 2: Developing and using rating scales for writing assessment

Monday, July 17, 9:00–4:00, Mario Laserna—ML—512; Tuesday, July 18, 8:30–3:30, Mario Laserna—ML—512

Sara T. Cushing, Georgia State University, USA

In this hands-on workshop we will cover fundamental considerations for developing, using and validating rating scales for writing assessment, both in classroom settings and for large-scale tests. In the first part of the workshop we will discuss the role of the rating scale in writing assessment in terms of scoring validity (Shaw & Weir, 2007) and discuss the advantages and disadvantages of different types of rating scales (i.e., holistic vs. analytic scales). We will generate our own criteria for scoring writing for a specific assessment purpose with a specific group of learners and we will experience different methods of rating scale construction, including the exploration of online resources for scale development. We will also use different published rating scales to evaluate second language writing samples and discuss the benefits and challenges of using different types of scales. We will discuss best practices in training, monitoring, and providing feedback to raters in large-scale assessments and we will discuss how rating scales can be used in classrooms to both lessen teachers' marking load, provide useful feedback to students, and make use of self-assessment and peer assessment. Depending on participant interests and experience, we will cover the following additional topics: aligning scales to standards such as the CEFR, investigating and reducing various types of rater bias, and interpreting and using published quantitative and qualitative research on rating scales to inform our own research and practice.

Reference: Shaw, S. D., & Weir, C. J. (2007). Examining writing: Research and practice in assessing second language writing (Vol. 26). Cambridge University Press.

Sara T. Cushing is Professor of Applied Linguistics and Senior Faculty Associate for the Assessment of Student Learning at Georgia State University. She received her Ph.D. in Applied Linguistics from UCLA. She has published research in the areas of assessment, second language writing, and teacher education, and is the author of Assessing Writing (2002, Cambridge University Press). She has been invited to speak and conduct workshops on second language writing assessment throughout the world, most recently in Norway, the United Kingdom, South Korea, and Thailand. Her current research focuses on assessing integrated skills and the use of automated scoring for second language writing.

Workshop 3: Current and Emerging Applications of Information Technologies for Language Assessment

Monday, July 17, 9:00–4:00, Mario Laserna—ML—516; Tuesday, July 18, 8:30–3:30, Mario Laserna—ML—516

Erik Voss, Northeastern University, USA

The role of technology in language learning and assessment is expanding rapidly. Computers are used to develop, deliver, and automatically score language tests. Teachers, test developers, and language testing researchers can benefit professionally from learning about potential capabilities and challenges of assessing language through computer technology (Chapelle & Douglas, 2006). This workshop will explore information technologies that make it possible to assess language learners in technology-enhanced language learning environments. We will focus on how current and developing computer technologies limit or contribute to assessment task design and score interpretation.

Part I: We will begin by examining information technologies in use today such as multimedia, social media, audio and video hardware, video conferencing software, and mobile computing that are currently used for assessing language in online education and classroom-based programs. Discussion topics will include human-computer interaction, digital literacy and anxiety, game-based approaches, social media, and natural language processing. Hands-on activities will contribute to the discussion of how different language technologies may affect language performance and how technology can alter the construct definition or ability being measured.

Part II: Building on the discussions and activities from part one, part two will include examples of sophisticated and emerging technologies (e.g., speech recognition, facial recognition, and virtual reality) that could advance current language assessment frameworks, task types and forms of human-computer interaction. Consideration will be given to the use of commercial technologies and data analytics for expanding knowledge of testing concepts and future directions (especially in terms of artificial intelligence and augmented reality). This part will also address criteria, principles, and argument-based validation approaches for evaluating assessment using technology.

Reference: Chapelle, C. A. & Douglas, D. (2006). Assessing language through computer technology. Cambridge: Cambridge University Press.

Erik Voss is an Associate Teaching Professor in NU Global at Northeastern University. His research interests include language testing and technology, and validation research. Dr. Voss

has worked at Iowa State University as assessment coordinator for the Intensive English and Orientation Program and as a SPEAK/TEACH rater and instructor in the Speaking Skills for International Teaching Assistants. He is currently the webmaster for the International Language Testing Association (ILTA) and has served as secretary of the Midwest Association of Language Testers (MwALT).

Los Andes Language Assessment Outreach Workshop Leaders

English (sponsored by CaMLA)

Gad Lim

Gad leads research on IELTS at Cambridge English. He has taught in higher education and trained language teachers in Australasia, North America, and Europe, and has worked as examiner, item writer, test designer, and test development manager. He has presented and published on language teaching, EAP, performance assessment, test validation, standard setting, and on the CEFR. His PhD is from the University of Michigan.

Stephen O'Connell

Stephen is an assessment manager at CAMLA. He has taught EFL in South Korea, ESL in New York, and has worked as a test developer of high-stakes standardized tests for more than ten years. Stephen holds an MA in applied linguistics and is a PhD candidate in the second language acquisition program at the University of Maryland, College Park.

French / français

Eliane Lousada

Eliane is Professor-Researcher at the University of Sao Paulo (USP), Brazil. She is specialist in Didactics of French as a Foreign and Second Language. She worked for fifteen years at the Alliance Française de Sao Paulo, where she participated in the application of language certifications such as DELF / DALF, TCF and TEF and where she, as a teaching coordinator, was in charge of teaching EFL didactics and assessment. After completing her doctorate in 2006, she was a professor at the Bachelor's and Master's level in French Studies at the University of Guelph in Canada. She currently teaches and conducts masters and doctoral research at the Department of Modern Letters at the USP. She has published numerous articles in the field of Didactics of French as a foreign language, especially on the subject of teacher education.

German / Deutsch (co-sponsored by Goethe-Institut)

Heike Neumann

Heike Neumann has a PhD in Second Language Education, specializing in language assessment, from McGill University in Montreal, Quebec, Canada, and has been at Concordia University in Montreal as head instructor for English as a second language and Test Content Developer for the Concordia Comprehensive English Placement Test (ConCEPT) since 2010. She has worked with German over the years as a teacher for adults at Concordia University and the Goethe Institut in Montreal and for high school students at the German Language Schools Quebec. As part of her work with the German Language Diploma classes, she also worked as a rater for the oral exams of the German Language Diplomas I and II (Deutsches Sprachdiplom der Kultusministerkonferenz, Stufe I und II). Her research interests include language assessment, second language writing, English for academic purposes, and collaborative learning. Her research has been published in the Canadian Modern Language Review, Journal of Second Language Writing, Journal of English for Academic Purposes, TESL Canada Journal, TESOL Quarterly, and Writing & Pedagogy.

Japanese / 日本語

Yasu-Hiko Tohsaku

Ph.D. in Linguistics from University of California, San Diego. Currently, Professor at School of Global Policy and Strategy, University of California, San Diego. Director of the School's Foreign Language Program and Undergraduate Japanese Language Program. Research interests are second language acquisition, foreign language pedagogy, and language testing. Publications include "Yookoso!: Invitation to Contemporary Japanese", "Yookoso!: Continuing with Contemporary Japanese" (both McGrawHill), "Doraemon no Dokodemo Nihongo" (Shogakukan), "Nippon 3.0 no Shohosen" (Kodansha), and so forth. Published numerous papers related to foreign language teaching and testing, etc. Participated in the development of the National Standards for Learning Foreign Languages in the United States. Presently, the Japanese language representative to the Executive Board of the National Standards Collaborative Project. Former Board Member of the Joint National Committee for Languages, the lobbyist group for foreign language education in the United States. Former President of Association of Teachers of Japanese and American Association of Teachers of Japanese. Currently, President of Computer-Assisted System of Teaching and Learning Japanese (CASTEL/J). Project Director, Japanese Global Articulation Project (J-GAP)

Portuguese / Português

Matilde Scaramucci

Matilde holds a degree in Portuguese / English Literature from the University of Vale do Paraíba (1974); Master's Degree in Linguistics (TESOL) from San Jose State University (1980), California; PhD in Linguistics from the State University of Campinas (1995); as well as postdoctoral studies from the University of Melbourne, Australia (2008). She is currently a professor at the Department of Applied Linguistics at the State University of Campinas. She was the Director of the Institute of Language Studies from 2010 to 2014 and editor of the Journal of Applied Linguistics between 2006 and 2014. She is a member of the Technical Technical Committee of the Certificate of Proficiency in Portuguese for Foreigners (Celpe-Bras) (1993-2006) and (2012-present), having been part of the team who developed this exam. Her main area of research is assessment in contexts of teaching / learning of languages, with numerous publications and orientations in this area. She has a long history of research on the retroactive effects of foreign language examinations in language teaching. She participated actively in the conceptual development of the Certificate of Spanish Language and Use (CELU) in Argentina and helped in establishing benchmarks between this exam and the Celpe-Bras.

Spanish as a foreign language / Español como lengua extranjera (ELE)

Claudia Forero González

Claudia is a Colombian teacher with more than 20 years of experience teaching Spanish as a foreign language (ELE). She holds a Ph.D. in Linguistics applied to the teaching of Spanish as a Foreign Language from the Antonio de Nebrija University. From the same University she earned a Diploma of Advanced Studies in the teaching of ELE. Her areas of interest include didactics, methodology, curriculum and materials design, as well as assessment. Her cum laude doctoral dissertation was focused on assessment and the development of a mechanism that is aimed at auditing ELE tests. This instrument was also implemented in the construction of the English placement tests for admission to the undergraduate and postgraduate candidates of the National University of Colombia between 2011 and 2016. Over 20,000 students were ranked nationwide.

Effects of rating method on oral English proficiency assessment: An investigation of rater orientations

10:15–10:45 · Mario Laserna—ML—604

Xiaoyi Zhang, Shanghai Jiao Tong University, China

Jin Yan, Shanghai Jiao Tong University, China

With the recognition of construct-context interaction in language testing (e.g., Bachman, 2007; Chalhoub-Deville, 2003; Chapelle, 1998; Fulcher, 2015), importance has been attached to rating method for its function in mediating the task and rater effect in performance assessment (Schoonen, 2005). Meanwhile, the heightened awareness of the necessity to relate rating scales to test construct attracts further attention to rating method. As argued by McNamara, Hill and May (2002), the ways in which rating scales and criteria are interpreted by raters represent the de-facto test construct of performance assessment. In this sense, rating method, an essential aspect featuring rating scales (Fulcher, 2003), also plays a part in defining test construct. Given a few exceptions (e.g., Barkaoui, 2008; Schoonen, 2005), however, research on how raters react to and perceive interactions among rating method, context of performance, and test construct appears scarce, although discussions over scale utility and validity of different rating methods can be abundant (e.g., Fulcher, 2003; Weigle, 2002).

The present study, therefore, compared how raters performed in speaking performance rating using two rating methods of different degrees of interaction with the construct to be measured. The comparisons were made regarding the effects of rating method on 1) score outcomes, 2) components of score variance, and 3) raters' judgmental behaviors. The context of the study was the rating of a national oral English proficiency test administered to tertiary-level learners of English in China. With a counter-balanced partially-crossed design covering two rating rounds separated by one week, twenty-six raters used the same set of rating criteria and scales on "speech intelligibility", "language use", "content elaboration", "discourse management", and "communication appropriacy", to score 100 test-takers' performances on the entire test (criterion-based scoring) and on individual tasks (task-based scoring), which include Reading aloud, Picture description and Pair discussion. Semi-structured interviews were conducted with 10 raters to elicit their perceptions of the rating methods.

Correlation analyses indicated a stronger halo effect in criterion-based scoring than in task-based scoring. Paired-sample t-tests showed criterion-based scoring resulted in significantly higher total scores than task-based scoring among higher-achievers, while the situation for lower-achievers was the opposite. Multi-faceted Rasch analyses identified better rater consistency and a wider use of scale steps in task-based scoring, while criterion-based scoring outperformed task-based scoring in discriminating test-takers with different levels of oral proficiency. Bias analyses only elicited a few cases in which component scores were significantly biased against rating methods. Semi-structured interviews revealed a unanimous preference for task-based scoring with which the raters felt more confident in giving scores. However, in task-based scoring, the raters were found to penalize test-takers on certain "core" aspects of speaking such as "speech intelligibility" across all tasks, although different "focal" evaluation criteria were

assigned to different tasks. Results of the study indicated language assessment literacy among raters could vary by their rating experience and the rating method being used. It was therefore recommended that raters be trained for a fuller understanding of different rating methods to enhance the validity of test score interpretation and use.

An investigation into the lexical-grammatical item type

10:15–10:45 · Mario Laserna—ML—606

Kaitlyn VanWagoner, Brigham Young University, USA

Mariah Krauel, Brigham Young University, USA

Braden Chase, Brigham Young University, USA

Troy Cox, Brigham Young University, USA

Judson Hart, Brigham Young University, USA

A meta-analysis of over 76 studies found that Elicited Imitation (EI) tasks demonstrate a strong ability to discriminate between speakers across proficiency levels (Yan, Maeda, Lv, & Ginther, 2016). While EI focuses on processing and automaticity with oral skills, could the same principles function with literacy skills? This study reports the use of a novel item-type called lexical-grammatical. This item type applies the same concept as EI to the reading-writing modality. The item mimics the unbroken string of phones processed aurally with EI to a written modality by stripping a sentence of all white space, capitalization, and punctuation. The examinee then reads the prompt, deconstructs it into its constituent lexical and grammatical parts and must reconstruct it by typing out the sentence. To imitate the transient nature of EI, time limits are applied to each prompt so the examinees must access their internalized language processing systems and use automaticity in responding. The item can then be scored a number of different ways ranging from correctly identifying word boundaries to inserting correct punctuation and capitalization. In a pilot exam, a 38 item test was administered to 252 ESL students as part of an online battery of tests for admittance into an online program. The items were scored dichotomously and the internal reliability was found to be quite high (Cronbach alpha = .94).

While the reliability is high, it is still unknown what aspects of the item type make it so. While it is hypothesized that vocabulary knowledge, grammatical competency, and writing conventions all contribute to some extent, typing ability might also play a role. To investigate the role of the aforementioned language skills/subskills, a lexical-grammatical test will be administered in conjunction with separate tests of typing, grammar, vocabulary, and writing mechanics. Analyses will then be run on the results to determine what the lexical-grammatical item type is assessing and to determine the extent to which this item type, like EI, could discriminate between readers/writers across proficiency levels.

Working towards professional standards for EFL test developers in China: An investigation into stakeholder perceptions of language testing practice

10:15–10:45 · Mario Laserna—ML—615

Jinsong Fan, Fudan University, China

Li Wang, Xi'an International Studies University, China

Xun Yan, University of Illinois at Urbana-Champaign, USA

Recent years have witnessed the development and implementation of a host of professional standards (e.g., codes of ethics or guidelines for practice) in the field of language testing both internationally and regionally (e.g., EALTA, 2006; ETS, 2014; ILTA, 2000; 2007; JLTA, 2007), which reflect collective efforts to raise quality and professionalism of language testing practice worldwide (Davies, 2004). In order to develop professional standards, one must first understand the status quo of testing practice in a particular context (Alderson, 2011). The present study is situated in a larger research project to investigate the current EFL testing practices in China in the hope to inform relevant educational and examinations authorities in their endeavour to develop professional standards for language test developers. Specifically, this study investigated how students and teachers, the two primary stakeholder groups, perceived the EFL testing practices in regards to test development, administration, and use.

Following three steps, we first administered an open-ended questionnaire to 248 university students and 80 teachers, to solicit their general views on good and bad English language tests. Second, semi-structured interviews (n = 15) and focus groups (n = 12) were conducted to students with a view to understanding more clearly how they perceived the current EFL testing practice in terms of test design, administration, and use. Third, semi-structured interviews were conducted to English teachers (n = 12) to understand their views on the current EFL testing practice.

All data in this study were coded and analysed in NVivo (QSR, 2012), using the grounded theory and inductive analytic approach; cluster analysis in NVivo was subsequently performed to compare students' and teachers' views. Overall, five themes emerged from the iterative coding process, including: 1) test design; 2) test impact; 3) test use; 4) test fairness; and 5) test administration. On the whole, both groups demonstrated a similar pattern in their perceptions, though teachers were found to have more positive perceptions. Both groups commented least positively on test design, believing that EFL tests in China tended not to measure communicative language use but rather relied on multiple-choice questions. In addition, both groups expressed the view that speaking ability had been somewhat under-represented. Regarding test impact and use, both groups commented that the tests had significant impact on them, and many tests had been overused or misused. In light of test fairness and administration, the transparency of test-related information, test security, and the comparability of difficulty of the same test across different administrations constituted the most important concerns.

Through investigating stakeholders' perceptions on the current EFL testing practice, this study has significant implications for the future development and reforms of EFL tests in China. Given the importance of understanding the status quo in developing professional standards (Alderson, 2011), this study provides timely reference to the relevant educational and examinations authorities in their current endeavour to develop professional standards for language test developers in China. Stakeholders' roles in language testing as well as the responsibility of language testers to inform policy will be discussed.

Authentic Arabic speaking testing: The evolution of a test to meet policy and construct best practices

10:50–11:20 · Mario Laserna—ML—604

Rachel Brooks, FBI, USA

Eleonore Saleh, FBI, USA

Since the Arabic language incorporates both the formal version (Modern Standard Arabic, MSA) and colloquial Arabic (dialect), assessment of a speaker's overall ability is often considered complex (Gebril & Taha-Thomure, 2013). Procedures for testing Arabic speaking have had to balance dialect and MSA both in government and academic contexts (ACTFL, 2012; Alosh et al., 2007; Brau, Brooks, & Saleh, 2011). At the FBI, the Arabic speaking tests were originally conducted in MSA-only, leading to artificial situations where examinees were asked to perform tasks that are unnatural in MSA. In 2007, the FBI began a policy to test Arabic dialects separately from MSA to meet the operational need for measures of proficiency in the dialect. This second era of Arabic testing also had limitations, as it was unnatural to completely separate the two. In 2011, consultation with Arabic and testing experts and stakeholders led to a third era of testing. The FBI changed to a combined test of dialect and MSA. The combined Arabic test best meets the measurement needs of policy makers and conforms to the construct of how Arabic is naturally used.

The FBI combined Arabic test has been in place for five years, prompting a reevaluation of the impact these testing policy changes have had on the FBI and examinees, including language learners and language position applicants (Fulcher, 2012). An analysis was conducted of over 400 Arabic test examinees that have retested between 2000 and 2016 in MSA and dialects. Scores from the same examinee taken in the three testing eras (MSA-only, MSA and dialect separately, and combined test) were analyzed with ANOVAs to see the impact of the changing test protocol and construct. In analysis 1, scores from era 2 were compared with those from era 1 to examine if a more comprehensive inventory of ability was acquired. Analysis 2 compares ratings from the combined tests with ratings of MSA and dialect. Analysis 3 examines whether learners' tests of the same dialect are affected across eras. Consideration was made in the analysis for examinee testing reason (end of training, applicant), acquisition method (native or non-native speaker), and proficiency level. A subset of tester reports was reviewed to analyze the justification for the scores given.

Results of the study have shown that a test that combines MSA and dialect most often results in a significantly higher rating as it allows examinees to better and more authentically demonstrate their overall proficiency in Arabic. With the current test, FBI policy makers have more accurate information to make hiring decisions and give assignments and examinees have more ways to demonstrate their Arabic speaking skills and qualify for positions, bonuses and assignments. The results have implications for Arabic testing within the government as well as in academic contexts, where issues in testing dialect versus MSA are relevant.

Developing a grammar scale linked to functional communication needs

John de Jong, VU University Amsterdam, Netherlands

Simon Buckland, Pearson, United Kingdom

10:50–11:20 · Mario Laserna—ML—606

In spite of its shortcomings signaled by researchers in language testing (e.g., Fulcher, 2004, 2010; Hulstijn, 2007) the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001) has been adopted worldwide by researchers, teachers and language policy makers. This universal adoption in spite of its regionally marked name seems to indicate a level of usefulness that is unprecedented and as of yet unsurpassed. Not all adoptions may be based on a full understanding of the framework and some adopters may add regional aspects (e.g., Tono, Y., & Negishi, M., 2012).

Notwithstanding this success there still remains substantial room for improving the effectiveness of its application to language learning teaching and assessment. The publication itself with its 260 pages is not directly accessible for people seeking to improve their learning or their methodology for teaching or assessing a language. This is not to be seen as a shortcoming of the publication, it was never intended to fulfill such a role. Users of the framework are recommended by the Council of Europe to create, for each language, inventories of lexical and grammatical elements known as "Reference Level Descriptions" (RLD; http://www.coe.int/t/dg4/linguistic/DNR_EN.asp). The RLDs are defined as "inventories of the linguistic realizations of general notions, acts of discourse and specific notions / lexical elements and morpho-syntactic elements" which are characteristic of each level (Council of Europe, 2005, p.5). In relation to this recommendation many users of the framework seem to be unaware of - or simply overlook - the fact that the CEFR directly builds onto a series of syllabus documents predating its 2001 publication from which these RLDs can be gleaned. Of these the Threshold level (Van Ek, 1975) is probably the best known. The Threshold level figures as the first 'independent' level in the CEFR, labeled 'B1'. It was later followed by Waystage (A2, 1979) Vantage (B2, 2001) and finally Breakthrough (A1, 2001). No such document exist for the C-levels. These documents spell out the linguistic elements in language-independent terms that are needed to be able to perform the functional communicative language tasks that define each of the CEFR levels. It is important to note that the lexical and grammatical elements are thus presented as ancillary to enable the language user to perform the communicative language tasks.

In this paper we present a scaling of grammatical elements for English that parallels the functional scale represented by the well-known 'Can do' statements that define the CEFR scale. The scaling is based on combining expert and teacher judgements on the usefulness of over 400 grammatical elements to perform language tasks at increasing levels of functional communicative proficiency. Reliability of judgements was assessed by evaluating agreement within and across each of the two rater groups. Finally validity was evaluated by relating the grammar scaling the functional communicative 'Can do' statements in the CEFR.

Enhancing language assessment literacy through teacher involvement in high-stakes test development

10:50–11:20 · Mario Laserna—ML—615

Benjamin Kremmel, University of Innsbruck, Austria

Kathrin Eberharter, University of Innsbruck, Austria

Franz Holzknecht, University of Innsbruck, Austria

Eva Konrad, University of Innsbruck, Austria

Involving teachers in the development of high-stakes language tests certainly holds advantages for an exam: teachers' classroom expertise can add to the validity of the test and can give credibility to an exam through a sense of ownership of a main stakeholder group. However, it also offers considerable potential for professional development and fostering language assessment literacy (LAL) among key players in the educational system. This is even more important when this involvement concerns a major exam reform, such as setting up a national standardized school-leaving exam, which offers a great opportunity for professional teacher development through item writer training, scale development sessions, rater training, and centralized marking sessions.

This presentation will share insights from training a large cohort of language teachers to become item writers for a state-wide language exam in a European country. It will first report on the teachers' roles, tasks and trainings in the test development project. It will then present the results of a study that retrospectively surveyed how being involved in the project has contributed to the participating teachers' development of LAL. Based on an expansion of Fulcher's (2012) questionnaire, teachers who were trained and acted as item writers responded to an online survey (N=56). The survey was designed to capture a) the teachers' reasons for their initial and ongoing involvement, b) their attitudes towards teacher involvement in the development of such tests and c) which dimensions of LAL were enhanced through their participation. The results show which assessment skills and knowledge the teachers were able to develop from participating in the exam reform project and how relevant they considered these to be for classroom practice.

Impact of language requirement policy: Students' perceptions about the use of TOEIC as an exit test

11:25–11:55 · Mario Laserna—ML—604

Ching-Ni Hsieh, Educational Testing Service, USA

Previous research on washback and impact of large-scale language tests have shown that high-stakes tests would bring about critical consequences for test takers and other stakeholders concerned (Alderson & Hamp-Lyons, 1996; Bailey, 1996; Cheng, 2005). This study investigated the use of the Test of English for International Communication (TOEIC) as an exit test within the context of higher education in Taiwan where students are required to take an external, standardized English language proficiency test and achieve a satisfactory level of proficiency before graduation. The TOEIC, a test focusing on workplace English, has become one of the most widely used tests for exit purposes in

Taiwan (Pan, 2014). However, little research exists that investigates stakeholders' views regarding the use of the TOEIC as one of the exit tests for higher education and how the relevant language policy impacts students whose graduation is directly influenced by its implementation. The study examined these issues by surveying TOEIC test takers' views regarding the use of the test and the implementation of the language exit requirement policy.

An online survey was developed and sent to more than 22,000 Taiwanese university students who had taken the TOEIC. A total of 1,528 students representing more than 100 universities responded to the survey with valid data. Twenty-six survey respondents also participated in one-on-one, follow-up phone interviews with the researcher to further explore the factors that influenced their perceptions about the language policy and the use of the TOEIC for graduation requirement. Descriptive statistics and cross-tab analyses were performed to analyze the survey data. The interview data were systematically coded and thematically analyzed using the software NVivo11. The survey responses showed that the great majority of the students considered that the English language requirement policy was necessary and important to ensure that university graduates had the English language skills required to be successful in the job market. The students who took the TOEIC to meet the graduation requirement predominately believed that they benefited from preparing to take the test and that obtaining a high score on the TOEIC would give them a competitive edge in the job market—the intended use of the TOEIC test. Results of the cross-tab analyses indicated that students' perceptions were influenced by their language proficiency. Low-proficiency students showed a stronger preference for the implementation of the language policy. High-proficiency students reported that they would prepare to take the TOEIC regardless of the exit requirement. Results of the interview data revealed that the students generally approved the language exit requirement policy because of its perceived positive impact on students' language learning. The students also viewed the use of the TOEIC for exit purposes favorably because of its high score reliability and the wide recognition of its scores for job applications. The study results have implications for similar Asian EFL contexts, where English language proficiency tests, such as the TOEIC, play a significant role in making decisions about university students' graduation and qualifications for the job market.

Developing a writing rubric using a hybrid approach

11:25–11:55 · Mario Laserna—ML—606

Jing Wei, Center for Applied Linguistics, USA

Meg Montee, Center for Applied Linguistics, USA

Tanya Bitterman, Center for Applied Linguistics, USA

Jennifer Norton, Center for Applied Linguistics, USA

The purpose of this study is to investigate the usefulness of different approaches to develop rubrics for large-scale tests. Historically, two approaches have been proposed: theoretical and empirical (Fulcher, 1996; Upshur & Turner, 1995). Proponents of the theoretical approach contend that rubric makers draw upon the theoretical framework of the targeted construct to create scoring criteria. In spite of its advantage in construct fidelity, research has shown that theoretically-oriented rubrics often use impressionistic or relativistic wording that are subject to individual rater interpretation and may result in low inter-rater reliability (Knoch, 2007, 2009). On the other hand, the empirical

approach argues that the scoring criteria should be derived from test-taker responses, which ensures that the scoring criteria capture characteristics of responses. However, rubric descriptors may lack generalizability across task types and also may vary depending on the background of the experts that are used to make the rubric (Turner & Upshur, 2002). The current study is unique in that it proposes a new hybrid approach for rubric development that is aligned with the validation framework of Evidence-Centered Assessment Design (ECD) (Mislevy, Almond, & Lukas, 2003): it suggests first creating rubric descriptors through domain modeling and domain analysis, and then using information from assessment implementation and record to hone the rubric. It also provides empirical evidence for the effectiveness of this hybrid approach by examining inter-rater reliability statistics for the newly developed rubric.

This study investigates the research question of how to create descriptors for a writing rubric by using a hybrid approach of rubric development. It has three phases: rubric development, refinement and validation. In Phase I, rubric descriptors were generated by domain experts using theoretical conceptualizations about the proficiency levels that characterize the construct of English Language Development (ELD). Six writing experts gathered together and brainstormed the descriptors by drawing upon the ELD Standards. The rubric went through an iterative process of tryout, review and revision. After the rubric was finalized and put into operational use, feedback from testing experts and professional raters suggested that the plus levels needed to be more clearly defined in order to be applied reliably. Therefore, in Phase II, rubric descriptors were further refined through a process of empirically sorting responses into specified levels and summarizing the features that typify responses at those levels. In Phase III, rubric validation, the six writing experts double scored a set of 100 responses using the refined rubric. Inter-rater reliability statistics were computed in order to show to what extent the refined rubric can be reliably applied to score writing responses.

This study has significant implications for rubric development in language tests. By incorporating both approaches for rubric development in one study, it compares their strengths and weaknesses. In addition, it shows that neither approach is sufficient to satisfy all the demands in a large-scale test development context. Rather, they should be combined in order to complement each other. Also, this study shows that rubric development should not be a one-time endeavor. It requires ongoing "retrofit" (i.e., upgrade and change) so that it can be aligned to the changing needs of test development (Fulcher & Davidson, 2009).

Seeking the construct of interactional competence in aviation: A meta-synthesis

1:30–2:00 · Mario Laserna—ML—604

Candace Farris, McGill University, Canada

The purpose of this study is to develop an empirically-based construct of interactional competence for radiotelephonic communications in aviation. We report the results of a mixed-methods meta-synthesis of 22 studies (qualitative, quantitative and mixed-methods) of naturally-occurring air traffic controller (ATC)-pilot discourse. The study is conducted within a socio-cognitive theoretical framework, according to which interactional competence is determined based on the individual's performance in interaction with the task, other individuals, and dynamic factors inherent in the communicative environment.

English is a lingua franca in radiotelephonic communications; therefore, pilots and ATC worldwide must demonstrate English language proficiency in order to operate in contexts where the use of a lingua franca may be required. Despite the high stakes implications of aviation language testing for safety and personnel licensing, there is little validity evidence to support the current construct outlined in the International Civil Aviation Organization's (ICAO's) Manual on the implementation of the ICAO language proficiency requirements (ICAO Document 9835, 2010). There have therefore been calls from within the language testing and aviation communities to revise the ICAO LPRs (e.g., Farris & Turner, 2016; ICAO Document A39-WP/249; Kim, 2012; Knoch, 2009).

Implementation of the current language proficiency requirements presents challenges for stakeholders (e.g., civil aviation authorities, test takers, test developers). Based on a content analysis of ICAO Document 9835 and the authors' professional experience with aviation language testing, we attribute these challenges to six overlapping tensions reflected in the current construct: standard phraseology vs. plain language; English as a lingua franca vs. other languages; native speakers vs. non-native speakers; context-specificity vs. generalizability; language proficiency vs. interactional competence; and routine vs. non-routine situations. We argue that underlying these tensions is an inadequate description of the construct of interactional competence. We therefore explored these tensions through a synthesis of the observations and findings of controller-pilot discourse studies conducted across a variety of contexts.

The data for this meta-synthesis are 22 studies which used consecutive recorded hours of naturally-occurring ATC-pilot discourse as data, representing over 650 hours of recordings across eight countries and 51 ATC facilities. For the development of global standards, a generalizable construct of interactional competence is required. Therefore, meta-synthesis, which allows for the inclusion of indigenous data from a variety of contexts, was deemed appropriate for our purpose.

The data were coded iteratively along two dimensions using NVIvo for Mac (Version 11). Initially, categories were not imposed in order to ensure that construct-relevant data were not excluded based on a preconceived framework. Observations and findings deemed relevant to the construct of interactional competence in accordance with our socio-cognitive framework were coded into categories which were continually reorganized. Next, for the purpose of comparing the empirically-based construct to the current construct, data from the emergent categories were categorized into the six tensions outlined above.

Findings suggest misalignment between the current construct and the empirically-derived construct that emerged in this study. Implications for future changes to aviation language policy and test development are discussed. Follow-up studies for ongoing and future development of the construct are described.

Beyond English proficiency scores: Understanding academic performance of international students in the first year at an English-medium university

1:30–2:00 · Mario Laserna—ML—606

Heike Neumann, Concordia University, Canada

Nina Padden, Concordia University, Canada

Kim McDonough, Concordia University, Canada

Past research into the relationship between scores and score profiles on English proficiency tests (EPTs), such as the IELTS and the TOEFL, has shown that there is no clear relationship between those scores and students' subsequent academic achievement. Research on the TOEFL revealed weak or no correlations between TOEFL scores and the students' subsequent grade point average (GPA) (Cho & Bridgeman, 2012; Hill, Storch, & Lynch, 1999; Light, Xu, & Mossop, 1987; Van Nelson, Nelson, & Malone, 2004; Wait & Gressel, 2009). Similarly, only weak or no correlations between students' test scores and their subsequent GPA were found in research on the IELTS (Arrigoni & Clark, 2010; Cotton & Conrow, 1998; Feast, 2002; Hill, Storch, & Lynch, 1999; Kerstjens & Nery, 2000). At the same time, Banerjee (2003) found that admissions decisions based on IELTS test scores are highly accurate because admissions counsellors base their decisions on other information in addition to the IELTS scores. In a similar vein, Rea-Dickins, Kiely, and Yu (2007) recommend that institutions "base IELTS admissions requirements on patterns of success ... [and] provide language support for students with weak IELTS profiles" (p. 62). Information about students' academic self-concept may provide additional helpful information when predicting future academic success. Research on academic self-concept has consistently shown a positive relationship between students' academic self-concept and subsequent academic achievement and education attainment both in school and higher education settings (Boulter, 2002; Cokley, 2000; Guay, Larose, and Boivin, 2004; Huang; 2011). The purpose of the current study was to examine if a measure of academic self-concept aids in the prediction of future academic performance of international students. The study focused on first year international students in undergraduate programs at the business school of an English-medium university in Canada; the business school was selected because it takes in the highest percentage of students whose first language is not English compared to the other faculties. The following information was collected about the student participants in the study: grades in English as second language and degree program courses, annual GPA, and EPT scores (including subscores). In addition, students completed an academic self-concept scale (Reynolds, Ramirez, Magriña, & Allen, 1980). To obtain additional information about what is required to succeed in business courses in the first year, instructors in two required first-year courses were interviewed about academic and language requirements in their courses and the profile of successful students in their courses. Correlations between the students' course grades, GPA and academic self-concept score were calculated. The instructor interviews were analyzed using the methods for developing grounded theory (Corbin, & Strauss, 2008). The findings from all data sources were triangulated to better understand the relationship between students' academic self-concept, and the requirements for success in first year business courses. The results will be presented, and the implications of the findings for admission of international students and in-program support of these students at university will be discussed.

Validation research of preschool language assessment for dual language learners: Collaboration between educators and test developers

1:30–2:00 · Mario Laserna—ML—615

Ahyoung Alicia Kim, WIDA, University of Wisconsin-Madison, USA

Mark Chapman, WIDA, University of Wisconsin-Madison, USA

Carsten Wilmes, WIDA, University of Wisconsin-Madison, USA

M. Elizabeth Cranley, WIDA, University of Wisconsin-Madison, USA

Tim Boals, WIDA, University of Wisconsin-Madison, USA

Despite the increase in the number of preschool-aged dual language learners (DLLs) in the U.S., there has been a general lack of options for assessing the language development of this population. To address this need, WIDA has developed a suite of formative language assessment tools (a.k.a., Language Preview) for preschool-aged DLLs. The Language Preview incorporates the following characteristics of valid and authentic assessment for DLLs (e.g., Bachman & Palmer, 1996; NAECY, 2009): (1) Culturally and linguistically appropriate assessment; (2) use of authentic materials and context; (3) promotion of family engagement; (4) interactive assessments promoting meaningful communication; (5) learner-based activities providing children opportunities to make choices and lead activities.

The goal of this paper is to describe the dynamic collaboration among educators and test developers in the creation of the Language Preview. The contributions of educators were particularly important during the pilot, bias and sensitivity review, and field test stages of development. The paper presents the findings from the field test. Participants were 37 DLL educators from various preschools in the U.S. and international preschools. The test developers consisted of five team members with varying expertise in language assessment, preschool education, and instructional design. Educators administered the Language Preview to a total of 145 DLLs, ages 2.5-5.5. The Language Preview consists of three instruments: Family Questionnaire (a survey for examining and assessing DLLs' home language and English use, to be completed with/by parents), Language Snapshot (individualized play-based activities), and Language Observation (classroom observation tool). Afterwards, teachers completed an online survey to share their feedback with the test developers. For analysis, parents' and educators' ratings of the children were examined in terms of descriptive statistics, correlation analysis, and factor analysis. Survey data were qualitatively analyzed according to emerging themes.

Overall, results show the benefits of dynamic collaboration with stakeholders, particularly educators and parents for developing a valid language assessment in a highly specific context. Findings of interest included that parents' and educators' evaluations of the children's language ability were similar in their average means and were significantly correlated. In addition, the Language Preview in general measured a unitary factor of English language ability. These findings support the validity of the assessment.

Moreover, educator survey findings suggest the usefulness of the Language Preview. Educators found the Family Questionnaire to be very helpful for better understanding the child's language use at home and they enjoyed communicating with the parents' regarding their goals for children's

language learning. Also, educators were able to identify DLLs' receptive and expressive language levels by using the Language Snapshot and Language Observation. In addition, educators liked the informal nature of the Language Observation, which allowed them to observe the DLLs in natural rather than formal test settings. However, there were challenges, for example, measuring children's receptive language or assessing children's language when they spoke their home language rather than in English. Findings inform DLLs' language development and provide practical guidance on the development of language assessment tools appropriate for preschool settings.

Skimming, scanning, search reading and reading comprehension: An exploration of constructs through exploratory and confirmatory factor analyses

2:05–2:35 · Mario Laserna—ML—604

Ángel Arias, University of Ottawa, Canada

Beverly Baker, University of Ottawa, Canada

Amelia Hope, University of Ottawa, Canada

Skimming and scanning are emphasized in the classroom and in textbooks as important and useful reading strategies. Intuitively, the importance of emphasizing such strategies seems justified, on the grounds that people need to quickly digest large amounts of reading material in order to cope with the information explosion in academic and professional contexts. However, there is almost no research on the construct(s) underlying skimming and scanning, as well as their relationship to overall reading proficiency.

This study, informed by an argument-based approach to test score validation (Kane, 2006), was undertaken to collect theoretical, empirical and practical evidence to decide upon the continued use of the skimming/scanning subtest of the CanTEST assessment. This subtest is currently distinct from the reading comprehension subtest. The CanTEST is a standardized English proficiency test used to determine whether or not candidates meet the language requirements of Canadian postsecondary institutions and Canadian professional licensing associations.

Viewed historically, the definitions of skimming and scanning have evolved, but a close relationship between the two exist: skimming and scanning are both types of speeded reading. Both also require selective processing, but the purposes of both types of reading are different: while scanning is the search for specific explicit information within a text, skimming requires inferencing, often to determine gist (Maxwell, 1972; Duggan & Payne, 2009; Rodgers, 2009; Kayleigh et al., 2015). In both skimming and scanning, lack of careful reading means that some comprehension is inevitably sacrificed.

In conducting this study, firstly, a literature review was conducted of empirical studies of skimming and scanning to explore how both activities have been conceptualized and operationalized in instruction as well as assessment. Secondly, statistical evidence was collected to examine the claim that skimming/scanning activities and reading comprehension activities tap into two distinct constructs. An exploratory factor analysis (EFA) was conducted on the tetrachoric correlation matrix of dichotomously scored skimming, scanning and reading comprehension items from a test form of the CanTEST (n = 1,549). The results suggested the retention of two factors, where skimming and scanning items loaded heavily on a different factor from reading

comprehension items. Subsequently, confirmatory factor analyses were conducted to further explore the results from the EFA analysis. The specified model fit the data well, supporting the results from the EFA analysis. Results also suggest that while search reading and scanning are different activities, they may be tapping into a similar construct related to speeded selective processing while reading.

This work has important implications for the language assessment literacy (LAL) development of both classroom teachers and assessment professionals, regarding three of the elements of LAL discussed by Taylor (2013): knowledge of theory, especially relating to possible differences in cognitive processing associated with each type of reading; knowledge of principles and concepts related to underlying constructs and the operationalization of these constructs in skimming and scanning tasks; and language pedagogy, specifically related to reading strategy instruction and how skimming and scanning are presented in materials for language learners.

Classroom educators and the design of K–12 ELP assessments: Impact on stakeholder language assessment literacy

2:05–2:35 · Mario Laserna—ML—615

Maurice Cogan Hauck, Educational Testing Service, USA

Perhaps the most extensive and impactful type of language assessments within the public educational system of the United States are K-12 English Language Proficiency (K-12 ELP) assessments. US federal legislation requires the English language proficiency of students from Kindergarten through Grade 12 who may be English learners to be assessed within 30 days of their beginning schooling for purposes of classification as English learners (ELs) or as Initially Fluent English Proficient (IFEP), and requires the English language proficiency of all students classified as ELs to be assessed annually to measure progress and inform decisions related to exit from EL services. (NCLB, 2002; ESSA, 2015).

Design and development of K-12 ELP assessments to support these uses involves a range of stakeholders including (1) state departments of education, who commission and oversee the development of such assessment systems; (2) testing companies, who work at the direction of state departments of education to develop, administer, score, and report on such assessments; and (3) educators, who will be central users of the assessments and assessment scores.

This paper focuses on the role of classroom educators as a crucial stakeholder groups commonly involved in the design, development, and administration of K-12 ELP assessments. The paper attempts to propose a conceptual framework for the educators' roles during the various stages of large-scale, high-stakes K-12 ELP assessments. The framework is drawn from empirical evidence collected from K-12 ELP assessment design efforts, including the on-going development work for a new K-12 ELP assessment system called the English Language Proficiency Assessments for California (ELPAC). This system will serve 1.4 million English learners and their educators in California public schools.

The proposed conceptual framework describes the nature and importance of the educators' language assessment literacy (LAL) at key stages, including: (1) evaluation of pilot task types; (2) item writing; (3) review of test items to ensure alignment to standards; (3) review

of performance level descriptors (PLDs); and (4) identification of exemplar responses to Speaking and Writing tasks for use in scoring and scoring training during the scoring and rubric development. At each stage, the test design is influenced by the experience of the educators, and the educators gain assessment LAL through their direct involvement in the assessment development process.

The paper provides concrete examples of how educators help to ensure alignment among educational standards the ELPAC assessments, and classroom practice, thus strengthening the validity argument for the ELPAC. Further, the paper discusses specific ways in which the LAL of the educators was increased, drawing on specific examples from the various development stages at which their involvement allowed educators to "see behind the curtain" of the often-secretive process of designing and developing an assessment system. The paper concludes with a discussion on opportunities to improve educators' LAL.

Investigating second language learners' use of linguistic tools and its effect in writing assessment

8:30–9:00 · Mario Laserna—ML—604

Saerhim Oh, Teachers College, Columbia University, USA

Advances in technology in recent years have greatly reformed the way we write (Goldberg, Russell, & Cook, 2003; Lunsford, 2006; Purpura, 2016). When we write with a computer, we can easily check our spelling and grammar, and find the meanings of uncertain words using linguistic tools (e.g., spell check, grammar check, dictionary, and thesaurus). While these linguistic tools are resources commonly used by all writers, second language (L2) learners especially make great use of them when they write, and these linguistic tools are an essential part of their writing process.

To assess students' writing in conditions parallel to how they write in real life, some assessments (e.g., National Assessment of Educational Progress and Oregon State Writing Assessment) have started to include commonly available linguistic tools in writing tests. However, these linguistic tools are prohibited in most L2 writing assessments based on the assumption that using them in tests provides an inaccurate measure of L2 learners' writing ability. Nonetheless, if L2 writing ability were re-conceptualized to include these tools, the assessments would more closely simulate writing behaviors in the real world, and we would be able to generalize writing performance in the assessment contexts to their writing ability in authentic contexts (East, 2008; Weigle, 2002).

Accordingly, this study investigated 39 adult L2 learners' use of linguistic tools in an online English writing assessment. The purpose of the study was threefold: (a) to examine the difference, if any, between L2 test-takers' writing scores with and without linguistic tools across different proficiency levels; (b) to identify the nature of L2 test-takers' use of linguistic tools in writing tests across different proficiency levels; and (c) to understand L2 test-takers' perceptions of using linguistic tools in writing tests. To achieve these goals, test-takers' written products on two tasks—with and without linguistic tools—were scored on a five-component rubric, and the difference between the two was compared using multiple paired t-tests, ANOVAs, and Many-Facet Rasch Measurement (MFRM). In addition, the entire writing process was recorded and analyzed. Lastly, survey results on test-takers' perspectives of using linguistic tools were analyzed.

The results indicated that in general, test-takers performed better with access to linguistic tools. In addition, there were clear differences among the three proficiency levels across the five components of the rubric for both settings. Bias analyses also revealed that there were no significant interactions (a) between the assessment conditions (i.e., with and without linguistic tools) and the proficiency levels and (b) between the assessment conditions and the rating scale components. The results also showed that test-takers mostly used linguistic tools when they displayed lack of knowledge of lexical form and meaning. Additionally, the survey results present a conflict between the extent to which test-takers want to use linguistic tools in the test, and their perception of test fairness. The implications of these findings will be discussed in terms of the role that linguistic tools may play in second language writing assessment.

Comparing the cognitive processes in the reading sections of IELTS and TOEFL

8:30–9:00 · Mario Laserna—ML—606

Nathaniel Owen, Open University, United Kingdom

Studies concerned with direct linking of different tests usually focus on the technical procedures of linking or comparing scores. Technical procedures of linking or comparing tests have received significant attention in the literature, with the aim of interpreting scores from different scales interchangeably (e.g., Dorans et al., 2010). The aim is to facilitate decision-making by stakeholders regarding test takers who have taken different, but supposedly 'equivalent' tests for the same purpose, such as IELTS or the TOEFL iBT.

However, these studies rarely go beyond technical considerations of score linking or test equating to compare the understanding of the construct embedded in the two (or more) tests. Tests which purport to measure the same construct (such as the above-named tests) are constructed differently and offer different item types, despite being offered for the same purpose, which is to gain admission to a programme of academic study offered with English as the medium of instruction. The extent to which test design differences are the result of institutional differences, business models or because they have different understandings of the construct is underexplored.

This paper investigates whether IELTS and the TOEFL iBT have similar or differing conceptions of the construct of reading for academic purposes from a cognitive validity perspective. Reading was selected as the focus of investigation due to the central role that this skill continues to occupy in the IELTS and TOEFL tests, despite the significant changes that have occurred in major language tests over the last 30 years.

The reading sections of IELTS and TOEFL iBT were compared through a series of twelve stimulated-recall interviews with six Chinese students who each completed one part of an IELTS and TOEFL iBT. All participants had successfully enrolled on postgraduate courses at an English university. Together, the interviews provided a complete record of one IELTS test and one TOEFL iBT. Stimuli included a video-recording of participants' observable strategic interactions with the tests which formed the basis for subsequent interviews. Participant verbalisations were coded based on a framework developed from Khalifa and Weir's (2009) model of reading. Outcomes of the research were presented in the form of detailed cognitive specification matrices (Buck, 2001) which illustrate the test developers' conception of the domain and how this is encoded in the tests and in specific item types.

The research revealed similarities and differences between IELTS and the TOEFL iBT and limitations in their conception of the construct. Both IELTS and TOEFL produced evidence of higher and lower-level processing according to the levels of Khalifa and Weir's model. One crucial aspect of the construct that both tests omit is 'creating an intertextual representation'. Neither test requires participants to combine information and form an understanding across multiple texts, suggesting that both tests have an incomplete definition of the construct of reading for academic purposes. A significant difference between the tests is that IELTS items appear to elicit a range of cognitive processes, whereas TOEFL item types target specific processes. The finely-grained use of Khalifa and Weir's reading model also suggested the boundary between higher and lower-level processing in the model is better conceptualised at the clause boundary, rather than the sentence boundary.

Towards a comprehensive, empirical model of language assessment literacy across different contexts

8:30–9:00 · Mario Laserna—ML—617

Benjamin Kremmel, University of Innsbruck, Austria

Luke Harding, Lancaster University, United Kingdom

Several scholars have suggested different models and components of language assessment literacy (LAL) for teachers, testing practitioners, and other stakeholders (Brindley, 2001; Inbar-Lourie, 2008; Fulcher, 2012; Taylor, 2013). With the exception of Fulcher (2012), however, these models have been developed through a method of abstracting from theoretical concepts to arrive at prescribed sets of components. As such, these models remain theoretical in nature, and represent the perspectives of language assessment researchers rather than stakeholders themselves. The present study aims to address this issue by developing a comprehensive understanding of LAL needs across stakeholder groups around the world through the use of an empirically-validated survey instrument which will provide evidence of (a) the distinguishability of hypothetically different dimensions of language assessment literacy, and (b) the needs, lacks and wants of different stakeholder groups across various international contexts with respect to identified dimensions.

The paper will report on findings from the development of a web-based language assessment literacy questionnaire (Kremmel & Harding, 2016), which was administered to respondents from a range of different stakeholder groups across the world to gauge their perceptions of the LAL needs of their peers. It will describe the results of the expert review and pretesting stages, and present emerging results of respondents' views of their peers' LAL requirements across three different professional groups: language teachers, language testing professionals and language testing researchers. Results from the large scale administration (estimated N>500) are expected to reveal variations in perceptions of needs according to professional group and geographical context. Using Mokken analysis to arrive at a model of empirically-based LAL factors, the paper will also illustrate how such an empirically-grounded model compares to theoretical LAL models, such as that proposed by Taylor (2013).

Defining EFL teachers' language assessment literacy and its promotion through virtual learning teams

9:05–9:35 · Mario Laserna—ML—604

Mitra Janatifar, Alzahra University, Iran

Seyyedeh Susan Marandi, Alzahra University, Iran

Esmat Babaei, Kharazmi University, Iran

Despite being trained in pre-service teacher education programs, most EFL teachers are underprepared when faced with language assessment-related activities. Part of the problem emanates from the fact that Language Assessment Literacy (LAL) as a construct has not been well defined by experts. With the aim of elaborating on the concept of LAL in the Iranian EFL context and promoting it, a two phase

study was conducted. In the first phase, an adapted version of the Language Assessment Literacy Survey developed by Fulcher (2012) was administered, which made use of two types of constructed and closed response items to help identify the EFL teachers' LAL components. The participants were 280 English language teachers from seventeen different provinces of Iran. Both Exploratory and Confirmatory Factor Analyses were used to analyze the results and define the constituents of Language Assessment Literacy as a construct. Furthermore, qualitative data analyses were employed to analyze the data obtained from the constructed-response items. During the second phase of this study, a Virtual Learning Team (VLT) was formed and used as a means of improving the Iranian English language teachers' LAL levels based on the results of the first phase of the study. A pre- and post-course survey was developed to evaluate EFL teachers' LAL levels both prior and after taking part in the VLT, in which a do-it-yourself Learning Management System (DIYLMS) was used as the technological venue. Apart from evaluating the outcome of the VLT project, post-course semi-structured interviews with the participants were also conducted to evaluate the challenges and rewards of using VLTs as a means of improving English language teachers' LAL levels in an EFL context. The results of this study can have direct implications for future teacher education programs with the aim of enhancing EFL teachers' Language Assessment Literacy levels.

Embarking on developing a meaningful measure of graduate-level L2 reading comprehension: The role of primary stakeholders' understanding of the construct

9:05–9:35 · Mario Laserna—ML—606

Catherine Baumann, University of Chicago, USA

Ahmet Dursun, University of Chicago, USA

Nicholas Swinehart, University of Chicago, USA

James McCormick, University of Chicago, USA

University of Chicago graduate students are required to demonstrate their ability to read in a foreign language in order to conduct research and participate in an international community of scholars. Like many institutions, the University had asked graduate students to show evidence of this ability through a translation exam. The long-term implementation of this exam brought a number of concerns from various stakeholders, primarily that a translation exam failed to accurately measure the skills required in the research domain. Faculty found that some students who received a high pass could not conduct research in the secondary language, or that some students who could conduct research failed the exam. Many students resented spending time and effort on a skill – translation – they knew they were unlikely to use and were further frustrated (and in some cases stalled in their progress) when unable to pass the exam, even after multiple attempts. Many international graduate students lacked the grammatical nuance required for translation into English. Instructors of newly designed reading and research courses had to deal with negative backwash in order to both prepare students to read secondary literature in the L2 (the stated goal of the course) and pass the existing external assessment measure (the translation exam). However, some departments were reluctant to change the exam format because of the translation exam's long history. Deans of students found themselves responding to concerns from all these different groups of stakeholders.

To address these concerns, the University of Chicago Language Center initiated development of the Academic Reading Comprehension Assessment (ARCATM). Reading and Research courses were modified to deliver more meaningful reading comprehension proficiency, culminating in the new exam. Ultimately, this led to campus-wide re-examination of language requirements. This presentation details the steps taken to enact this change, beginning with transferring administration of the exam to the Language Center and then conducting meetings with each department to discuss the construct of reading for research purposes, introduce the format of the exam, and convince faculty and deans of its validity. Next, we discuss the results from follow-up focus-group interviews with these key stakeholders to explore their understanding of the theoretical and pedagogical rationale underpinning the construct of academic reading comprehension and task types in the ARCATM and its impact on revisiting their language requirements.

Investigating the types and uses of feedback for middle-school English language learners' academic writing performance

9:05–9:35 · Mario Laserna—ML—617

Mikyung Kim Wolf, Educational Testing Service, USA

Yuan Wang, Educational Testing Service, USA

Saerhim Oh, Teachers College, Columbia University, USA

Fred Tsutagawa, Teachers College, Columbia University, USA

Academic writing skills are increasingly important for school-age children. The writing standards of the Common Core State Standards in the United States, for example, require that students be able to write coherently and effectively in various genres in order to be ready for college and careers. While much research has been conducted on mainstream K-12 students and adult L2 learners concerning academic writing performance, relatively little empirical research is available on the academic writing ability of K-12 English learners (ELs). The present study aimed to address this gap and offer suggestions to support ELs in their academic writing development within the context of argumentative writing.

Specifically, we focused on examining the types of feedback given for middle-school ELs' argumentative writing as well as the relationship between writing quality and students' use of feedback. The data for the study were collected from a classroom-based writing assessment task. The ultimate goal of the study is to use the assessment results to help teachers by providing useful suggestions about giving feedback that meets ELs' needs.

A total of five teachers and their students participated in this study (106 ELs and 24 non-ELs as a comparison group). In order to generate automated diagnostic writing feedback, the Criterion® Online Writing Evaluation tool was utilized. The students first wrote an argumentative essay for a given prompt. The students then received immediate feedback from Criterion, followed by their teacher's feedback a day later. The students revised their writing based on the feedback they received. The students' pre- and post-feedback essays were scored on a 4-point holistic scoring rubric. The Criterion and teacher feedback types were analyzed by a pair of researchers based on a coding scheme that categorized the types of feedback into language use, organization, and content development. Descriptive statistics and multiple ANOVA analyses were conducted to

examine the quality of the pre- and post-essays and degree of feedback usage across students at the different score levels.

Results indicated that teachers tended to provide the majority of their feedback on grammatical errors in EL writing. The second most frequently commented area of feedback regarded content development (e.g., clear claims and supportive evidence). Grammatical feedback was also the most prevalent category from Criterion, particularly for ELs with scores of 1 and 2. With respect to student feedback use, on average, students attempted to address the feedback from both Criterion and teachers in all areas (e.g., content, organization, grammar, vocabulary, and mechanics). Generally, students improved their essays after feedback. ELs at the higher level tended to improve their syntactic accuracy whereas ELs at the lower level improved in vocabulary. However, the level of argumentation was not found to differ significantly in post-feedback essays. In this presentation, we will present these results in detail and discuss implications for teachers' understanding about the argumentative writing construct and nature of feedback given to ELs in their writing, as they are related to the conference theme of assessment literacy at the teacher level.

Video-conferencing and face-to-face delivery of a speaking test: Implications for different assessment stakeholders

9:40–10:10 · Mario Laserna—ML—604

Vivien Berry, British Council, United Kingdom

Fumiyo Nakatsuhara, University of Bedfordshire, United Kingdom

Chihiro Inoue, University of Bedfordshire, United Kingdom

Evelina Galaczi, Cambridge English Language Assessment, United Kingdom

Mina Patel, British Council, United Kingdom

Face-to-face tests for assessing spoken language ability offer many benefits, particularly the opportunity for reciprocal interaction. However, face-to-face speaking test administration is often logistically complex and resource-intensive, and the face-to-face mode may therefore be impossible to conduct in geographically remote or politically sensitive areas. This paper reports on a 3-phase validation study, the findings of which have implications for a wide range of different assessment stakeholders including test-providers, test-takers, examiners, raters, and test-score users, in addition to technicians and staff who facilitate the administration and marketing of speaking tests.

Phase 1 consisted of a small-scale initial investigation into what similarities and differences in scores, linguistic output and test-taker and examiner behaviour could be discerned between the two formats. Phase 2 of the study was a larger-scale follow-up in which 99 test-takers took two speaking tests under face-to-face and computer-delivered conditions. Performances were rated by 10 trained examiners. A convergent parallel mixed-methods design was used to allow for collection of an in-depth, comprehensive set of findings derived from multiple sources. The data collected included an analysis of feedback interviews with test-takers as well as their linguistic output during the tests (especially types of language functions) and score ratings awarded under the two conditions. Examiners responded to two feedback questionnaires and participated in focus group discussions relating to their behaviour as interlocutors and raters.

MFRM analysis of test scores indicated that there were no significant differences in scores awarded on the two modes, although some qualitative differences were observed in test-takers' functional output and examiners' behaviour as raters and interlocutors. 71.7% of test-takers preferred the face-to-face test; 39.4% reported that there was no difference in the difficulty of the two modes. The majority of examiners reported that under the video-conferencing condition test-takers had more opportunity to demonstrate their level of English proficiency and that it was easier for them to rate test-taker performance. Subsequent bias analysis indicated that there were a number of significant interactions in both modes of delivery which need to be considered.

Phase 3 of the study aimed to investigate and test a technological solution that could be used to deliver the Speaking test in video-conferencing mode. Eight examiners in Bogota and Buenos Aires conducted video-conferencing speaking tests with 89 test-takers in Medellin, Mexico City and Caracas. All test sessions were double-marked by different examiners using video-recorded performances. Initial findings from a MFRM analysis of scores support the scoring validity of the video-conferencing tests conducted in this phase of the project, although comments from examiners during focus group discussions and test-takers' responses to a questionnaire suggest that there are still a number of issues to be addressed before video-conferencing can be introduced as a viable alternative mode of delivery for speaking tests.

The presentation will conclude with a discussion of the comparability of the construct(s) measured by the two delivery modes, and the implications this has for the interpretation of the scores obtained from the respective modes; it will also highlight the importance of collaboration with all assessment stakeholders.

Using assessment information: How young English learners interpret assessment information on score reports

9:40–10:10 · Mario Laserna—ML—606

Alexis A. López, Educational Testing Service, USA

Jonathan Schmidgall, Educational Testing Service, USA

Ian Blood, Educational Testing Service, USA

Jennifer Wain, Educational Testing Service, USA

One of the key elements in learning-oriented assessments is using assessment information to enhance learning (Carless; 2009; Turner & Purpura, 2015). Assessment feedback is vital to the learning process and it is a critical aspect of learning environments (Hattie & Timperley, 2007; McNamara, Jackson, & Graesse, 2010). If young English learners (ELs), especially students with low English proficiency, are able to understand assessment feedback, they could use this information to enhance their learning. In this study, we examine how to provide assessment information on score reports so that all students are able to understand it. In particular, we focus on how to provide clear assessment feedback to middle school ELs and their perceptions about the usefulness of the feedback. We addressed these specific research questions: 1) What are the main characteristics of effective feedback for middle school ELs? 2) What type of feedback do

they need to support their learning? 3) How do they use this feedback? And 4) What perceptions do they have of the assessment feedback they get?

The study was conducted in three sequential phases. Phase 1 attempted to answer the first research question and focused on investigating the main features of effective student assessment feedback (i.e. clear, meaningful and actionable feedback). We reviewed the literature on student feedback and score reporting, and also reviewed 80 existing score reports to identify the main characteristics of effective feedback for young learners. In Phase 2, we developed four mock-up score reports based on the findings in Phase 1. Each report presented information in different ways (e.g., using visuals, tables). We conducted cognitive interviews with 20 middle school EL students to examine the usefulness of different types of diagnostic information to enhance learning. We got information about whether the students knew how they performed, what the scores meant, and how they would use the information to enhance learning. We also asked students about their perceptions of the different types of feedback.

In Phase 3 we conducted eight focus group meetings with middle school EL students. The purpose of this phase was to have a better understanding on how to provide feedback to young ELs and how they use this information to enhance learning. There were four students in each focus group meeting and each group was made up of a diverse group of students in terms of English language proficiency, home language, gender, number of years in the United States, and grade level. Questions for the group addressed issues such as the types of feedback they get, how and when they get the feedback, how they use the feedback, their perceptions about the feedback, and the type of feedback they would like to get. All cognitive interviews and focus group meetings were audio recorded and transcribed for research purposes. Data was analyzed qualitatively in reference to the research questions being addressed in this study. In this presentation we will report our findings, including some key issues to consider in providing assessment feedback and developing score reports for young ELs.

Washback on classroom testing: Assessment literacy as a mediating factor

9:40–10:10 · Mario Laserna—ML—617

Doris Froetscher, Lancaster University, United Kingdom; Austrian Ministry of Education, Austria

Assessment literacy plays a vital role in teachers' ability to manage challenges in classroom assessment. Teachers necessitate a repertoire of methods and the knowledge to adapt, adopt or develop them for particular classroom situations (Turner, 2012). While external exams or curricula have increasingly taken criterion-referenced approaches and adopted standards such as the CEFR (Katz & Gottlieb, 2012), classroom assessments offer the possibility to tap into aspects of students' mastery which standardized tests, due to their inherent limitations, are unable to gauge (Llosa, 2011). However, this complementary relationship may be jeopardized by washback from the external exam on classroom assessment (Wall & Alderson, 1993).

By focusing on the under-researched area of classroom testing, the research reported in this paper addressed the intersection between washback and assessment literacy. It is situated in a context where a high-stakes secondary school-leaving exam, previously teacher-developed, was reformed into a professionally developed CEFR-linked standardized national exam. The present study looked into assessment literacy as a mediating factor influencing the washback effect of the new exam on summative classroom testing practices.

To this end, actual classroom test reading tasks as well as teacher perspectives were examined in a two-phase mixed-methods study, investigating differences in task characteristics and practices between the periods before (pre) and after (post) the reform. More precisely, the first phase consisted of a detailed analysis of 173 classroom test reading tasks using a specially-designed and validated instrument, and a subsequent statistical analysis against teacher variables, for example training attended. The second phase of the study employed an exploratory teacher questionnaire as well as semi-structured interviews with nine teachers, coded thematically using MAXQDA. As the data was not only collected through teacher self-reports but also through test task analysis, this design meets the need for more objective measures for washback.

This study provides insights into the nature of washback on classroom testing, which is under-researched to date. Results show a lack in teachers' agency regarding their classroom tests, and participants largely made the exam reform accountable for their changed classroom testing practices. These included, for instance, a shift in test methods and a turn to using existing tasks, particularly past exam papers, rather than designing tasks themselves. Participants stated they did not have the training and know-how in task design to be able to design their own classroom test tasks. Thus, specifically relevant to the LTRC 2017 theme, the corroborated findings from all three kinds of data collected in this study highlight the importance of assessment literacy as a mediating factor in washback on classroom assessment. The behaviour of orientating classroom test tasks too much towards the external exam poses the risk of narrowing the construct; it would thus be highly desirable for teachers in our context to build agency in classroom testing through acquiring more assessment literacy.

An evidence-based approach to generating the Language Assessment Literacy profiles of diverse stakeholder groups

10:40–11:10 · Mario Laserna—ML—604

Sheryl Cooke, British Council, China

Colin Barnett, British Council, China

Olena Rossi, Lancaster University & British Council, United Kingdom

The unitary model of validity (Messick, 1989) and Kane's (2012) chain-of-inferences argument have highlighted the range of stakeholders who can affect test validity at different stages in the test cycle from construct definition and test design through to score use and decision-making (Fulcher, 2012; O'Loughlin, 2013; Inbar-Lourie, 2013; Taylor, 2013). This suggests that stakeholder Language Assessment Literacy (LAL) is a crucial component of test validity. The construct of LAL is elusive, however, in particular because the needs of different stakeholders are so diverse and, while LAL profiles have been proposed for certain groups, these remain hypothetical and an evidence-based approach is needed.

This paper presents a model to operationalise the definition of the LAL construct across different stakeholder groups. The model uses a participatory, collaborative instrument and grounded-theory approach to generate specific LAL profiles and engage stakeholders, raising language assessment awareness in the process. Stakeholders were invited to attend face-to-face LAL workshops that targeted their particular interest group. Input on the concept of 'validity', with content tailored for each specific group, was delivered. Participants then collaboratively worked

through a series of tasks designed to relate their role directly to test validity and to identify, in each group's case, what areas of knowledge are necessary to safeguard the value of the test and mitigate potential stakeholder-specific threats to overall validity. Questionnaires were administered before and after the workshops. The grounded-theory analysis was conducted on data obtained during the workshops and from the surveys. At least 100 stakeholders have participated so far and workshops are on-going.

The presentation describes the tool and the underlying rationale and presents worked examples of the model with three different stakeholder groups in East Asia: language teachers, language assessors, ministry officials and those engaged in the sales and marketing of tests – the test marketeers. We demonstrate the LAL profiles for each group and how these are generated. The model is reviewed and recommendations for future iterations are made; the presentation concludes with a proposal for the potential use of the instrument with different groups to contribute to a cross-stakeholder LAL inventory that could be used as the basis for LAL literacy raising material and training development as well as to inform the generation of LAL profiles for diverse stakeholder groups.

Influence of contextual specificity on interaction between language ability and specific-purpose content knowledge: The case of the OELPE ESP test

10:40–11:10 · Mario Laserna—ML—606

Eunice Eunhee Jang, Ontario Institute for Studies in Education, University of Toronto, Canada

Andrea Strachan, Touchstone Institute, Canada

Jeanne Sinclair, Ontario Institute for Studies in Education, University of Toronto, Canada

Elizabeth Larson, Ontario Institute for Studies in Education, University of Toronto, Canada

Juliana Gallo, Touchstone Institute, Canada

With an unprecedented record of migrations around the world, assessing English (or other target languages) for specific purposes (ESP) is increasingly used for professional and workplace-related purposes and also for supporting migrant professional workers' needs. While ESP is not different from other language assessments in terms of measurement requirements as well as validity and ethical standards (Douglas, 2013), its highly specialized language use varying across context (e.g., nursing, optometry, aviation, business) requires a deep understanding of the characteristics of target language use (TLU), contextual variation, and the degree of dependency of language competence on background content knowledge for specific purposes (Chapelle, 1998; Davies, 2001).

Despite general consensus that context is a part of ESP construct (Jacoby & McNamara, 1999), there remain questions of how to ensure the (simulated) authenticity of specific-purpose language performance. Building on Douglas' view of authenticity as a 'continuum of specificity' (2005), the present study investigated the relationship between the degree of ESP test specificity and test performance on an optometry-specific language proficiency assessment. Specifically, we examined two key dimensions of ESP assessment: task authenticity and interaction between language knowledge and specific-purpose content knowledge. It was hypothesized that task difficulty and cognitive complexity are associated with the degree of authenticity in the form of a continuum of specificity.

The present study involved test performance data from the Optometric English Language Proficiency Assessment (OELPA), which is designed to guide admission and language training placement decisions for international optometric graduates (IGOs). Approximately 80 candidates were recruited from IGOs currently enrolled in a bridging program. The participants responded to the listening (35 items), reading (46 items), and writing (2 tasks) subtests of OELPA. A subsample of 30 candidates were selected to complete the OELPA speaking test (4 tasks). We sought input from 4 subject matter experts (SMEs) certified professional optometrists) as well as linguistic experts via questionnaires about the degree of specificity and cognitive complexity required by items and tasks. The characteristics of authenticity were examined by developing task/item/passage specifications based on target language use (TLU) domain, tenor, field, genre, text type, cultural specificity, and concreteness of information. We further examined the extent to which test taker performance is accounted for by task specificity variables.

The study results suggest that when tasks and stimuli are sufficiently specified, requiring a high level of optometric content knowledge, IGOs are more likely to demonstrate specific-purpose language ability. However, a high degree of specificity resulted in lowering task discrimination estimates when tasks tended to elicit basic comprehensions and descriptions of factual information. The study findings were further substantiated based on the SMEs' qualitative input. The study highlights the importance of developing ESP tasks that are sufficiently specific and elicit cognitive and linguistic features required for authentic TLU contexts.

How teachers' conceptions mediate their L2 writing assessment practices: Case studies of ESL teachers across three contexts

10:40–11:10 · Mario Laserna—ML—617

Antonella Valeo, York University, Canada

Khaled Barkaoui, York University, Canada

Assessment practices occupy a large portion of ESL teachers' time and have a great impact on instruction and students' engagement and learning in the ESL classroom (Leung, 2005; William 2001). These practices are in turn strongly influenced by teachers' conceptions of assessment and second-language (L2) learning and teaching and the contexts in which they work. Teachers' conceptions of assessment, defined as how teachers understand or view the roles, purposes, and impact of assessment and grading in the classroom, are closely related to the concept of teacher language assessment literacy (LAL) as defined in the literature (e.g., Fulcher, 2012; Taylor, 2013). Teachers' conceptions of L2 learning and teaching (i.e., how and why L2 are learned and should be taught) are central to examining teacher LAL and classroom assessment practices. Contextual factors such as learner characteristics and goals, program policies, and institutional culture have a significant impact on how practice is mediated by teacher conceptions.

Nevertheless, most previous research has focused on surveying ESL teachers' assessment skills, training and needs (e.g., Brown & Bailey, 2008); examining ESL teachers' classroom assessment practices (e.g., Cheng et al., 2008); and/or telling teachers how assessment ought to be done (e.g., Boyles, 2005; Hoyt, 2005; Popham, 2004; Stiggins, 2007). There is little research examining how ESL teachers actually conceptualize and conduct assessment in the ESL classroom or the

relationship between how teachers understand L2 assessment, learning and teaching, on one hand, and the decisions they make when assessing their students, on the other (Leung, 2005). Consequently, this study examined how ESL teachers' writing assessment practices in the ESL classroom are mediated by their theories and conceptions of assessment and L2 learning and teaching, as well as constraints and factors embedded in their teaching contexts.

Using a case study approach, we investigated the assessment practices and conceptions of 12 ESL teachers of adults in three teaching contexts in Canada: immigrant settlement programs, university academic preparation programs, and undergraduate credit-bearing ESL programs. Case studies were developed for each of the teachers using data gathered from classroom observations of writing assessment and instruction, analyses of teacher assessment of students' writing, stimulated recalls by teachers about their assessment practices, and in-depth interviews exploring teachers' conceptions about learning, teaching and assessing L2 writing (including task design, feedback, and grading). Data analysis examined how teachers' assessment practices are shaped by their conceptions of assessment, the nature of L2 writing, and L2 learning and teaching, as well as factors and constraints in their instructional contexts.

Preliminary findings suggest that (a) teachers hold varying conceptualizations of the nature, purposes and quality of writing assessment in the ESL classroom; (b) teachers' conceptions of L2 learning, teaching and assessment play a complex role in mediating how they develop, use and interpret assessments to evaluate and support their students' writing development; and (c) context has a varied impact on teachers' assessment practices. We discuss the findings and their implications for research on LAL, classroom practice, and the design, and implementation of assessment-related policies and ESL teacher professional development.

Developing assessment literacy in a dynamic collaborative project: What teachers, assessment coordinators, and assessment researchers can learn from and with each other

11:15–11:45 · Mario Laserna—ML—604

Claudia Harsch, University of Bremen, Germany

Sibylle Seyferth, University of Bremen, Germany

Anikó Brandt, University of Bremen, Germany

We report on an assessment literacy project situated in a languages center serving four universities, where over 60 teachers provide courses for 21 languages. The courses' learning outcomes are aligned to CEFR-levels. Teachers develop course-based achievement tests, which have to reflect the CEFR-based learning outcomes, while they have to be graded on a university-specific grading scale. Many teachers expressed their need for support when developing these tests, since most teachers have not had any training in assessment. Moreover, students complained that tests often were not comparable across courses and languages.

Our project follows Taylor's (2013) view of differentiated assessment literacy needs amongst different stakeholder groups, using an approach where all stakeholders involved bring their experience, abilities, skills and knowledge to the table. The project team comprises teachers (varying in numbers), five course/assessment coordinators whose job requires a thorough understanding of assessment principles and impact, and two researchers with a background

in assessment (one PhD researcher, one senior researcher). We see the development of assessment literacy in our context as a long-term project, and will report on insights gained during the first 18 months.

The project is needs-based and teacher-driven. Our teachers traditionally organize a monthly lunch to discuss pedagogical issues; assessment issues arising from the lunches were taken up in three pedagogical conferences and a series of seminars so far. Insights from the lunches, conferences and seminars are put into action in working groups where teachers, coordinators and researchers work and learn with and from each other.

The project is regularly evaluated, not only to monitor effects and impact, but also to instigate reflection and self-regulated professional development in our stakeholder groups. We use the following instruments: anonymous online surveys between terms to monitor different stakeholders' perceptions of the project, and to explore their assessment beliefs, practices and needs; interviews with stakeholders (individual and focus groups, semi-structured) to explore change processes and the factors which mediate change; reflective diaries to document and reflect changes in practice, beliefs and perception; student evaluations at the end of each term (including their perceptions of assessment practices and the final tests).

Aspects covered so far include principles of valid test design; CEFR and constructive alignment of learning, teaching and assessment; test-/task formats suitable for communicative assessment. The practice-oriented sessions explored local practices in assessing writing across languages and CEFR-levels; we are currently co-developing test specifications and writing tasks across languages for different CEFR-levels; and we benchmarked written performances taken from a German-as-foreign-language course, thus relating the university grading scale to CEFR-levels.

We will present selected results from the analyses of survey, interview and diary data, to examine the ways beliefs and practices are influenced by the project, and to explore factors which facilitate or hinder change. Furthermore, we present one example to illustrate our approach: the aforementioned benchmarking of written performances allowed us to align the university-grade-system with course-specific learning goals and CEFR-levels - an innovation which would not have been possible without the different stakeholders listening to and learning from each other.

How to move from indigenous criteria to a usable scale? Evolving guidelines for ESP test developers

11:15–11:45 · Mario Laserna—ML—606

Ute Knoch, University of Melbourne, Australia

Catherine Elder, University of Melbourne, Australia

Annemiek Huisman, University of Melbourne, Australia

Eleanor Flynn, University of Melbourne, Australia

Elizabeth Manias, Deakin University, Australia

Tim McNamara, University of Melbourne, Australia

Robyn Woodward-Kron, University of Melbourne, Australia

It is generally acknowledged that the knowledge base constituting 'language assessment literacy' may differ for different assessment purposes – whether general or specific – and for different parties - whether language teachers, score users or test developers themselves (Inbar-Lourie, 2012). We would also emphasize that this language assessment knowledge base is not static and may need to evolve in response to new challenges in our field. This paper describes how a team of test developers, health professionals and language raters responded collaboratively to one such challenge in the relatively under-researched area of language for specific purpose (LSP) rating scale development.

The criteria used to assess performance on language for LSP tests tend to be generic rather than profession-specific in nature, thereby limiting test authenticity and validity. For this reason, researchers have recommended involving domain experts in the rubric design process by eliciting profession-relevant, indigenous criteria and applying these to assessing performance on the LSP test (see e.g. Jacoby, 1998; Jacoby & McNamara, 1999; Douglas, 2001; Pill, 2016). However, these indigenous criteria, derived as they are from people outside the assessment field, cannot always easily be converted to a rating scale for a language test because they (a) do not adequately differentiate proficiency levels and (b) may not always be relevant to the more abstract and decontextualized testing domain. Previous work in this area has been scarce and there are therefore few models to guide test developers.

The current paper addresses this problem with reference to the writing component of the Occupational English Test (OET), an LSP test designed to assess the English communication skills of overseas-trained health professionals who apply to practise in Australia. While OET tasks are derived from a needs analysis of workplace writing demands, the assessment criteria have no such basis in empirical research. The aim of our study is to better align the OET writing assessment criteria with what members of the health profession perceive as important for written healthcare communication.

The paper outlines the different stages of the study from eliciting the indigenous criteria from health professionals of nursing and medical backgrounds through to the creation of draft checklist indicators based on their perceptions and experiences, the involvement of OET raters in converting these indicators into a useable rating scale for language assessment purposes, and the trialling of this new scale to assess a sample of OET writing

scripts. By presenting data from all these stages, we outline one possible process for the creation of professionally-relevant assessment criteria, which may be used as a blueprint to guide language testing practitioners in scale development on other specific-purpose tests. Implications are drawn for the notion of assessment literacy as a dynamic construct which needs to be adapted for different audiences and purposes in light of new knowledge and experience in our field.

Designing L2 writing assessment tasks for the ESL classroom: Teachers' conceptions and practices

11:15–11:45 · Mario Laserna—ML—617

Khaled Barkaoui, York University, Canada

Antonella Valeo, York University, Canada

This qualitative study examined how ESL teachers design tasks to assess their students' second-language (L2) writing abilities. Most of the literature tends to focus mainly on guidelines for developing writing tasks for large-scale assessment (e.g., Bachman & Palmer, 2010; Weigle, 2002) and/or advising teachers on how to design writing tasks (e.g., Kroll & Reid, 1994; Reid & Kroll, 1995). Consequently, little is known about how ESL teachers actually design and use tasks to assess their students' L2 writing abilities and the factors and reasoning involved in this process. As several authors have argued, in order to address this gap, the focus should shift from building theoretical models of task design and then trying to apply them to the classroom, to exploring what teachers consider and do when they design, select and use assessments in the classroom, what they look for when they assess, and how these practices are shaped by their conceptions and assumptions about the nature of L2 learning, teaching and assessment (e.g., Leung, 2005; Rea-Dickins, 2001, 2004; Wiliam, 2001). Such research can provide significant insights concerning teacher language assessment literacy (LAL) as defined in the literature (e.g., Fulcher, 2012; Taylor, 2013), particularly in relation to how teachers understand the roles, purposes, and impact of assessment as well as their conceptions of the nature of L2 writing ability and learning.

This study investigates and compares the decision making processes of 12 ESL teachers teaching in three ESL teaching contexts for adults in Canada: immigrant settlement programs, university academic preparation programs, and undergraduate credit-bearing ESL programs. The three contexts vary widely in terms of learner characteristics and goals, program policies, and institutional culture. Each teacher was interviewed about his/her educational and professional background and experiences, as well as his/her conceptions about learning, teaching and assessing L2 writing. Each teacher was also asked to bring samples of writing assessment tasks s/he had designed and used and to discuss the rationale, considerations, and processes involved in developing and using each assessment task as well as his/her evaluation of the qualities and effectiveness of each task. Data analysis focused on the factors that influenced teachers' decisions when designing and using writing assessment tasks, particularly teachers' conceptions concerning the nature of (good) writing, how writing is learned and should be assessed, what constitutes fair and effective assessment, and contextual factors and constraints at play in their instructional settings.

Findings suggest that (a) teachers hold varying conceptualizations concerning how and why to design and select writing tasks for the ESL classroom; (b) a variety of individual and contextual factors shape how teachers design and evaluate the quality of writing tasks; and (c) teachers' conceptions and contexts significantly mediate how they design and select writing tasks. The findings contribute to our understanding of teachers' LAL, teachers' writing assessment practices, and how they are influenced by teachers' conceptions and contexts. These findings have implications for classroom practice and writing theory and research in multilingual contexts.

DEMO: Automated task-based feedback in a formative assessment of workplace English speaking skills using a spoken dialog system

11:50–12:20 · Mario Laserna—ML—604

Eugene Tsuprun, Educational Testing Service, USA

Keelan Evanini, Educational Testing Service, USA

Veronika Timpe-Laughlin, Educational Testing Service, USA

Ian Blood, Educational Testing Service, USA

Automated spoken dialog systems (SDS) are a promising method for formative assessment and Computer Assisted Language Learning (CALL) applications since they enable language learners to practice speaking in interactive dialogs in the target language without the need for a live instructor. However, in order to be useful for the learner, SDS-based applications need to provide meaningful feedback that can be used to assess the learner's performance in meeting the demands of a given task and improve relevant linguistic skills. The feedback provided by most speech-based CALL systems to date has been limited to the domain of pronunciation, as this type of feedback is easily generated using an automatic speech recognizer. However, in order in order to increase the potential usefulness of SDS-based learning and assessment applications, research and development efforts should focus on generating feedback on other aspects of the learner's performance—including grammatical and pragmatic phenomena—that are also important for meeting the communicative demands of a given task.

This demo presentation describes an SDS-based formative assessment prototype that was designed to provide feedback on grammar, pragmatics, and task completion. Several interactive, scenario-based tasks were developed using an open-source SDS framework for the domain of global workplace English. Additionally, automated capabilities were designed to provide feedback on how well learners meet the communicative demands of workplace tasks such as placing an order with a customer service representative, requesting a refund, scheduling a meeting, and asking a supervisor to review a document. We will first show examples of non-native English speakers interacting with several of these tasks, highlighting the different types of linguistic feedback provided. We will then describe in greater depth how grammatical and pragmatic feedback are provided in one example task. This example will document the end-to-end design and development process including (a) the definition of the speaking construct, (b) the communicative goals targeted by the task,

(c) the conceptualization of the stimulus materials and dialog components, (d) the design and implementation of a branching conversation flowchart using a graphical editor that is linked to the SDS, (e) the initial specification of expected responses and associated semantic categories required by the speech processing components of the SDS, (f) the deployment of the task through a video-based SDS on a crowdsourcing platform, (g) the iterative refinement of the task design and the components of the SDS through ongoing data collection, and (h) the feedback that is provided to the learner based on the path taken through the conversation and a linguistic analysis of the content of the learner's responses. Finally, results of a user perception survey about the usefulness of the SDS system's feedback will be presented.

To summarize, this demonstration will provide attendees with an in-depth look at the types of automated feedback that can be provided in a task-based SDS language learning application and how tasks can be designed to optimize the performance of the SDS in processing a learner's response and providing accurate feedback.

Assessment competencies for effective feedback: Psychological and environmental factors influencing language assessment

11:50–12:20 · Mario Laserna—ML—617

Maryam Wagner, McGill University, Canada

Eunice Eunhee Jang, Ontario Institute for Studies in Education, University of Toronto, Canada

Feedback is central to learning. It may be used by learners to "confirm, add to, overwrite, tune, or restructure information in memory, whether that information is domain knowledge, meta-cognitive knowledge, beliefs about self and tasks, or cognitive tactics and strategies" (Butler & Winne, 1995, p. 275). Providing effective feedback demands that teachers possess the ability to observe, interpret, measure, and describe students' knowledge and skill development. This set of diagnostic competencies (Edelenbos & Kubanek-German, 2004) intersects with the language assessment literacy that teachers need to effectively evaluate students, and subsequently use to advance their language learning (Fulcher, 2012; Inbar-Lourie, 2013; Malone, 2013; Taylor, 2009). Further, there are psychological and environmental factors that influence assessment and hence, the provision of teachers' feedback. The purpose of this investigation was to identify integral factors influencing students' assessment based on a descriptors-based language-assessment framework in a K-12 learning context, and how the information can contribute to understanding facets essential to augmenting teachers' language assessment competence for providing feedback.

We tracked 42 teachers' assessment of 159 students' language development using a descriptors-based language assessment framework across three school boards. We gathered multiple types of evidence to understand factors contributing to teachers' use of the framework drawing on classroom observations, interviews, and use of tasks to elicit linguistic behaviors. Using surveys, we gathered background information (e.g., home language environment, engagement in literacy activities, age), and motivation (e.g., goal orientations) about students to understand the contribution of individualized factors influencing students' language development as measured by the assessment framework.

Exploratory factor analysis of goal orientation data revealed a five-factor structure, three of which were congruent with the theoretical three factor structure in the goal-orientation literature (mastery, performance prove, performance avoid (Dweck, 1986; Harackiewicz et al., 2002; Nicholls, 1998). It was hypothesized the additional two factors were indicative of the complexities associated with language learning, and were identified as: academic achievement-oriented, and attitude to English. Together, the 5 factors explained approximately 50% of the variability that was observed in the participants' responses. We used the results of the factor analysis to use and group items for subsequent latent profile analyses (Collins & Lanza, 2010; Lazarsfeld & Henry, 1968) to determine if the variables could be classified into mutually exclusive subgroups to characterize learners and inform feedback and instruction. A four-class structure was identified comprising four groups (classes) of students: academic achievers, performance-prove oriented, mastery oriented, and possessing negative attitudes to English.

Thematic analysis of interviews and classroom observations revealed three primary facets of the classroom environment that contributed to the assessment of learners in the context of a language assessment framework: characteristics of the assessment task, and teachers' assessment competence, and the availability of instructional support. This paper elaborates on the complexities of environmental and cognitive factors that influence students' language learning, and its implications for advancing teachers' assessment competence, and specifically, the associated knowledge and skills necessary to provide effective feedback.

Defining the "just qualified" speaker: Comparisons from two CEFR linking panels

1:30–2:00 · Mario Laserna—ML—604

Stephen O'Connell, Cambridge Michigan Language Assessments, USA

Sharon Pearce, Cambridge Michigan Language Assessments, USA

A key aspect of successful standard-setting studies is defining the minimally competent or just-qualified candidate (Cizek et al., 2004). Failure to take the just-qualified learner into account can result in cut scores that are too high (Lim et al., 2013; Papageorgiou, 2010), which can have serious consequences. As such, much of the literature on standard setting emphasizes the importance of training panelists to envision candidates who are just acceptable enough to be classified at a given performance level (Cizek & Bunch, 2007; Mills et al., 1991). The purpose of said training is that the subjective differences that standard-setting panelists inevitably bring to their decisions should, to the greatest extent possible, stem solely from the features of the test items or the test performances that they are considering, and not from their individual perceptions of just qualified (Mills et al., 1991).

However, despite its importance and despite multiple reports of panelists struggling with the concept of just qualified (Hein & Skaggs, 2010; Mills et al., 1991; Papageorgiou, 2010), documentation on how to address the issue in practice is scarce. This is the case even when the performance-level descriptors are from the widely used (though sometimes criticized as under-specified) CEFR. This is the gap that this research attempts to address: a detailed exploration of how just-qualified definitions were developed from the CEFR descriptors and an investigation of how those definitions may have influenced standard-setting panelists.

In this presentation we first explain how just-qualified CEFR B1-, B2-, and C1-level speakers were defined by the facilitators in the context of two CEFR linking studies: one conducted for a commercially available English speaking test, and one that was conducted for a high-stakes secure speaking test of academic English. We then discuss data collected from 25 judges (12 in the first linking study and 13 in the second). The judges completed pre- and post-study activities that asked them to provide their definitions of just-qualified speakers. Following methods used in analysis of verbal protocol data (e.g., Green, 1998; Brown, 2007), in two rounds of coding, we tagged our judges' just-qualified definitions for tokens of both "limiting" language (explicitly negative language) and "boosting" language (explicitly positive language). Our hypotheses were that the facilitators' repeated focus on the concept of just qualified would result in participants' increased usage of "limiting" language in their post-study just-qualified definitions, and decreased usage of "boosting" language in their post-study just-qualified definitions (in comparison to pre-study definitions).

Findings from the first linking study panel that supported those hypotheses (a trend of increased use of limiting language and a decreased use of boosting language) will be compared with results from the second panel. The degree to which panelists seem to absorb the facilitators' definitions and the effect that has on cut scores will be discussed, as will, more generally, the implications and challenges of using the CEFR in this way.

Mind the gap: Bringing teachers into the language literacy debate

1:30–2:00 · Mario Laserna—ML—606

Vivien Berry, British Council, United Kingdom

Susan Sheehan, University of Huddersfield, United Kingdom

Sonia Munro, University of Huddersfield, United Kingdom

Teachers' attitudes and beliefs are frequently cited as exerting a powerful role in shaping their decisions, judgements and behaviour (Borg, 2006; Kagan, 1992). Consequently, exploring teachers' levels of assessment literacy may help teacher educators to better understand the factors which promote or prevent effective assessment, and thus contribute to more targeted teacher education. Much previous research into teachers' assessment literacy has relied on survey data (Fulcher 2012, Plake & Impara 2002)). The research to be discussed in this presentation focuses on the sociocultural context in relation to actual assessment literacy practices in the language classroom, since an investigation into what is happening in classes may be of little value without exploring why it is happening. With the exception of a case study following three Chinese University teachers (Xu 2015), no teachers have been asked directly about their attitudes to assessment or their specific training needs. This project sought to bring teachers more directly into the assessment literacy debate in order to provide them with training materials which meet their actual stated needs.

The initial phase of the project consisted of a series of interviews and observations of experienced teachers, conducted at the international study centre of a British university. The interviews drew on Davies' (2008) components of assessment literacy which, following Stiggins (1991, 1997) he defined as Skills + Knowledge but with the important addition of Principles. In the interviews, teachers were invited to estimate their understanding of the components of the assessment process and asked to indicate how much they would like to learn about each individual component. Observations were then conducted which focused on teachers' assessment practices in the classroom. Post-observation interviews were subsequently conducted with the teachers, in which they were asked to reflect on their observed classroom practice. In the second phase of the project, focus group discussions were held with experienced teachers at teaching centres attached to a major international organisation in two European countries. These teachers taught a variety of different English language classes across a range of ages and proficiency, including kindergarten, elementary, secondary and tertiary level students, plus special-purpose classes for organisations. These discussions confirmed the findings from the initial phase of the project, culminating in the creation of a set of on-line training materials.

Four key findings from the project will be presented relating to the teachers': 1) previous training in assessment; 2) attitudes to language testing and associated theory; 3) understanding of assessment in its broader sense; 4) understanding of 'language assessment literacy'. From this research it would seem that the gap between teachers and those who research and write about language testing is considerable. This research project sought to narrow the gap by giving teachers a stronger voice in the debate, which, in turn, may have important implications for the development of future teacher training courses.

National college entrance examination reform and the effort to raise assessment literacy of stakeholders in Japan

1:30–2:00 · Mario Laserna—ML—607

Keita Nakamura, Eiken Foundation of Japan, Japan

Educational reform efforts are observed around the world. These educational reform initiatives increasingly employ high-stakes accountability tests to improve the so-called quality of education (Chalhoub-Deville, 2016). In the case of Japan, the ministry of education is conducting a series of reforms which include the reform of the college entrance examination system.

In the reform, the ministry started endorsing the use of several standardized assessments that can measure the four skills of English. In this new system, instead of taking the traditional tests that are usually prepared by universities, test takers can choose one of the standardized tests and submit the result when they apply to universities of their choices. Each university thus needs to set the cut score of each of the tests so that test takers can check if they have reached the required score for the admission when they apply. Universities usually set the cut scores based on the reference table in which scores of each test are comparable in reference to their CEFR (Common European Framework of Reference) band.

In order to facilitate the movement to use the above new system, the ministry had held meetings where relevant stakeholders were invited to exchange ideas and express their honest opinions to each other. Those stakeholders included the representatives of high school principals, university administrators, university language testing researchers, researchers of testing organizations, and CEOs of leading companies. These two-hour public meetings took places every six months from 2014 to 2016.

As one of the participants of the meetings representing the researchers of testing organizations, the author conducted a qualitative study based on Pill and Harding (2013) on the data collected from the official minutes of the three-year meetings provided by the ministry. Of the 144,471 letters of written minutes, the author at first isolated the 69,964 letters of relevant Q&A sessions where stakeholders made comments related to the standardized tests of English, analyzing their qualitative characteristics, and finally categorizing them into similar concepts.

The results were found that those comments could be mainly categorized to 1) practical perspective, 2) validity perspective, and 3) reliability perspective. When cross-referenced the types of stakeholders with the categories of comments and questions, it was found that a) high school principals commented mainly on the practicality issues such as the test fee and the test locations, b) university administrators commented mainly on the practicality of the new system and also on the reliability of tests such as the comparability of scores from the different tests, and c) language testing researchers commented mainly on the validity of the new system such as the difference of the tested construct by different tests.

The implications would be discussed in terms of the value system behind the need of each stakeholder and how to reach for the needs of all of the stakeholders. In the presentation, the author would share the new website which was created by the participating testing organizations for the purpose of fostering the language assessment literacy of other stakeholders.

Understanding temporal fluency: The influence of the ACTFL OPIc speaking prompts on L2 speakers of Spanish

2:05–2:35 · Mario Laserna—ML—604

Gregory Thompson, Brigham Young University, USA

Troy Cox, Brigham Young University, USA

Alan Brown, University of Kentucky, USA

One of the hallmarks of proficiency testing centered on the ACTFL guidelines is the need to assess floor and ceiling performance based on the major levels of the proficiency scale. Thus, a student who is in the intermediate range alternates between responding to intermediate or "floor" level prompts in which appropriate performance is sustained and advance or "ceiling" level prompts in which there is linguistic breakdown. ACTFL has defined breakdown to include dysfluencies, grammatical errors, reverting to L1, silence, etc. However, there has been little published research validating the extent to which prompts targeting a particular communicative function characteristic of a major proficiency level achieve their desired aim, that is, to establish the floor or ceiling of the speaker's proficiency vis-à-vis the major proficiency levels. Furthermore, the relationship between measures of temporal fluency of a candidate's response to a particular prompt (speech rate, mean length of run, number of silent pauses, etc.) and her final rating has not been examined.

To this end, over 150 students with ACTFL proficiencies ranging from Intermediate to Superior in Spanish were administered oral proficiency interview computerized exams (OPIc). The OPIc is ideal for examining the impact of prompt level and type on students' responses as they are consistent across candidates. ACTFL assigns each prompt to a major level, e.g., Intermediate, Advanced, Superior and each exam consists of prompts that span at least two adjacent levels. Temporal fluency measures of the speech samples were analyzed using PRAAT to determine how examinees at each major level performed on their floor prompts (those prompts that corresponded with the examinee's final rating) and their ceiling prompts (those prompts that corresponded to the next major level). These findings have implications for test designers and those interested in using temporal fluency measures as a rough estimation of speaking proficiency.

Do workshops really work? Evaluating the effectiveness of training in language assessment literacy

2:05–2:35 · Mario Laserna—ML—606

Yan Jin, Shanghai Jiao Tong University, China

Wei Jie, Shanghai Jiao Tong University, China

Since the introduction of assessment literacy into the field of language testing and assessment, research has largely been focused on the conceptualization of the construct and the need for broadening its scope (Fulcher, 2012; Inbar-Lourie, 2008; Taylor, 2009, 2013), ignoring to some extent the training of language instructors to make informed decisions throughout the assessment process (Malone, 2008). One feasible solution is to offer language instructors with opportunities to attend professional development workshops. Though expensive and time-consuming, workshops are frequent approaches to help language instructors to supplement their formal training and improve their classroom effectiveness (Malone, 2008). The perceived effectiveness of such workshops, however, is rarely evaluated and assessed in a principled way. To this end, a study was conducted to explore a systematic approach to measuring the extent to which workshops can help participants develop their language assessment literacy.

A training program for tertiary-level English language teachers was designed and implemented by a language testing organization, which has gained rich experience in working with language teachers in test development over the past three decades. A series of workshops were conducted by 15 language testing specialists on topics of school-based language assessment, large-scale language assessment, item writing, test data analysis, rating scale development, and EAP/ESP assessment. To evaluate the effectiveness of the workshops, a quasi-experiment design (Rubin, 2000, 2008) was adopted and data were collected from a control group (non-participants of the program, n=95) and an experiment group (program participants, n=68). An initial survey indicated that the two groups were comparable in terms of the subjects' gender, age, academic background, teaching experience, and interest in training. The experiment group completed six questionnaires (S1-S6) at the beginning, the midpoint and the end of the program. Both groups were surveyed two months after the program using a questionnaire (S7) to measure their level of language assessment literacy.

Cronbach alphas indicate a satisfactory level of internal consistency (r=.69-.90). Point-biserial correlation coefficients suggest that previous training experience has an effect on participants' expectations and evaluation of the workshops. Within-group difference tests on responses to anchored questions indicate the usefulness of the workshops for the various aspects of participants' career development. Comparisons between the control and the experiment groups on their responses to assessment literacy questions (S7) provide further evidence supporting the effectiveness of the training program in improving participants' level of language assessment literacy.

In the study, a package of questionnaires have been developed and administered before, during and after the program, which has enabled evaluators to track and compare participants' responses on different components of the program. More importantly, a systematic and objective evaluation procedure has been developed for evaluators to make causal inferences about the effectiveness of the training program. The evaluation employs a quasi-experimental design with propensity score matching plus difference in differences methods to match the control and the experiment groups which are non-equivalent by nature. Future research is needed to refine the instruments and test the procedure in different contexts of training in language assessment literacy.

A language program evaluation using an argument-based approach: A case study in Chile

2:05–2:35 · Mario Laserna—ML—607

Mónica Stella Cárdenas-Claros, Pontificia Universidad Católica de Valparaíso, Chile

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

Katherine Rick, Lincoln College International, Saudi Arabia

In this presentation, we put forward a theoretically and empirically based proposal to use an argument-based approach to blended language program evaluation. Grounded in the theory of argumentation, our proposal uses a four-stage design that involves planning of an argument, gathering of evidence, presentation of the argument, and argument appraisal. In our proposed framework, we have also adapted the principles of blended language learning and program evaluation to encourage flexibility and stakeholder involvement. Accordingly, we suggest conducting focal program evaluations across any or a combination of three levels--macro-level (program administration), meso-level (departmental concerns), and micro-level (classroom-based practices)--and with a range of stakeholders. We also used blended language learning theory with a focus on key considerations such as sustainable and appropriate uses of technology, a purposeful deployment of resources, and a recognition of multimodal activities.

We have implemented the proposed framework in a range of contexts. In this presentation, we discuss an evaluation project conducted at the micro-level in an English Language Teaching (ELT) program in Chile. Focusing on multimodal practices in the classroom or the combination of approaches, materials, tools and technology into two or more ways to enhance variety, this evaluation entailed the development of an argument consisting of five inferences (i.e., domain definition, evaluation, explanation, utilization, and ramification) and underlying warrants and assumptions about blended learning in English language classes. To gather evidence to support the inferences and assumptions in the argument, we collected data from a range of stakeholders (five teachers and 41 students) and from a variety of sources (official documents, in-depth interviews, questionnaires, and classroom observations).

The evidence supporting the assumptions associated with the evaluation and explanation inferences showed that contextual factors (i.e., syllabus, lack of policy, and limited resources and structure) along with teachers' factors (i.e., lack of preparation and confidence, comfort zone, and disenchantment) and students' factors (i.e., limited use of technology for academic purposes, limited selection of tools, and overreliance on teacher's guidance) affected multimodal interactions in the classroom. We will present measures taken to appraise the argument and provide suggestions for further research.

Contextual variables in written assessment feedback in a university-level Spanish program

2:40–3:10 · Mario Laserna—ML—604

Kathryn Hill, La Trobe University, Australia

Ana María Ducasse, RMIT, Australia

A number of authors have highlighted the need for the research community to take greater account of teachers' language assessment literacy (LAL), e.g. professional experience, perspectives and knowledge (e.g., Freeman & Johnson, 1998; Leung, 2005, 2014) in context with regard to assessment in the service of learning (Carless 2011; Mcmillan 2010). This paper reports on a collaborative dialogue (Scarino, 2016) between a teacher and researcher regarding written feedback practices in a Spanish as a foreign language program at an Australian university. Starting with Hill's (forthcoming) framework, designed as a heuristic for promoting teachers' assessment literacy, the study sought to explore the teacher's 'tacit' expertise with the aim of uncovering how the researcher's theoretical knowledge and the teacher's practice might inform each other.

The study investigated the following questions:

- What does the teacher do (types of feedback)?
- What does she look for (features of writing, quality & standard)?
- How do aspects of the assessment context influence the nature and content of feedback?

Participants comprised a language assessment researcher, an 'expert' Spanish as a foreign language lecturer and 15 students from beginner (CEFR A1), intermediate (CEFR B1), and advanced (CEFR C) levels in a university-level Spanish program. Data comprised written feedback on writing tasks for each of the three levels collected over a 12 week semester as well as recordings and transcripts of discussions between teacher and researcher regarding feedback decisions. Data were analysed using Aljaafreh and Lantolf's (1994) regulatory scale and thematic content analysis. Following Turner and Purpura (2016), relevant contextual variables will include student attributes (specifically CEFR level and prior language learning experience) and institutional constraints, in addition to task and topic.

The collaboration using the framework both facilitated critical reflection on the ways in which contextual variables impacted written assessment feedback and advanced the teacher's practice.

Understanding language assessment literacy profiles of different stakeholder groups in China: The importance of contextual and experiential factors

2:40–3:10 · Mario Laserna—ML—606

Xun Yan, University of Illinois at Urbana-Champaign, USA

Jinsong Fan, Fudan University, China

Cong Zhang, Shandong University, China

The burgeoning interest in language assessment literacy (LAL) has led to widened recognition of the importance of developing LAL for all stakeholder groups of language assessments. However, recent investigations also suggest that stakeholder groups might differ in interests, needs, and expectations in assessment practice, resulting in different LAL profiles among them (e.g., Malone, 2013; Taylor, 2013; Pill & Harding, 2013). To further this line of research, this study examines assessment practice and LAL development of different stakeholder groups in China and explores contextual and experiential factors that shape their LAL profiles.

This study targeted three major language assessment stakeholder groups: language testers, language teachers, and graduate students in language studies programs. To examine the roles of contextual and experiential factors in shaping LAL profiles, semi-structured interviews were conducted to 20 participants, four participants from each of the following five subgroups: language testers, graduate students with and without a focus on language assessment, and EFL teachers working in secondary and tertiary contexts. All interview data were verbatim transcribed, coded and analyzed in NVivo (QSR, 2012), using an inductive approach. After an iterative coding process, five categories emerged: (1) assessment-related practices and background; (2) perceptions of LAL and LAL development; (3) experience and perceptions of assessment training; (4) general impression of the field of language testing and assessment; and (5) miscellaneous comments. Based on the emergent categories, LAL profiles and perceived assessment training needs for each participant were described and contrasted both within and across stakeholder groups.

Both individual and group differences were observed among the LAL profiles. At the individual level, each participant experienced different LAL development paths; their own perceived strengths and weaknesses in LAL were conditioned by the resources they had access to, and were associated with their experiences in assessment development and use and/or language testing research. At the group level, language-testing students demonstrated greater familiarity with testing terminologies and discussed a wider range of theoretical and technical topics related to LAL development. On the contrary, language teachers in general were unfamiliar with testing terminologies; however, when provided with definitions, teachers tended to be able to understand and link those concepts with specific examples from their assessment practice. As to assessment training needs, EFL teachers prioritized item writing and analysis skills over theoretical knowledge and technical skills in measurement, whereas language-testing students tended to express needs in improving statistical knowledge and research skills.

Taken together, the findings of this study suggest that assessment practice and training needs are highly contextualized and shaped by experiential factors for different stakeholder groups. These contextual and experiential factors provide important implications for assessment training, highlighting the need to tailor assessment training towards specific stakeholder group(s). As such, assessment training programs should consider the assessment context and prior assessment-related experiences of its target audience and develop realistic expectations for the content coverage and achievement goals for different stakeholders.

Internationalization and language testing in Brazil: Exploring the interface of TOEFL ITP and rankings at UFES

2:40–3:10 · Mario Laserna—ML—607

Kyria Rebeca Finardi, Federal University of Espirito Santo (UFES), Brazil

Gabriel Amorim, Federal University of Espirito Santo (UFES), Brazil

Claudia Kawachi-Furlan, Federal University of Espirito Santo (UFES), Brazil

This presentation reports results of a study (Finardi, Amorim, Kawachi-Furlan, in press) carried out to explore the interface between English language testing and internationalization indexes in a Federal University in Brazil. The study is based on the following assumptions: in the globalized world we live in, some knowledge of English is necessary to 1) expand access to online information (Finardi, Prebianca, Momm, 2013) and education (Finardi & Tyler, 2015); and 2) to drive internationalization of higher education forward (Finardi & Ortiz, 2015; Finardi & França, 2016; Finardi, Santos & Guimarães, 2016). Moreover, the study acknowledges the importance of the Brazilian government-funded program English without Borders (EwB) in the internationalization of higher education in that country. The EwB program offers three actions free of charge to the whole university community: 1) face-to-face English for specific purposes (ESP) classes, 2) an online English course (My English Online) and 3) the TOEFL ITP tests. This program with its massive testing of English within the academic community has had considerable wash-back effects in the Brazilian internationalization agenda and promotion of language policies in that country (Finardi & Archanjo, in press). Based on previous studies carried out in the Federal University of Espirito Santo (UFES)(Finardi & Ortiz, 2015; Finardi & França, 2016; Finardi, Santos & Guimarães, 2016; Amorim & Finardi, in press) which suggested that lack of proficiency in English was a serious caveat to the internationalization process in that institution, the study reported here (Finardi, Amorim & Kawachi-Furlan, in press) aimed at verifying whether proficiency levels and internationalization indexes were correlated after UFES joined the EwB program in 2013.

So as to verify this hypothesis, the study analyzed the interface between English proficiency levels measured with the TOEFL ITP test and internationalization scores measured in terms of rankings at UFES between 2013-2016. Results of the study suggest that during the period analyzed these two variables remained stable, thus confirming the hypothesis raised.

SYMPOSIA ABSTRACTS

TUESDAY, JULY 18th

Towards an ILTA policy on language assessment literacy: How to define the construct and for whom?

5:00–7:00 · Mario Laserna—ML—240, Auditorium B

Convenor:

Catherine Elder, University of Melbourne, Australia

Presenters:

Benjamin Kremmel, University of Innsbruck, Austria

Kathrin Eberharter, University of Innsbruck, Austria

Luke Harding, Lancaster University, United Kingdom

Dina Tsagari, University of Cyprus, Cyprus

Margaret E. Malone, Georgetown University, USA

Discussants: Micheline Chalhoub-Deville, University of North Carolina, Greensboro, USA

Tim McNamara, University of Melbourne, Australia

Interest in language assessment literacy (LAL) within the language testing community over the last decade has generated research and discussion on various aspects and areas of concern. Topics have included debate as to the nature of this entity, i.e., its definition and core components, the commonality or variability of the knowledge-base, as well as who can be considered qualified to disseminate it and how. An intriguing area for research and debate focuses on the relationship between the assessment knowledge, skills and principles required for making language assessment decisions "across stakeholder boundaries" as the theme of LTRC 2017 suggests. In addition and of equal importance is the degree to which assessment literacy in the area of languages can be considered unique, or whether it is indistinguishable from generic concepts of assessment literacy in terms of content and scope. Since these issues are central to the conceptualization of assessment knowledge they merit discussion and dialogue within the language testing community at large.

This two-hour symposium will bring together different perspectives on the above issues with the intention of starting a conversation amongst ILTA members and on the theoretical and practical aspects of language assessment literacy and how these might inform a LAL policy for

the organization. The first part of the symposium will consist of three papers by LT scholars with different takes on the LAL construct . The second part will begin with commentaries from two expert discussants, which will draw out the main themes and contentious issues raised by each paper and their implications for an ILTA LAL policy. The moderator will then open the floor for discussion by the audience of the issues raised.

Paper 1. Putting the 'language' into language assessment literacy

Benjamin Kremmel, Kathrin Eberharter, and Luke Harding

Language assessment literacy (LAL) has emerged as an important topic in the field of language assessment, with several studies addressing LAL requirements among language teachers, language testers, test score users and other stakeholders (see Taylor, 2013; Harding & Kremmel, 2016). Among the various suggestions for taxonomies of LAL skills and abilities, there has been little discussion thus far concerning the specific construct-related knowledge required to conceptualize, develop, administer, score and use language assessments, even though this type of construct knowledge may be crucial in newer types of assessments such as diagnostic assessment (Alderson et al., 2015) and dynamic assessment (Poehner et al., 2013). Indeed, specific knowledge of language constructs is arguably the distinguishing feature between LAL and more generic conceptualisations of assessment literacy in education.

This talk will therefore problematize the role of language in current models of LAL. It will do so by comparing prominent approaches to LAL, identifying the role of 'language' within each, and pointing to under-specification of construct knowledge in existing theories. Then, using the speaking construct as an illustrative example, scenarios based on empirical research and practitioner experience will be discussed in order to demonstrate instances where knowledge of language was crucial to LAL. The talk will conclude by arguing that understandings of language constructs should be at the heart of LAL, and suggestions for how ILTA initiatives could address this issue will be proposed.

Paper 2. The importance of contextualizing language assessment literacy

Dina Tsagari

The growth in the use of accountability systems, the influence of external frameworks and the assessment policies implemented in educational contexts have increased both the amount and the importance of' 'language assessment literacy'(LAL) (Fulcher, 2012; Harding & Kremmel, 2016; Taylor, 2013).

However, research has shown that in many educational systems across Europe, English Language Teachers (ELTs) LAL is not a straightforward matter (Tsagari, 2016). Teachers' acquisition and implementation of LAL seems to be a situated activity, located in particular contexts, each characterized by assessment practices compatible with the social and educational values and beliefs that the school's assessment culture upholds (Inbar-Lourie, 2008; 2013). Recently, studies have begun looking at LAL in particular contexts drawing attention to the intricacies in examining teachers' perceptions and knowledge about the learning and assessment they bring to the dynamic LAL acquisition process (Scarino, 2013; Xu, 2015).

The goal of this paper is to examine the notion of LAL on a constructive and interpretive epistemological basis taking into consideration the importance of 'context'. The paper will report results based on quantitative data collected via online questionnaires designed to investigate how teachers perceive and practice assessment, the types of assessment used, teachers' training needs and preferred modes of delivery of future training courses. The findings provide nuanced in-depth understanding of the specificity of LAL of ELTs which contributes to the identification of assessment priorities and the development of assessment training strategies that are contextually situated within effective modes of training.

Paper 3. Including student perspectives in language assessment literacy

Margaret E. Malone

This paper explores the concept of language assessment literacy (Inbar-Lourie, 2013; Taylor, 2009) and the centrality of including students in such efforts. Language assessment literacy refers to the knowledge, skills and understanding that users of language assessments have when interpreting the results of assessments and applying these results to inform students about their learning progress (or lack thereof) and to change, maintain or improve language teaching and learning. Although recent research focuses on the language learning classroom (Fulcher, 2012), the importance of language teachers in the assessment process (Hill and McNamara, 2012) and teacher beliefs about assessments and their uses (Cheng and Sun, 2015; Rea-Dickens, 2003), only limited research (Butler, 2016; Malone and Montee, 2014; Masters et al, 2010) investigates students' understanding of assessment, results and their own performances relative to general, course and their own learning goals.

Developing language assessment literacy requires understanding both how languages are learned and the fundamental principles of languages assessment. In describing such concepts to teachers, language testers may overemphasize the technical aspects of assessment and underplay the concrete, immediate impact for teachers (Malone, 2012). If such communication is difficult for adults, then language assessment literacy efforts for students represent still more complexity for the field.

This paper explores the challenges and opportunities inherent in extending the reach of language assessment literacy to the realm of student knowledge, skills and understanding, and includes methodologies and approaches for analysing and exploring student understanding of language assessment.

Human-machine teaming up for language assessment: The need for extending the scope of assessment literacy

3:20–5:20 · R 209

Organizers/ moderators:

Yan Jin, Shanghai Jiao Tong University, China

Tan Jin, Sun Yat-sen University, China

Presenters:

Veronica Benigno, Pearson, United Kingdom

John de Jong, Pearson, Netherlands

Lin Gu, Educational Testing Service, USA

Lili Yao, Educational Testing Service, USA

Larry Davis, Educational Testing Service, USA

Bo Zhu, IFLYTEK CO., LTD, China

Wei Wang, Sun Yat-sen University, China

Baichuan Li, Youmi Technology, China

Discussant:

Xiaoming Xi, Educational Testing Service, United States

Advancements in machine intelligence in recent decades have rapidly transformed and improved the practices of language assessment. The improvement is first evidenced in making language assessment more efficient, with the typical examples of computer-adaptive testing and automated evaluation systems (Chapelle & Douglas, 2006; Xi, 2010). Improvements have also been made through innovations to enhance methods and uses of language assessment, such as introducing new language constructs in machine-mediated contexts and increasing opportunities for learning when assessing (Chapelle & Voss, 2016). With recent progresses in the development of machines with human-like intelligence in learning, notably in understanding and producing human language, machine translation, spoken dialogue systems and machine reading (Hirschberg & Manning, 2015), it is envisaged that, in the decades to come, there is much room for machine to further improve assessment efficiency and validity.

While radically transforming the practices of language assessment, machines are "given objectives that don't take into account all the elements that humans care about" (Russell & Bohannon, 2015, p. 252). In establishing the alignment of machine intelligence with human objectives, various stakeholder groups, ranging from assessment developers to assessment users, from examiners

to test takers, from researchers to policy makers, are required to be updated with new knowledge, skills, and competencies in order to balance the benefits and risks of incorporating machines into assessment practices. The scope of assessment literacy (see Fulcher, 2012; Taylor, 2009, 2013), as a result, needs to be extended in order to address the implications of technological innovations on language assessment. Without a shared knowledge base and a good understanding of why and how humans and machines can utilize each other's strengths in the process of developing and using language assessment, the benefits are likely to be overstated by supporters of human-machine collaboration whereas the risks exaggerated by opponents.

This symposium therefore aims to discuss the need for extending the scope of assessment literacy in machine-mediated contexts by addressing critical issues regarding assessment development, implementation and scoring as well as the use of language assessments. To be specific, empirical studies focusing on the interface between human expertise and machine intelligence are to be presented and discussed from the perspectives of various stakeholders, including assessment developers, examiners and raters, and assessment users. Together these presentations will highlight research efforts that address emerging theoretical, methodological as well as ethical issues in the endeavor of integrating human expertise and machine intelligence and safeguarding against the potential pitfalls of intelligent machines.

To date, few attempts have been made to confront the challenges associated with human-machine collaboration in language testing and assessment. Therefore, we believe that the discussion at this symposium will contribute to an expansion of the scope of assessment literacy and promote a better understanding of the core knowledge, skills and competencies that underpin good quality assessment practices in machine-mediated contexts.

Paper 1. Multi-modality to assess language proficiency: Construct definition and score interpretation in PTE Academic

Veronica Benigno and John de Jong

This paper reports on test development and quality control procedures put into place by Pearson Language Testing to produce integrated-tasks in the Pearson Test of English Academic PTE Academic (PTE Academic). In the first section of the paper the investigators of the project, Dr. Veronica Benigno (Pearson, London) and Prof. John de Jong (Pearson, London and Vrije Universiteit, Amsterdam), will describe the language performance which is elicited by item types assessing skills in combination and therefore asking test takers to integrate information from more than one source text. This section will also describe the requirements set to select the item writers and the training materials (task specifications, performance samples, etc.), the procedures used to help item writers produce and review test items according to reference standards and guidelines, and the steps taken to mitigate construct under-representation and construct-irrelevance. The second section of the paper will look more into detail at how test takers are evaluated in terms of linguistic, sociolinguistic, discourse, and functional competencies. The two sections will focus on providing evidence of the validity of the inferences about the different components of the construct definition, and by consequence, on the validity of the score interpretation. This paper contributes to a greater understanding of the skills and competencies required from item writers to produce high-quality test items and calls for the need to apply rigorous quality control procedures to ensure assessment literacy among item writers.

Paper 2. Can machine-generated feedback have value in addition to human judgments for predicting speaking true scores?

Lin Gu, Lili Yao, and Larry Davis

Automated scoring technology is well-suited to producing specific and objective measurements of linguistic features of language performance. These scores can potentially be used by language learners in understanding the strengths and weaknesses of their performance. However, a common concern is that diagnostic sub-scores may provide little distinct information that is not already included in the total score (e.g., Haberman, 2008). This presentation aims at broadening the scope of assessment literacy by demonstrating how to evaluate psychometric quality, especially added value, of sub-scores. We examined machine-generated feature scores within the context of the speaking section of TOEFL Practice Online, a low-stakes practice test. Measurements of a small number of selected linguistic features were generated by SpeechRaterTM, an automated scoring system developed by Educational Testing Service. Following an analytic approach proposed by Haberman (2008), we evaluated whether each of the feature scores was reliable enough to be reported on its own and whether the selected features had value in predicting feature true scores and the human true score through a series of best linear prediction models. The analyses suggested that the feature scores indeed had value, in addition to human holistic judgments, for predicting the examinees' true speaking ability. We argue that the evaluation method demonstrated here can be used as a confirmation of technical quality for speaking testing programs that intend to report machine-generated spoken feature scores to test takers. We will also discuss how results from such an evaluation method can empower teachers and test takers to interpret and use scoring information generated by machines.

Paper 3. Writing to the machine: Challenges facing automated scoring in the College English Test in China

Yan Jin, Bo Zhu, and Wei Wang

This presentation discusses the need for extending the scope of assessment literacy to include test takers' understanding of how machines work when their performances are evaluated by an automated scoring system. The presenters will first introduce a collaborative study between a language testing organization and a high-tech company, which aims at using an automated scoring system to gradually replace human raters in the scoring of essay writing and paragraph translation (Chinese to English) in the College English Test (CET), an English language test of a very large scale and high stakes in China. Experiments using the CET operational test data have come to the conclusion that well-trained machine raters on the whole performed more consistently than human raters, although there is room for improvement in the accuracy of scoring of top-level performances. In addition to technological challenges, a major concern of the CET test developer is construct-irrelevant variances introduced by test takers' use of testwise strategies when they know that they are writing to the machine instead of human raters. The presenters will report on a large-scale survey and focus group interviews aiming at investigating CET test takers' perceptions of automated scoring, the strategies they are likely to use, and the relationship between strategy use and test performance. Results of the investigation have pointed to the need for machines to become more intelligent, and more importantly, the need to inform test takers that machines have developed counter test-taking strategies.

Paper 4. Data-driven adapting of source texts for test preparation: What can teachers learn?

Tan Jin and Baichuan Li

While there have been a considerable number of corpus-based studies informing the content of testing materials, direct explorations of corpora by teachers to adapt source texts (i.e., data-drive adapting of source texts) for test preparation remain a largely unexplored area. In this connection, this study examines the practice of language teachers in producing test preparation materials, focusing on the development of their knowledge and skills when using data-driven adapting. The data-driven adapting was realized through a text evaluation system, which was developed to level text complexity as well as to tag words and sentences for adapting purposes. Furthermore, an online course for language teachers on using the text evaluation system to adapt source texts was developed and provided. To investigate both the outcome and process of data-driven adapting, a convergent parallel design of mixed methods was employed. A quantitative study was first conducted through text analysis of the outcome to reveal differences between the original and the adapted texts. Subsequently, a qualitative study was conducted through semi-structured interviews to investigate the process the teachers employed to adapt texts. Results are discussed in terms of the effectiveness of data-driven adapting. Overall, the results provide strong evidence that teachers can learn and benefit from data-driven adapting. The study also highlights that certain knowledge and skills are needed to make best of the adapting actions by teachers. Implications are also given to further develop the assessment literacy of teachers in human-machine collaboration for producing test preparation materials.

Test preparation: A double-edged sword

3:20–5:20 · R 210

Organizer:

Liying Cheng, Queen's University, Canada

Presenters:

Dina Tsagari, University of Cyprus, Cyprus

Shahrzad Saif, Université Laval, Canada

Hong Wang, Mount Saint Vincent University, Canada

Discussant:

Elana Shohamy, Tel Aviv University, Israel

When decisions made based on language test scores are of consequence to them, teachers and students including school, university, and commercial test preparation centre administrators, tend to tailor their instructional practices and language learning to reflect the test's demands (Cheng, 2014, Green, 2006). The higher the stakes of the test, the stronger the urge to engage in test preparation practices intended to enhance test performance.

Appropriate test preparation practices may help test takers target their study and improve overall language ability: effective language learning is reflected in higher test scores. However, inappropriate test preparation may involve 'teaching to the test,' or narrowing the curriculum and learning to the content of the test only (rather than addressing the broader domain to which test scores are intended to generalize). Test preparation may also exploit construct-irrelevance: learners may find strategies for 'beating the test' and inflating their scores. Inflated test scores are of particular concern for score users such as university admissions staffs, immigration officials, and potential employers - all important stakeholders. These inflated test scores may prove problematic both for the test takers themselves and for those who work with them: university students with inflated test scores will most likely struggle in their first-year courses and may unduly strain the support services provided (e.g., Campbell & Li, 2008; Jepson, Turner, & Calway, 2002, May & Kettle, 2010).

Test preparation can thus be seen as a double-edged sword: it has the potential for both positive and negative effects on learning. Test preparation is a complex phenomenon, and takes different shapes in its nature, effects and perceived value situated in different sociocultural contexts. This symposium presents three empirical studies conducted in different countries to unpack the complex nature of test preparation, together with a concise introduction and critical discussion of this complex phenomenon from the symposium organizer and discussant. Tsagari explores multiple case studies of test preparation at the school and university level in Europe. Saif examines the outcomes and dynamics of a test preparation centre in Iran. Wang investigates the relationship between test-takers' obtained test scores and their academic achievements

at one Canadian university. These three empirical studies raise concerns about language test validity, and have implications for the assessment literacy needs of test score users. Taken as a whole, this symposium provides theoretical understanding of and methodological guidance to test preparation research.

In this 90-minute symposium, the organizer will first introduce the theoretical and methodological frameworks guiding test preparation research (10 minutes). This introduction will be followed by three empirical studies conducted in different countries, each of 15 minutes. At the end, the discussant will highlight recurring themes, and indicate future directions for studies on test preparation across stakeholder boundaries (10 minutes). The audience will have 25 minutes for questions and discussions.

Paper 1. Test preparation in high-stakes language exam contexts: A European perspective

Dina Tsagari

This paper considers test preparation from a pedagogical perspective looking at test preparation practices used from both a learning-oriented perspective, e.g., to support/enhance learning, and from a construct measurement perspective, e.g., to raise test performance at secondary or higher education in Greece and Cyprus. Motivated by constructivist and socio-cultural theories of learning (Brown, 2007; Lightbown & Spada, 2006) and the influence of Vygotsky's work (1978) in language teaching and learning, this presentation reports on four studies undertaken using a qualitative research design to explore teachers' perceptions and their relation to test face validity and classroom practices towards high-stakes tests. Teacher interviews and classroom observations were analyzed, in conjunction with Atlas.ti, using a coding scheme developed and informed by relevant literature on test washback, feedback and classroom interaction.

The results revealed a number of complexities involved in test preparation in high-stakes exam contexts, that is test preparation involves instructional, ethical, and validity issues such as the effectiveness of test preparation in raising test scores, the effect it has on test validity, the impact on learning and teaching, misunderstandings about test purpose and use, accountability issues, and lack of effective channels of communication between the examination boards and teachers. Based on the results of these studies, the paper will provide critical discussion of points derived from the case studies undertaken and make theoretical and pedagogical recommendations with the aim of unpacking the complex nature of test preparation and enhancing teachers' awareness and practices of exam preparation as part of their 'language assessment literacy' development (Inbar-Lourie, 2008; Fulcher, 2012; Taylor, 2013)

Paper 2. The learning outcomes of test preparation courses and the dynamics of the test centers

Shahrzad Saif

This presentation reports on a longitudinal study investigating the language learning outcomes of IELTS preparation courses offered by a major test preparation (TP) center in Iran. It addresses three questions: 1) Do TP activities reflect the dynamics of the test center? 2) Does TP enhance test performance as shown by increased test scores? 3) Does preparing for high-stakes tests also improve test-takers' proficiency?

Using a mixed-methods design, the study was conducted in two phases. In Phase I, qualitative data was gathered from 65 stakeholders (teachers, test center administrators, test-takers) using questionnaires, interviews, focus group interviews, and observations. Phase II examined the effects of 10-week Academic IELTS preparation courses on learners' test performance and English language proficiency. The results of the pre- and post-test analysis of the data, gathered from 201 learners in two homogeneous control (General English) and experimental (IELTS preparation) groups, revealed that both groups significantly improved their IELTS scores and proficiency test scores. However, the experimental group performed significantly better on the post-test implying that, by the end of the treatment, TP helped the IELTS group raise their test scores and, at the same time, improve their proficiency scores more than the General English group. The triangulation of results reveals that teachers' and learners' perceptions of the test content as well as context-specific factors — TP center's culture, program length, test-takers' expectations, motivation for learning English — influenced what was taught and learnt in TP classes. The implications for the validity and consequences of high-stakes test scores are discussed.

Paper 3. Exploring the relationship between test-takers' obtained test scores and their academic achievement

Hong Wang

Test preparation practices are prevalent in language testing (Alderson & Hamp-Lyons, 1996). However, there is concern that test preparation could inflate test scores: raising scores by exploiting construct-irrelevant test features without concomitant gains in English language ability (Koretz, 2005). In reality, English often is a major challenge for international students in English-speaking universities, even when they have achieved high test scores at entry. Having met the English language entrance requirements does not necessarily indicate that students are able to make the adjustments necessary to succeed in such educational system (Jepson, Turner, & Calway, 2002, May & Kettle, 2010). Therefore, the study investigated the relationship between test-takers' obtained test scores and their academic achievement in one Canadian university.

To provide "information-rich cases for study in depth" (Patton, 2002, p. 230), 151 university students from a range of countries participated in the study. The instrument was a structured questionnaire consisting of 41 statements rated on a five-point Likert scale and open-ended questions to collect and track participants' demographic data, language learning experiences, admission test scores, test preparation experiences, and academic achievements. Findings revealed that test preparation courses were valued by those that took them as well as a perception of a lack of support mechanisms for students' university studies. The study helps in providing insights into the literacy requirements of ESL students' university studies and the extent to which their scores reflected their ability to cope with the linguistic demands of studying in English-speaking universities.

FRIDAY, JULY 21st

The construct of multilingualism and language policies in language testing

8:30–10:30 · Mario Laserna—ML—240—Auditorium B

Organizers:

Jamie Schissel, University of North Carolina at Greensboro, USA

Micheline Chalhoub-Deville, University of North Carolina at Greensboro, USA

Presenters:

Elana Shohamy, Tel Aviv University, Israel

Constant Leung, King's College London, United Kingdom

Mario López-Gopar, Universidad Autónoma Benito Juárez de Oaxaca, Mexico

Nicholas Limerick, Teachers College, Columbia University, USA

Nick Saville, University of Cambridge, Cambridge English Language Assessment, UK

James R. Davis, University of North Carolina at Greensboro

Discussant:

Micheline Chalhoub-Deville, University of North Carolina at Greensboro, USA

The symposium seeks to bring together the burgeoning research in multilingualism within applied linguistics to delineate conceptualizations of multilingualism as constructs amenable for classroom and standardized testing systems operations while accounting for language policies. We acknowledge that multilingualism is not a new concept. Innovations such as translanguaging (García, 2009) or translingualism (Canagarajah, 2013), however, have garnered attention more in applied linguistics than in language testing. Previous calls for the need to address multilingualism in test constructs have been voiced through scholarship and panels examining content testing of multilingual populations (Shohamy, 2011; Solano-Flores & Li 2013) and the Council of Europe's (2007) discussions of plurilinguialism for language education policies in Europe. Yet these approaches remain too few.

Research in applied linguistics has amassed reasonable information for language testers to engage in considering an epistemological shift from a monolingual construct to multilingual langauging in language testing theories, practices, and validation for operational language tests, proficiency scales, and content tests. The symposium brings together scholars conducting quantitative and qualitative research from different regions of the world who draw from a variety of domains within applied linguistics and language testing. The projects investigate multilingual assessments used with high school students learning Hebrew as an additional language, the development of multilingual tests with and for English teachers in Oaxaca, Mexico, language proficiency testing of Indigenous languages in Ecuador, and language education policies and testing in Europe and. Across all domains, a theory of action (Chalhoub-Deville, 2015), which connects policy, design and development, decision and consequences will be highlighted.

Paper 1. Towards a deeper understanding of multilingual and translanguaging assessment

Elana Shohamy

Current trends in the language testing profession have been much more humane in research intentions, and that is manifested in a number of areas within the field. Compared to a decade ago, there are some positive movements in the field which are trying to "talk back" to contest homogenous testing methods and approaches of the past. Yet, multiple challenged need to be addressed before multilingual/translanguage tests can be designed and used. This paper presents research based on think aloud protocols of multilingual test takers while performing reading and writing subject-content tests. The sample consisted of 30 students from two high schools and included immigrants for whom Russian and French are their L-1s and Israeli Palestinians for whom Arabic is the L-1, all for whom Hebrew is a second language. The test takers were presented with multiple strategies to use to solve reading comprehension test questions. In these cases, students were encouraged to use other languages in addition to specifying how all these languages were employed and what their functions were. Similarly, a writing task required students to provide verbal reports while writing texts on three different types of genres and topics. In both situations test takers were encouraged to use as any of the languages they are proficient in. The analysis focused on the emerging patterns of multilingual and translanguaging in the two conditions. Attitude questionnaires reflecting on the experiences of using one versus two languages was also documented. Suggestions for using the findings to construct multilingual tests will be suggested based on these patterns to better reflect the construct of the multilingualism.

Paper 2. Task-based translingual assessments: A framework for classroom-based language assessments for linguistically diverse communities

Jamie L. Schissel, **Constant Leung, Mario López-Gopar, and James R. Davis**

Prioritizing a language ideology of heteroglossia that characterizes communicative competence as fluid and dynamic language practices are the norm has powerful implications for language assessment. Yet, despite calls from Shohamy (2011) and Otheguy, García, and Reid (2015) for language assessments to embrace approaches that integrate the language resources of learners, there remains little work in this direction.

We present an assessment framework that prioritizes learners' language resources and language practices within the test design for classroom-based language assessments. Weaving together theories and pedagogies of translanguaging and translingual practices with formative assessment practices and task-based language assessment, we present guiding principles for test design and preliminary findings from an ongoing project with English language teachers in Oaxaca, Mexico. As a community, the Oaxaca region is among the most linguistically and culturally diverse areas in Mexico and since 2003, the Mexican government formally recognizes 364 Indigenous languages, which, among other uses, allows for their use in educational contexts. This project broadly employs theories and research to intricately design language assessments to valorize learners' dynamic, flexible language resources and address the potentially life-long social consequences of performance on language assessments through collaboration with teachers and learners.

Paper 3. Indigenous language revitalization and the problems with testing native Quichua-speakers

Nicholas Limerick

Research on Indigenous language revitalization has shown how mainstream norms and institutions must be re-conceived if language maintenance is to be successful. This paper extends this assertion to consider the difficulties of language competency exams in Indigenous languages. Based on more than two years of ethnographic research including field notes, interviews, audio and video recordings and test results with intercultural bilingual education in Ecuador, it asks: How must one speak in Quichua in order to be declared a proficient speaker? Intercultural bilingual schoolteachers are required take an official Quichua exam in order to declared "bilingual" if they wish to teach in the system. Paradoxically, Quichua individuals sometimes fail the exam even though they have grown up speaking Quichua. This process leads to widespread discontent with Quichua exams, as well as the intercultural bilingual school system more generally, as native Quichua-speakers decry being told that they are not fluent in a language that they have always spoken. Through applying models of translanguaging to Quichua language use, this paper examines the processes through which native Quichua-speakers fail the exams. It concludes that language exams in traditionally marginalized languages like Quichua must take into account widespread linguistic diversity if they are to be successful, including the normalization of Spanish loan words in Quichua.

Paper 4. How can multilingualism be supported through language education in Europe?

Nick Saville

The language policy landscape in Europe has impacted education over the past 25 years, in particular, the policies of the Council of Europe (Strasbourg) and the European Commission (Brussels), as well as the work of the Association of Language Testers in Europe. The Council of Europe is responsible for the development of the Common European Framework of Reference and its underlying scale of proficiency was adopted by the European Union (EU) in an attempt to set a European Indicator of Language Competence. Results of two wide scale surveys commissioned by the EU to provide participating countries with comparative data on foreign language competence and insights into good practice in language learning will be discussed.

One conclusion is that despite many policy initiatives at European, regional and national levels, it has proved difficult to find an appropriate theory of action for developing relevant multilingual language skills through formal education. Part of the challenge is a difficulty in aligning macro and micro aspects of policy. Both the EU and the Council of Europe operate under the principle of subsidiarity in the area of education. This means that, while recommendations are made at a transnational level, education policies are devolved to Members States and are implemented by national education authorities. In order to address this problem, language professionals need to reimagine the 'ecology of language learning' as a theory of action to guide the development and implementation of more effective multilingual policies.

FRIDAY, JULY 21st

Assessing the academic literacy of university students through post-admission assessments

8:30–10:30 · Mario Laserna—ML—346, Auditorium A, "upstairs"

John Read, University of Auckland, New Zealand

April Ginther, Purdue University, USA

Slobodanka Dimova, University of Copenhagen, Denmark

Sophie Swerts Knudsen, University of Copenhagen, Denmark

Pia Osorio, Universidad del Norte, Colombia

Albert Weideman, University of the Free State, South Africa

With the worldwide growth in English-medium university education, institutions in many countries need to come to terms with the linguistic diversity of their student populations and the fact that many of their incoming students who have English as an additional language are not well equipped to meet the language demands of their degree studies in English. In fact, similar concerns are arising in universities where the medium of instruction is the students' own first language. Thus, it is timely to consider how language assessments of various kinds can help to identify students who should enhance their academic literacy and to diagnose their areas of weakness.

The assumption here is that the major English proficiency tests which are used to control the entry of international students to degree programmes ought to be complemented by post-admission (or post-entry) language assessments designed to meet the needs of the student population at particular institutions. Despite the homogenising influences generated by internationalization in higher education, universities vary in terms of the language and cultural backgrounds of their students, the extent to which the medium of instruction is the students' first language, and the adequacy of the students' preparation for academic study in their secondary education. Thus, local assessment initiatives are to be encouraged, and the proposed symposium takes a broad view of the situation, with papers from four different countries (and indeed four continents).

The first two papers look specifically at academic reading, a somewhat "hidden" skill that is less directly observable by university instructors than writing or speaking, and yet a lack of L2 reading ability can significantly hamper students in their degree studies. Paper 1 discusses the use of timed oral reading at a US university as a diagnostic assessment and an intervention tool to improve reading fluency, so that students can more adequately cope with the required reading in their courses. Then Paper 2 reports on a study of the reading abilities of graduate students at a Nordic university, employing both a formal reading test and interviews with the students and their professors. The results help to make the case for a post-admission reading test to assess whether students can access source texts at a level which meets the expectations of their academic programmes.

The third paper takes a wider view of the academic literacy needs of students entering a university in Colombia, covering both reading and writing in L1 and L2. The paper presents an ambitious initiative involving students across the whole institution. In this case, post-entry assessments will provide a basis for monitoring the development of the students' literacy skills through their first year of study.

Finally, growing out of the author's experience with academic literacy tests in South Africa, Paper 4 addresses the concern that local post-admission assessments may not be adequately conceptualized and operationalized if they focus on individual language skills. The paper discusses a theoretically defensible model of test design which considers the practical development of appropriate assessment tasks in relation to relevant theory in applied linguistics.

Paper 1. Developing L2 oral reading fluency to enhance academic literacy

April Ginther

A great challenge for incoming international students is the amount of required reading associated with entry-level university courses. Students report being overwhelmed by assigned readings and call on peers who share their L1 for both academic and social support (i.e., Chinese students rely on other Chinese speakers). While within-group support has many positives, it is likely to diminish opportunities for practice and exposure to the L2 at this critical entry point.

Timed oral reading is a widely used method to assess reading fluency in L1 instructional contexts; fluency is typically operationalized as the ability to read aloud accurately with appropriate expression (prosody, phrasing), at a speed such that the oral reading aligns with the meaning of the text. Fluency is key because it provides evidence that the reader understands what is being read. Furthermore, oral reading practice has been found to positively affect performance on free speaking tasks. Nevertheless, in L2 instructional contexts, the development of reading fluency and the use of timed oral reading for assessment (or instruction) is almost nonexistent.

This study investigated timed oral reading as both an assessment and intervention method for evaluating and improving L2 reading fluency. A pre- and post-test administered to 77 international students enrolled in an English for Academic Purposes course was found to demonstrate a statistically significant improvement, with a very large effect size, in reading speed. This presentation will discuss why development of oral reading fluency is a particularly promising approach to the development of matriculated students' academic literacy.

Paper 2. Slipping through the cracks: Students' academic literacy in EMI

Slobodanka Dimova and Sophie Swerts Knudsen

Nordic universities have noted increased uses of English textbooks, not only in the growing English-medium instruction (EMI) programs, but also in the programs taught in the local languages (Haberland & Risager, 2008). Concerns have been raised regarding these tendencies as research suggests that reading and learning from texts in L2 requires more time and effort than in L1 (Hellekjær, 2009; Mezek, 2013). Due to these difficulties, only 30% of students reportedly complete their reading assignments while many rely on other resources, like lecture notes and PowerPoint slides (Pecorari et al., 2012). Despite these trends, domestic students are not screened for academic English literacy because university admission requires only high-school English.

Given the contextual differences of EMI in non-Anglophone, as compared to that in Anglophone countries, the present study was guided by questions related to (1) the reading proficiency levels of incoming domestic students and (2) the characteristics of the literacy construct in the EMI context. For that purpose, the reading section of IELTS was administered to 159 incoming graduate students at a Nordic university, followed by interviews with students and professors regarding their reading practices. Results suggested that graduate students exhibited a wide range of proficiency levels, 27% of which were below 6.5 on IELTS. While students knew skimming and scanning, they experienced difficulties identifying essential information and inferencing. While professors reported application of activities that facilitated reading, these activities could support only higher-level students. The implications of these findings suggested implementation of post-entry reading tests based on local expectations for academic success.

Paper 3. L1 Academic literacy in Colombia: Administrative and academic issues

Pia Osorio

In recent years, the development of students' proficiency in a second language has received a great deal of attention in Colombia. However, within this timeframe, the development of academic literacy in students' L1 has also emerged as in need of focused and sustained attention. Despite the best efforts of the local K-11 Ministry of Education, incoming university students still experience difficulties in adjusting to the style and amount of L1 reading and the quality of writing that is required and expected at the undergraduate level. In response, the Language Department at Universidad del Norte is leading a university-wide initiative intended to provide students with necessary reading and writing skills in the first year of instruction that will support not only the development of academic literacy in the L1 but also L2 proficiency, their disciplinary academic paths, and by extension, their professional lives.

This presentation summarizes the experience of an L2 focused extension branch of the University on becoming an academic division that includes both L1 and L2 academic literacy responsibilities. The narrative will focus on the conception and development of academic literacy courses and associated assessment procedures, specifically, the construction of post-entry assessments for baseline evaluation and to track students' progress in the development of academic literacy skills in their first year. These development efforts are embedded in recent policies and initiatives by Ministry of Education and the University which are intended to produce positive results in the overall quality of instructional outcomes across the University.

Paper 4. A skills-neutral approach to academic literacy assessment

Albert Weideman

This paper will address a concern with post-admission assessments of academic literacy that take a skills-based view of language ability. It will build on the experience at four South African universities with the design of assessments of academic literacy over the past 12 years. It will argue that if the test designer views the ability to use language for academic purposes not as composed of separate and supposedly separable skills, but as an interactive and functional competence to achieve certain goals, there will be design gains. Examples will be given of how a definition of academic literacy may be more adequately conceptualized, thus contributing to the theoretical

defensibility of the design as a valid measure of academic literacy. The paper demonstrates how the test construct can subsequently be operationalized in the development of appropriate assessment tasks. It presents a model of test design as a process that has five distinct phases, of which the last two are iterative, providing an opportunity for further refinement. Not only does a skills-neutral approach offer gains in respect of the theoretical defensibility of the assessment instrument: data will be presented to demonstrate the consistency or reliability of the results obtained, and how empirical measures contribute further to the responsible interpretation of test results, as well as to the utility of the test in respect of diagnostic information. The paper will conclude by showing how the design principles of reliability, theoretical defensibility, interpretability and usefulness relate, with others, to a theory of applied linguistics.

WORKS-IN-PROGRESS ABSTRACTS

THURSDAY, JULY 20th

1. Exploring the possibility of using integrated assessments of science and language

2:00–3:30 · Mario Laserna—2nd floor—Calle del Saber

Lorena Llosa, New York University, USA

Scott Grapin, New York University, USA

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

The purpose of this study is to examine whether NGSS-aligned assessment tasks can be used to assess English learners' (ELs) language proficiency in addition to their ability to engage in science three-dimensional learning. Such "integrated assessments of language and content" would present several advantages over traditional English language proficiency (ELP) assessments: Language proficiency would be assessed in the context of an actual content area task. This would address the lack of correspondence between the ways in which traditional ELP tasks that aim to assess the language of the content areas engage language and the ways in which the content areas actually engage language. From a practicality perspective, integrated assessments could cut down on the amount of testing that students in school must do. Finally, these types of tasks could allow for a better understanding of how science and language proficiency develop and support each other. Drawing from data from the Science and Integrated Learning (SAIL) project, the study will address the following questions:

- To what extent can responses to NGSS-aligned science assessments be used to assess both science and language proficiency?
- At what level of students' English proficiency can these tasks be used to assess both science and language proficiency?

One hundred students, ELs and non-ELs, in five fifth grade classrooms in one district are participating in the SAIL project this academic year. The goal of the SAIL project is to develop NGSS-aligned fifth grade science instructional materials including end-of-unit assessments.

Students will complete several assessment tasks for two instructional units currently being field-tested. In order to address the research questions, first, we will score the responses to the tasks in terms of science since that is the main construct these tasks are designed to measure. Second, we will score the responses in terms of language using a draft rubric developed for this study. We will then examine the language scores and the characteristics of the language of the responses and the extent to which they differ based on students' science scores, their classification as EL, former EL, or non-EL, and EL's level of English proficiency. During the work-in-progress session, we will share the integrated tasks and samples responses produced by students at different levels of science and language proficiency. We will seek feedback on the draft language rubric and discuss challenges that emerge in scoring the responses and interpreting the findings.

2. Assessing young learners of English: Principles and challenges

2:00–3:30 · Mario Laserna—2nd floor—Calle del Saber

Veronica Benigno, Pearson UK, United Kingdom

The CEFR (Council of Europe, 2001) has been used in Europe and beyond for the purposes of standard-setting and curriculum development. Whilst the CEFR set out to address different domains (public, personal, educational, occupational), it largely addresses the needs of adult learners focused on general English. Adopting a constructive yet critical approach to the CEFR, this paper reports on two parallel and ongoing projects to identify and systematically describe the linguistic needs of young learners of English.

The first project is a large-scale project to create CEFR-inspired descriptors of young learner proficiency. The main goal of this project is to extend and make the framework relevant in the educational domain of primary and lower secondary learners of English. Using the same rigorous procedures applied by North (2000) to develop the original framework, we created a set of 410 descriptors for young learners covering the levels from below A1 to high B1 and targeting the age range six to fourteen. In a first stage the descriptors were created and rated by over 1,000 teachers to young learners, ELT authors and language experts from around the world. In a second stage the ratings were scaled through IRT analysis and the descriptors were aligned to a fine-grained scale of English which is directly correlated to the CEFR but allows the observation and the reporting of smaller increments of proficiency, within each of the six CEFR levels.

The second project aims to set vocabulary requirements for young learners at different levels of proficiency making use of methods drawn from corpus linguistics. A meaning-based and CEFR-aligned vocabulary database was developed by combining quantitative and qualitative analysis: frequency values and usefulness ratings (by teachers) of around 10,000 word meanings were combined into a weighted measure to rank word meanings according to their relative importance. A number of considerations related to the young learners' L1 and cultural background and topical and experiential use of language were taken into account to develop the database.

We believe this paper represents a contribution to the on-going debate on what young learners can do and what instruments can be used to assess their performance, while taking into account crucial factors affecting their learning experience, such as age, cognitive development, and linguistic, cultural and educational background. In defining YL-specific descriptors and vocabulary goals, we examine the peculiar language needs that arise in this domain, helping

to more accurately define the construct of English of young learners and to identify the main principles and challenges which should be considered in this particular context.

3. Towards a text-based framework for academic Spanish reading and writing assessment

2:00–3:30· Mario Laserna—2nd floor—Calle del Saber

Jesús Guerra, Universidad del Norte, Colombia

Nayibe Rosado, Universidad del Norte, Colombia

This presentation reports on recent developments in the construction of a framework for the assessment of academic reading and writing in Academic Spanish undertaken by the Spanish Department at Universidad del Norte, in Barranquilla, Colombia. A central consideration in this framework has been the foregrounding of text as the principal domain for the interpretation of reading and writing performance, as opposed to more mentalistic construct definitions which describe reading and writing mostly as the command of discrete cognitive skills. The theory of language underlying this framework comes from Systemic Functional Linguistics (SFL), a linguistic tradition which views language as a multilayered system of social semiosis which is constitutive of thought and which shapes (and is shaped by) social interaction (Halliday, 1978).

Concerning reading, a matrix of reading outcomes was constructed by crisscrossing SFL metafunctions (ideational, interpersonal and textual) with dimensions proposed by PISA and Saber, namely: access and retrieving, interpretation/integration and reflection/evaluation. The latter was intended to connect internal reading measurement with recognized external benchmarks. The matrix currently provides the basis for the design of criterion-referenced tests with diagnostic, formative and program evaluation purposes.

As regards writing assessment, challenges have been encountered in defining developmentally appropriate measurement criteria which validly assess freshmen's writing ability (without unduly assuming levels of performance expected of later stages of undergraduate education). An avenue of enquiry has been to analyze textual output across the curriculum in search of developmental patterns which could inform assessment at different stages of undergraduate education. An initial scheme for developmental progression in writing ability has been produced and is currently in the process of being validated by staff and external advisory. Another strategic line has been the analysis of genre-based rubrics currently in use with the aim of grading descriptors into more realistic performance indicators and defining these descriptors more in terms of textually observable features which promote inter-rater reliability.

Key highlights from this inquiry are its innovative approach to reading and writing construct definition from linguistically informed dimensions, its nature as a teacher-led form of reflection on learning and assessment and its concern with striking a balance between internal conceptions of language and learning and those implicit in external assessment frames.

4. Writing assessment training and Mexican EFL university teachers: A study of training impact

2:00–3:30 · Mario Laserna—2nd floor—Calle del Saber

Elsa Fernanda González, Universidad Autónoma de Tamaulipas/ University of Southampton, Mexico

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, 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 & 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 or practical knowledge to assess writing in their classrooms. Thus the need for teachers to be assessment literate is crucial. However, the benefit of writing assessment training and its actual impact in the assessment of writing is still unclear.

The present study has 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, two semi-structured interviews to the eleven teachers, and journal entries of four of these eleven teachers. Data obtained from the interview transcripts and the journal entries suggest that training had positive impact in three main areas: a) classroom teaching of writing, b) classroom assessment of writing and c) teachers´ self-awareness. 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. 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 reflected on the importance that writing should have in their assessment process and for students´ language development. A Writing Assessment Training Impact Inventory is proposed so as to classify the impact of training in EFL teachers. The presentation finalizes with a discussion of possible research implications for EFL classroom assessment as well as further research plans for the project.

5. Preservice teachers' language assessment literacy development

2:00–3:30 · Mario Laserna—2nd floor—Calle del Saber

Erika Restrepo, Fundación Universitaria Luis Amigó, Colombia

Diana Jaramillo, Fundación Universitaria Luis Amigó, Colombia

Although assessment plays an important role in any teaching and learning process, research has evidenced lack of teachers' Language Assessment Literacy (LAL) in Colombian contexts. For instance, Arias and Maturana (2005) found lack of knowledge and skills among teachers from two universities regarding appropriate implementation of assessment and testing, clear distinction between formative and summative purposes, precise definition of the linguistic construct, and deliberation of assessment tasks qualities in French and English courses. On the other hand, Herrera and Macías (2015) claim that teachers' gaps in language assessment literacy can relate to deficiencies in both teacher education and professional development programs in this context. Consequently, there is a need for appropriate development of teachers' LAL. As a teacher in charge of the assessment and testing course, and as the director of the English teaching program offered at Funlam college, in Medellín, we are really concerned about how preservice teachers develop language assessment literacy conducive to sound assessment practices in their classrooms. Therefore, we feel strongly motivated to plan and develop the current study.

We plan to conduct a qualitative case study embedded in an interpretive paradigm. Participants will be preservice teachers enrolled in the assessment and testing course offered as part of their training in the English teaching program at Funlam. Data will be collected through practical and personal documents (assessment and testing course syllabus and students' personal reflection notepads), focus group interviews and observation field notes. Data will be analyzed by using a thematic analysis method. Validity will be ensured by providing a detailed account of the focus of the study, the researchers' role, participants' position, basis for their selection and the context from which data will be collected, the triangulation of data and peer examination. The results will be presented in descriptive. We will answer two research questions: How do preservice teachers from the English teaching program at Funlam develop Language Assessment Literacy? How are the theories of Language Assessment Literacy communicated to the preservice teachers?

We propose two general objectives: To analyze how preservice teachers from the English teaching program at Funlam develop Language Assessment Literacy, and to establish a relationship between teacher educator's practices and the areas of Language Assessment Literacy developed in preservice teachers during the course. Three specific objectives are stated to achieve the purpose of the study: To contrast preservice teachers' understanding of key concepts related to language assessment literacy during initial and final stages in the course, to examine preservice teachers' ability to design procedures for language assessment and their ability to critically evaluate existing ones in a classroom-based assessment context, and to determine preservice teachers' perceptions about the content and methodology proposed in the course.

6. Judgment and decision-making in rating speaking: Exploring the role of cognitive attributes

2:00–3:30 · Mario Laserna—2nd floor—Calle del Saber

Kathrin Eberharter, University of Innsbruck, Austria; Lancaster University, United Kingdom

This study examines the role of cognitive attributes in the judgement and decision-making of pre-service teachers of English as a foreign language when rating speaking. As they interpret and apply a test's rating criteria, the raters become the arbiters of the test construct as expressed in the scores they award (McNamara, 1996). However, rater judgment is highly complex (Alderson, Clapham & Wall, 1995), influenced by numerous factors (Lumley, 2002) and a potential source of construct-irrelevant variance (Bachman & Palmer, 1996). In order to reduce such variance caused by the complex and often unpredictable 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). Nonetheless, rater cognition research is yet to engage in a fundamental way with the connected fields of judgment and decision-making research, despite clear synergies between these fields. Findings from cognitive psychology suggest that complex decision-making tasks are greatly influenced by a whole range of factors including processing capacities, perception, the interplay of deliberate and automated thinking, as well as metacognitive control (Newell & Bröder, 2008). Exploring these may prove useful with respect to the complex task of assessing second-language speech in real time.

This work-in-progress presentation will outline a mixed-methods study designed to investigate the cognitive processes involved in rating spoken English. In the context of a general English school-leaving examination that neither monitors nor systematically trains the teachers/raters involved, this study sets out to answer the following questions: Do cognitive attributes like experience, decision-making style and preferred cognitive style have an influence on rater accuracy and consistency? How do raters with different cognitive attributes use an analytic rating scale? During a first phase, 60 pre-service teachers will rate a set of speaking performances and then complete two self-report questionnaires on their decision making style (Scott and Bruce, 1995) and cognitive style (Pacini and Epstein, 1999). Multiple regression analysis will aim at establishing associations between measures of rater consistency and accuracy with measures of cognitive attributes. A second phase will then investigate features of rating behaviour in participants sampled from the first phase via verbal reports and eye-tracking. This study seeks to contribute to the field's understanding of the role of cognitive processes involved in rater judgment and how different raters might be managing these processes. This presentation will present preliminary results from the piloting and first round of data collection.

7. The assessment literacy of Brazilian language teachers

2:00–3:30 · Mario Laserna—2nd floor—Calle del Saber

Gladys Quevedo-Camargo, University of Brasília, Brazil

Matilde Scaramucci, University of Campinas, Brazil

Although recent and still under construction, the concept of assessment literacy (Stiggins, 1991, 1995; Fulcher, 2012; Malone, 2013; Inbar-Lourie, 2008, 2013; Taylor, 2009, 2013; Scaramucci, 2016) has become crucial for the development of (language) teachers working in different teaching contexts all over the world. According to Stiggins (2002), the concept refers to the understanding of principles that govern solid, well-informed and fair assessment practices. Despite the increasing importance assessment issues have had in the Brazilian educational system over the last years, there are no studies to date to gather information on how much assessment literacy Brazilian language teachers have and what the nature of this knowledge is. This mixed-method study aims at gathering data on this matter in order to inform teacher development policies and other initiatives. Questionnaires, interviews and document analyses will be used. Three are the specific objectives of the study: (a) to carry out an extensive literature review to understand how the concept has evolved since it first appeared in 1991, which its main aspects are and how the studies proposed for investigating the concept have been conducted; (b) to investigate which aspects are present in the Brazilian pre-service language teacher education; and (c) to investigate which aspects have not yet been taken care of and, consequently, which gaps in the Brazilian in-service language teacher education need to be filled. As this is an ongoing research, only the data related to the first two objectives will be presented and discussed.

8. China's Standards of English: The uses of English speaking activities in China

2:00–3:30 · Mario Laserna—2nd floor—Calle del Saber

Jun Wang, Shanghai Jiao Tong University, China

Xiaoyi Zhang, Shanghai Jiao Tong University, China

In the project of constructing China's Standards of English (http://cse.neea.edu.cn), within the speaking sector, developing Scales for Speaking Activities (SSA) is a rather important part. SSA will include not only the typical types of English speaking activities from various contexts, but detailed descriptions of how well the activities are performed by learners with different levels of proficiency using easy-to-understand descriptors. To achieve such purpose and provide empirical support for scale validation, a needs analysis to describe the Target Language Use (TLU) domain, identify typical speaking activities and evaluate the representativeness of them is one of the essential prerequisites for developing SSA.

Same as in the Common European Framework of Reference for Languages (Council of Europe, 2001), language activities in this study are categorized according to domain, but the domain classification is adjusted into educational and non-educational ones basing on China's education-dominated foreign language use situation. The educational domain is concerned with pedagogical activities within teaching, learning and testing contexts, and the non-educational one particularly embraces real-life activities within overseas-living and foreign-related working contexts.

This study comprises two research phases. The first phase aims to generate an initial list of typical English-speaking activities from stakeholders who are most familiar with the TLU domain and materials with high professional recognition. Specifically, 9 small-scale teacher workshops and a series of document analyses were conducted. The teacher workshops invited 44 Chinese English teachers teaching from elementary to tertiary level, and both questionnaires and interviews were applied to collect not only the commonly used speaking activities in the classroom but also teachers' descriptions of how these activities were actually used. Document analyses consist of two parts: textbook and test analyses involving 35 English teaching textbooks, 36 English speaking tests and 13 speaking test syllabuses. The results of these formed an initial list of typical speaking activities which was then applied into questionnaire design in the next phase.

Second research phase attempts to supplement the initial list of speaking activities and evaluate these activities' representativeness through two large-scale survey investigations for educational and non-educational domain respectively. Three demographic groups of participants were invited, namely teachers teaching English from elementary to tertiary level, people with overseas living experiences and employees who often need to deal with foreign-related affairs. Overall the study collected 468 valid questionnaires. Both descriptive and inferential statistics were used in data analyses. Results merit in classifying speaking activities according to their frequency of real-life applications, perceived importance to educational or non-educational purposes and task difficulty estimates aligning with teaching grades. Basing on these results, credible decisions could be made regarding the following steps of SSA development, such as selecting suitable activities for scale construction and collecting or writing performance descriptors for different levels. In short, this study reveals considerably what and how English-speaking activities are used in China. Essential information is gathered for the development of SSA, which provides strong evidential support for scale validation and enriches empirical explorations of needs analysis research.

9. Measuring heritage language learners' proficiency: Validating the Korean C-test for research purposes

2:00–3:30 · Mario Laserna—2nd floor—Calle del Saber

Young-A Son, Georgetown University, United States

Heritage language learners (HLLs) have become a focus of increasing interest in research on language acquisition and assessment, especially in the US. This interest is partly due to an influx of migrants to the US which, as multiple scholars (e.g., Leeman & King, 2015; Kagan & Dillon, 2012) note, has contributed to higher enrollment rates in some foreign language programs. Research in this domain, however, has been challenged by the lack of consistent conceptualization of HL proficiency, and considerable variability in assumptions about the way HL proficiency might best be operationalized in assessments. As Norris and Ortega (2012) have pointed out for research in SLA more generally, this variability across studies has hindered the systematic accumulation of knowledge and the generalizability of results. In the case of HL research, the lack of tools to measure language proficiency consistently has added another layer of complexity in the comparability of findings across studies.

Although efforts to address these critical gaps in SLA research have motivated the development of shortcut measures that assess general language proficiency, these tests have yet to be evaluated for their use in assessing HLLs. A large body of academic work has found that C-tests, for example,

are practical in providing a quick measurement of language learners' global proficiency (Eckes & Grotjahn, 2006), which is useful for research purposes and may provide a solution to lack of generalizability of research findings. Validation studies on this shortcut measure, however, have mainly focused on adult Foreign Language Learners (FLLs). Therefore, more research is needed to focus on how these measures behave with different learner groups. That is, studies should carefully evaluate the extent to which assessments such as the C-Test elicit similar or different language abilities from these distinct learner groups.

Accordingly, this study seeks to examine the development of an innovative Korean C-test and validate its use for assessing two different learner populations, that of HLLs and FLLs. The validation study evaluates both decisions made during assessment development and the interpretation of C-test results for the intended use of providing a measure of Korean learners' general language proficiency for research purposes. It follows Kane's (2006, 2011, 2013) Argument-Based Approach to validity by (1) sketching out the interpretive argument chain based on the proposed interpretation and uses of the test, and (2) developing a validation argument that evaluates the coherence of such interpretations by providing relevant evidence to support them. The study examines six inferences, namely, authenticity, evaluation, generalization, explanation, extrapolation, and utilization. Each of the inferences are presented with warrants and evaluation questions. Furthermore, the study proposes a methodology to gather evidence needed as backing for the warrants, which involves considerations during item and test form development, as well as the administration of the Korean C-test together with four criterion measures, namely listening, reading, speaking, and writing tests, to 100 Korean language learners, both HLLs and FLLs. The analyses of the results will be discussed in terms of how they can provide the evidence to back each warrant.

10. Stakeholders' voices: Alignment or discrepancy between proposed and perceived inferences based on English test scores in international business contexts

2:00–3:30 · Mario Laserna—2nd floor—Calle del Saber

Gwan-Hyeok Im, Queen's University, Canada

International English test scores are instrumental in determining an employee's capability of working in global business contexts. The Test of English for International Communication (TOEIC), designed by Educational Testing Service (ETS) in the United States, measures listening, reading, speaking, and writing skills of individuals whose first language is not English in international business contexts (ETS, 2015). For over 30 years, TOEIC scores in South Korea have played an important role in determining an employee's English proficiency for international business. However, as the test is designed and referenced upon the norms of only Standard English (Jenkins & Leung, 2014), concerns about the accuracy of decisions based on TOEIC scores have been raised within international business contexts where English is used as a communicative medium for speakers of different first languages, i.e., English as a lingua franca (ELF) context (Seidlhofer, 2011). However, there are limited empirical studies that address the issue of how well such English language communication is represented by the TOEIC (Im & Cheng, in press).

Messick (1989) points out that validity does not pertain to a test itself, but to the appropriateness of inferences derived from test scores. This calls for investigating how the intended and perceived

inferences are aligned, because misalignment may lead to inaccurate decisions about test takers. Therefore, this study investigates how intended and perceived inferences related to the TOEIC in the ELF context are aligned. Guided by Chapelle et al.'s (2008) argument-based validation approach, this study will not only lay out inferences intended by ETS, but also collect and analyze data regarding the intended inferences and synthesize the findings about the actual perceptions of stakeholders.

A multi-method design consisting of two phases will be used. In Phase 1, documents published by ETS will be collected to lay out intended inferences about English proficiency and test use for the TOEIC. In Phase 2, ten employers involved in hiring and promotional decisions at international companies in South Korea will participate in interviews. Along with this, a questionnaire will be administered to 300 employees at the same international companies who have a minimum of three-years work experience and who use English to communicate with those who have different first languages and native English speakers in the workplace.

The document data collected in phase one will be analyzed deductively using content analysis (Hsieh & Shannon, 2005) to organize information into categories related to inferences based on the validation framework (Chapelle, 2008). Interview data will be analyzed using thematic analysis (Braun & Clarke, 2006). Quantitative data from the questionnaire in phase two will be analyzed using an exploratory factor analysis to reduce the number of variables. Following this, an analysis of variance will check for any significant differences between groups by gender, TOEIC scores, and duration of overseas work experience.

It is expected that this study contributes much needed empirical evidence from stakeholders regarding inferences about score interpretation and use of the TOEIC.

11. Towards a principled approach to the design of rater training and norming protocols

2:00–3:30 · Mario Laserna—2nd floor—Calle del Saber

Daniel J. Reed, Michigan State University, USA

Koen van Gorp, Michigan State University, USA

Aaron Ohlrogge, Michigan State University, USA

Heekyoung Kim, Michigan State University, USA

Dan Isbell, Michigan State University, USA

Jessica Fox, Michigan State University, USA

It is widely recognized that both initial rater training and subsequent norming before actual rating occasions are crucially important to the validity and success of operational testing programs and research programs that employ judges (cf. Elder et al., 2007; Knoch et al., 2007; Knoch et al. 2016). However, the theoretical and empirical underpinnings of particular rater-training designs and the details of related activities (e.g., norming or recalibration, monitoring during operational rating) have not been clearly established in the fields of language testing and second language studies (research). Fundamental questions include what constitutes adequate familiarization with rubrics and scales, how many samples are needed at various scale levels, how raters should

be assessed in order to qualify for a particular rating assignment, how they should be monitored, and what feedback should be given to them. Additionally, important questions include how the benchmark samples should be chosen, who selects and rates these benchmarks and additional "calibrated" sets, whether it is better to use samples from the center of a band or ones on the fringes (borderline cases) or a mix of the two, how much of a justification (explicit statement of an assigned rating) is needed, how much "discussion" would help or hinder the process, and whether it can be done online as effectively as in person (Wolfe et al, 2010).

With the goal of developing a principled approach to the design of rater training and norming protocols (materials and procedures), the testing team in an English program for international students at a major mid-western university undertook a series of steps. The first step was to conduct a literature review in order to survey what current practices have been documented. The investigators were interested in both theory and practices and so searched both scholarly journals as well as documents such as the APA/NCME/AERA Standards, state assessment reports, and test manuals. The next step was to conduct a survey and follow-up interviews with key individuals involved in rater training for major test providers. In addition, the investigators examined the details of the rater training and norming protocols for the English exams at their own institution as well as data from the administration of several thousand English exams administered in Greece (The Michigan State University Certificate of English Language Competency and the Michigan State University Certificate of English Language Proficiency). Research questions included whether current training practices resulted in acceptable levels of agreement, whether agreement was stronger at particular scale levels or for particular test levels (e.g., a B2 versus C2 essay exam). The investigation culminated in a framework for guiding research and practice, with indications of parts of the framework that needed additional work from a principled or theoretical basis as well as additional backing data. The results of this work will be presented along with a recommendation for allotting an increased role for e-learning systems in training examiners and raters.

12. A corpus-driven receptive test of collocational knowledge

2:00–3:30 · Mario Laserna—2nd floor—Calle del Saber

Ivy Chen, The University of Melbourne, Australia

As collocations are essential and ubiquitous in all genres of English, learners (and even native speakers) find these often seemingly arbitrary word combinations especially difficult. When used incorrectly, collocations not only negatively affect fluency but also mark speakers otherwise highly proficient in English as non-native. Unfortunately, no validated test of collocations exists in L2 testing (Webb & Sasao, 2013). Tests found in the literature suffer from at least one of the following drawbacks (a) a small number of items, (b) unsystematic selection of items, (c) inclusion of only one collocation type (usually restricted verb-noun), (d) unreported or unsystematic test specifications, (e) unreported reliability information and (f) if a corpus was used, there was no comparison with other corpora and often item selection was based on the frequency of individual words rather than that of the collocations. The current work-in-progress aims to to create a corpus-driven and reliable receptive test of the knowledge of the word parts (i.e. what words make up the collocation) of high-frequency collocations in English (to be used with adult learners from intermediate proficiency onwards) by taking into account how the different properties of

collocations (e.g. frequency) affect lexical processing, and in turn, item difficulty; and by using Purpura, Brown, and Schoonen's (2015) up-to-date and language-specific version of Kane's (2006) argument-based approach to validity. Participants will primarily be L1 Mandarin speakers (one group living in Taiwan, another group of students studying English in Australia in preparation for tertiary study in Australia and another group of students studying at an Australian university), with other smaller groups of participants with other L1s and a group of native Australian English speakers. Test development will include the exploration of the natural distribution of collocations and a hypothesized model of collocation properties, which relates to the domain description. This model includes five under-researched properties of collocations: (a) phonology (length in number of syllables); (b) collocation type on a functional continuum (optional and unpredictable grammatical collocations to lexical collocations to fixed pragmatic collocations linked to particular social events); (c) frequency (mediated by genre); (d) degree of coherence; and (e) semantic transparency. Other issues addressed as part of the validation process include: effect of test mode (online vs. pen-and-paper); effect of question format (2 formats compared); calculation of item facility and discrimination; comparison between participants of differing general English proficiency levels and backgrounds (second-language vs. foreign language learning experience, different L1s); and correlations with criterion measures of general English proficiency, spoken English fluency, knowledge of routine formulae. This work-in-progress will check the internal structure of the test to see whether the assumption of a single construct is supported when a larger range of collocations are incorporated into a test or whether collocations seem to differ widely by where they lie on the functional continuum, for example.

13. The washback effect of EPLIS on teachers' and learners' perceptions and actions

2:00–3:30 · Mario Laserna—2nd floor—Calle del Saber

Paula Ribeiro e Souza, University of Campinas, Brazil

English has been used as the common language in pilot and air traffic controller radiotelephony communications for international aviation. However, the insufficient English language proficiency on the part of both professionals has been pointed out as a contributing factor in accidents and incidents in several research reports, which has led the International Civil Aviation Organization (ICAO) to reconsider their existing provisions for the English language use in this context. In 2003, ICAO introduced a set of language proficiency requirements (LPRs) to ensure that air traffic control personnel and pilots involved in international flight operations are proficient in the use of English in order to deal adequately with unexpected or complicated situations. Since then, Brazil has developed its own examination for assessing the English language proficiency of air traffic personnel in order to meet ICAO LPRs. EPLIS (English Language Proficiency Exam for the Brazilian Airspace Control System), a high-stakes test by any account, was first implemented in 2007. For the first 7 years, it was delivered only to in-service controllers and aeronautical station operators, but in 2014, pre-service air traffic controllers started sitting EPLIS in the last semester of their two-year training program.

This study aims at investigating the washback effect of EPLIS on teachers' and students' perceptions and actions in an Air Traffic Control Initial Training Program. Adopting a mixed method sequential explanatory design (Creswell e Clark, 2011; Terrel, 2012), this study comprises two consecutive phases. In phase 1, the focus is to identify some patterns and trends of

perception and behavior among students and teachers in relation to EPLIS, as well as the impact of the exam on aspects of English teaching and learning in the training program. A 64-item questionnaire was designed and divided in 4 parts: general Information; knowledge about EPLIS; test preparation; and views and perceptions of the impact of EPLIS. First, descriptive statistical analysis of the demographic variables will be conducted. Then, exploratory factor analysis and principal component analysis will be employed to identify the major constructs present in the questionnaires. Finally, correlation analysis will be used to plot the significance of the relationships between variables. The hypotheses raised in Phase 1 will guide the collection of qualitative data, which will be conducted through focus group interviews with students, individual interviews with teachers and classroom observations. The aim of Phase 2 is to expand on, to refine and to explain the initial results in depth. All data will be triangulated in order to provide a holistic account of the phenomenon under investigation. This work-in-progress section will focus on the research findings obtained in the pilot study and in Phase 1.

Given the high stakes of EPLIS and its introduction as a mandatory test for pre-service air traffic controllers, its effects are expected to have been maximized within the context of this study. This study intends to provide separate and combined results from two major stakeholders, students and teachers, on the impact of the test and will contribute to a more robust theory about the washback effect – what its nature is, how and in what circumstances it acts, how it can be measured and what forces intervene in the process.

14. Pairs in written assessments

2:00–3:30 · Mario Laserna—2nd floor—Calle del Saber

Sean Dunaway, George Mason University, USA

To be valid, assessments must be authentic, practical, and reliable; however, many assessments are solely individual. Having left academia, many tasks in learners' fields require effective teamwork. Collaborative learning in education brings success in academic achievement and overall student motivation (Felder et al, 1998; NSSE, 2006); therefore, assessment tasks that provide opportunity for collaboration may better predict learner success after completing their education.

The concept of dyad or paired testing has been used in oral proficiency tasks (Grubor, 2013; Bennett, 2010) and in graduate nursing programs (Rossignol, 2004), but is not commonly used in written assessments (e.g. reading, vocabulary, and grammar assessments). The presenter first used dyad testing in an EFL context to reduce cheating, but has found advantages in using dyad testing for written assessments at his university's Intensive English Program. In taking assessment as pairs or groups of three, test-related anxiety was significantly reduced, and students engaged in productive negotiations of meaning that may have enhanced uptake of course content. Overall academic performance did not seem to be significantly affected (i.e. tests that learners had taken as individuals had similar scores to those taken in dyads), but when assessments were returned graded to students, more discussion and uptake seemed to take place as learners reminisced on shared episodic memories of their decision-making processes. Instances of predatory collaboration, where one test taker copied another's answers with little communication, were infrequent. Even weaker learners seemed eager and confident to share their ideas with their more proficient partners.

All in all, results seem promising, but are limited by the summative nature of final course grades. As much as students must work together for greater achievement and attainment of learning objectives, the final piece of feedback in many educational settings is a letter grade for the individual. The presenter will present current evidence and anecdotes from his action research on pair testing of written assessments, and hopes to collaboratively develop further principles of effective dyad testing during this presentation

15. In VIVO Vs. in VITRO: The journey of a secondary English public examination in Bangladesh

2:00–3:30 · Mario Laserna—2nd floor—Calle del Saber

Nasreen Sultana, Queen's University, Canada

The presentation aims at discussing the ongoing tension, how a test may work differently or may not work as it has been thought once it starts operating in the context. In this regard, Zumbo's (2015) argument about in VIVO vs. in VITRO will be used to discuss bringing examples from a unique context such as Bangladesh, where various social, political and economical forces work together or against each other in determining the success of any examination. Zumbo (2015) argues assessment as something in vivo rather than in vitro, which means, when examinations are designed by the test designers and policy makers, it is like testing something in the lab or in a glass in a controlled environment, which is called in Vitro. Contrarily, the same examination becomes different when it is actually taken by the test takers in the context in a living educational structure, which is in VIVO. Thus, the purposes as well as the perceptions of the examination could be different when it is in use unlike the way it has been thought while designing at the policy level.

The context of the study is the secondary public examination in Bangladesh, known as the Secondary School Certificate (SSC) examination, a prestigious examination in the social context of the country. According to the National Curriculum (2012), the SSC English examination assesses the creativity of the students that how far they can use English in their own context. However, Imam (2005) finds out that in Bangladesh the English proficiency level of the undergraduate students is equivalent to the standards of grade seven students. Further, Poddar (2016) reports that due to teachers' teaching to the test classroom practice, most of the students pass the SSC English examinations with good grades, although their level of English remains very poor. Even as a teacher of many years in Bangladesh, I often find it frustrating that students having A+ (top grade) in their SSC fail to exhibit the required knowledge of English language in their real life. So, somehow the examination is not working as it is supposed to work. Hence, I intend to trace the understanding of the policy makers and test designers about this crucial examination and how it has been perceived by the test users such as teachers, parents and students to realize the reasons behind the failure of the examination.

Thus my research questions are: a) how are the purposes of SSC English examination perceived by the policy makers, test designers, teachers, students and parents? and b) is there a working social factor/context that might work in building those perceptions? I will use mixed method (Creswell, 2013) technique to collect data from the mentioned stakeholders. The proposed study will be the first to investigate the journey of the secondary English examination in Bangladesh from the policy makers/test designers' home to the test users' abode and how they may react different in different environments.

16. How literate in language assessment should English teachers be?

2:00–3:30 · Mario Laserna—2nd floor—Calle del Saber

Sonia Patricia Hernández Ocampo, Pontificia Universidad Javeriana, Colombia

When discussing about test design, it is frequent to hear teachers say: "I am not an expert. Why do something that has already been done by experts?" This "leave it to the experts" position cannot be a commonality among language teachers. Who, if not them, should be in charge of designing tests to measure their students' learning? Why is it that language teachers are considered non-experts in language testing?

This work-in-progress presentation aims to show to what extent English teachers should be literate in language assessment in our context, teaching English to pre-service language teachers in a teaching degree program at a university in Bogotá. For this to be done, the group of English teachers will be interviewed about their assessment practices and the knowledge they require to carry them out. So will the English component coordinator, member of the group of teachers in charge of revising the exams designed in order to ensure the tests meet the specifications and the courses programs. A sample of student teachers at different stages in their major will also be interviewed about the knowledge on language testing they expect their teachers to have, and the type of training in assessment they need. Results are to be contrasted with what the theory says about the level of literacy this group of stakeholders is expected to have. This will hopefully impact English teachers on their beliefs about assessment practices, test design particularly, and will empower them to actively design tests that suit the population they work with.

17. Reducing the number of options in multiple-choice items: Does the option reduction method matter?

2:00–3:30 · Mario Laserna—2nd floor—Calle del Saber

Michelle Y. Chen, Paragon Testing Enterprises / The University of British Columbia, Canada

Jayanti Banerjee, Paragon Testing Enterprises, Canada

Zhi Li, Paragon Testing Enterprises, Canada

It is common for multiple-choice items to be accompanied by four options. This is predominant because, when an item is well-written, the presence of four options ensures that test takers are not able to easily guess the correct answer. This claim, however, rests on the assumption that all four options are equally plausible. In operational tests, this is not always the case. Typically, multiple choice items with three distractors may have one of them that is weak or obviously wrong. That weak distractor may also take item writers more time to write. It is therefore argued that three-option multiple-choice items offer practical advantages such as efficiency in item writing and test administration (Haladyna & Downing, 1993; Lord, 1977). Empirical studies also indicate that three-option items are as psychometrically effective as their four-option counterparts in many cases (Lee & Winke, 2011; Rodriguez, 2005; Shizuka, Takeuchi, & Yashima, 2006; Thanyapa & Currie, 2014).

To date, the studies comparing three- and four-option multiple-choice items have primarily relied on removing distractor(s) from the items originally with four options. The three common option-reduction methods are: removing the least plausible distractor; the least popular distractor; or, the least effective distractor. In the first method item writers/reviewers identify the least plausible distractor. The latter two methods use option analysis with response data. These different option-reduction methods make different assumptions on both item development procedures and test taker behaviors, although these assumptions are rarely discussed or tested. This project aims to address this gap through: 1) examining the distractor effectiveness of a high-stakes English proficiency test; 2) comparing the psychometric properties of the three-option items generated from each option-reduction method; 3) discussing the effectiveness of item writer judgments of option plausibility; and, 4) exploring the effect of each option-reduction method upon the test construct.

In this exploratory study, we will focus on listening items from a high-stakes English proficiency test. Each of the option-reduction methods described above will be used to compile three versions of listening test items. Then, instead of administering these versions as part of an experimental design, we will draw on existing response data. The item response theory (IRT)-based item parameters (difficulty, discrimination, and guessing) will be compared across the three modified versions with the original version to investigate the possible effects of the option reduction methods. In addition, we will invite a panel of item writers to discuss their judgment of option plausibility and the effects of reducing number of options on the construct. We will pay special attention to compare the option evaluations made by item writers/reviewers with the two data-based methods.

During the work-in-progress session we will reflect upon our preliminary results and seek feedback on our plans for investigating the effect of option reduction upon the test construct. We hope this project will contribute to our understanding of how options function in multiple choice items.

18. A comparison of scenario-based and traditional achievement tests

2:00–3:30 · Mario Laserna—2nd floor—Calle del Saber

Jorge Beltrán, Teachers College, Columbia University, USA

This study was conducted in an effort to explore the comparability of the usefulness and validity of a Spanish achievement test when delivered in a scenario-based format, as opposed to the traditional test format. Scenario-based testing has shown potential as an alternative to the traditional approach to standardized testing, for instance, in assessing test-takers reading comprehension while also promoting learning throughout the test (O'Reilly, Weeks, Sabatini, Halderman, & Steinberg, 2014; O'Reilly & Sabatini, 2013). Nonetheless, there is not enough research in the foreign classroom context, and a comparison of the processes and information collected through this type of assessment should be compared to that of traditional (non-scenario) tests.

The two versions of the test were criterion-referenced tests, with content sampled from the curriculum of a beginner Spanish class (A1 according to the CEFR). Both versions covered the four skills, as well as grammar and vocabulary. The texts for the reading section were selected based on their equivalence, as determined by the readability indexes Coh-Metrix and Lexile. Similar

procedures were followed for other skills, for example, the listening materials were determined to pertain to the A1 level of the CEFR and selected from a well-known Spanish textbook.

Fifteen students took both versions of the test. The order in which they took the test was randomly assigned. Each test took about 45 minutes to be completed by each student. After each test, test-takers completed a questionnaire, and a focal group was selected to further inquire about their experience with the test. Test scores, questionnaire responses, and the attitudes and perceptions as described in the focal group session will be compared and analyzed in order to provide an account of the possible similarities and differences in test-taker performance and attitudes towards each of these two types of tests.

19. Lexical competence in Japanese and the definition of a test construct for teachers of Japanese as a foreign language

2:00–3:30 · Mario Laserna—2nd floor—Calle del Saber

Monica Jessica Aparecida Fernandes Yamamoto, State University of Sao Paulo (UNESP), Brazil

Douglas Altamiro Consolo, State University of Sao Paulo (UNESP), Brazil

This ongoing study is part of a larger research project about the assessment of foreign language (FL) teachers' proficiency for pedagogical purposes by means of the EPPLE, a language proficiency examination designed especially for teachers and teachers-to-be. The EPPLE is under development at the State University of Sao Paulo (UNESP), in Brazil, and it aims at indicating teachers' linguistic proficiency in oral and written language, as well as causing a positive impact on teacher education courses in Brazil and elsewhere. Our study, nested in the aforementioned project, aims at developing a proficiency test that focuses on the receptive lexical competence of Japanese language teachers in Brazil. In this WIP presentation we deal with the aims, the initial findings and the research plan leading to the development of the test. Given the fact that lexicon underpins the four linguistic skills, a number of researchers consider lexical competence a key factor in effective language use for communication (Laufer, 1998; Nation, 1990, 2001; Schmitt, 2000; Webb, 2005). Due to the relevance of describing proficiency levels to test FL teachers, and the role of lexicon in language proficiency, we are developing a test to verify whether or not teachers certified from Letters courses (undergraduate courses for language and literature teacher education in Brazil) reach a minimum required level of lexical knowledge to teach Japanese as a foreign language (JFL). In order to define the test construct and its domain, a literature review of the JF Standard for Japanese-Language Education (JF Standard), a document produced by the Japan Foundation, an organization linked to the Ministry of Foreign Affairs in Japan that foster cultural interchanges and educational cooperation between Japan and other countries, and of the guidelines of the Common European Framework of Reference (CEFR) for languages is included in the theoretical framework. Surveys conducted by the Japan Foundation in 2012 revealed the existence of 3.98 million students of Japanese, 63, 805 teachers and 16,046 teaching institutions registered with the Foundation (Japan Foundation, 2012, p. 3). Such panorama demanded the establishment of an infrastructure for the quality and spread of JFL worldwide and the release of the JF Standard in 2010 to support Japanese language teaching, learning and testing. In this study, the CEFR and the JF Standard, and a literature review in language assessment and testing, have led to a comparative analysis of those documents, with focus on the role of vocabulary in

JFL. Such analysis supports the test construct and provides guidelines for proficiency scales for teachers of JFL's performance in our test and the specifications for lexical proficiency. Test items are being developed and the test will be piloted with experienced teachers of JFL at colleges and universities in Brazil.

20. The performance of Brazilian air traffic controllers in radiotelephony communications: Analyzing EPLIS

2:00–3:30 · Mario Laserna—2nd floor—Calle del Saber

Natalia de Andrade Raymundo, Brazilian Air Force/ UNICAMP, Brazil

This study aims at presenting initial research on the on-going validation process of the Brazilian Air Force English proficiency exam for Air Traffic Controllers, named EPLIS. 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 considered to be a high stake 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. In this presentation, we will make a parallel between what the ICAO rating scale considers as operational in Air Traffic Control regarding English proficiency, which is assessed by EPLIS, and the language skills Brazilian Air Traffic Controllers need in order to deal with language complications. We will analyze a 5-hour-sample of a ground-air interaction in English of a controller who has achieved the operational level in EPLIS, in order to observe if controllers who have been considered to be operational in EPLIS can communicate in situations in which phraseology is not enough. In addition, we will analyze how supervisors and controllers construct discursively the operational level. First results have shown that the controller level 4 may not be able to deal with all the complications that he/she might face during a work shift.

21. Examining the cognitive dimension of academic speaking ability through a scenario-based speaking test

2:00–3:30 · Mario Laserna—2nd floor—Calle del Saber

Yuna Seong, Teachers College, Columbia University, USA

Strategic competence has been acknowledged as an integral facet of language ability (e.g., Bachman & Palmer, 1996; Canale & Swain, 1980) in language assessment literature. Although there are varying approaches in its conceptualization and empirical investigation, literature on language learning strategies (e.g., Cohen, 2011; O'Malley & Chamot, 1990; Oxford, 2011) and language assessment (Bachman & Palmer, 1996; Phakiti, 2007; Purpura, 1999; 2014) concede that metacognitive and cognitive strategy use is essential in understanding strategic competence necessary for language learning and use.

Metacognitive and cognitive strategy use is particularly important in the academic domain and in understanding the nature of academic language ability because students are expected to use their language knowledge to perform tasks that require complex thinking skills (Chamot &

O'Malley, 2004; Zweir, 2008). For instance, students would have to comprehend and synthesize listening and reading materials in order to respond either through speaking or writing.

This study aims to examine the nature of academic speaking ability and its cognitive dimension using a specially designed academic speaking scenario-based assessment. The online Scenario-based Academic English Speaking Test (SBAEST), designed to replicate real world academic speaking demands, requires students to listen to various audio and video materials on a certain topic and provide oral responses to questions by summarizing and synthesizing the materials and responding to them with their own opinions. In order to examine the cognitive dimension of the test takers' academic speaking ability, the test also includes strategy tasks specifically designed to elicit students' actual use of metacognitive and cognitive strategies such as planning, predicting the content of the listening material, and recalling key points.

This study is a work in progress and the test is scheduled to be administered to advanced ESL students in academic listening and speaking courses at an American university. In addition to examining the SBAEST's usefulness (Bachman & Palmer, 1996) in measuring academic speaking ability, the main purpose of this study is to examine the nature of academic speaking ability and its cognitive dimension (i.e., strategic competence) by examining the takers' performance on the strategy tasks and how it relates to their speaking performance. Closer examination of the students' performance on the strategy tasks could also shed light on the characteristics of effective or less effective strategy use helping us gain insight into whether measuring strategic competence could even be possible. If so, this could address the needs of many academic language programs that are in search for better ways to assess international students' academic language ability. If separate assessment of language knowledge and strategic competence becomes possible, this may help us in getting a step closer in terms of knowing in which area the student needs assistance or further instruction: language knowledge or using the language knowledge to perform advanced academic tasks.

POSTER ABSTRACTS

FRIDAY, JULY 21st

1. Our TALE: Developing online teacher assessment literacy training materials

10:45–12:00 · Mario Laserna—2nd floor—Calle del Saber

Dina Tsagari, University of Cyprus, Cyprus

Karin Vogt, University of Heidelberg, Germany

Ildiko Csepes, University of Debrecen, Hungary

Anthony Green, University of Bedfordshire, United Kingdom

Nicos Sifakis, Hellenic Open University, Greece

Research undertaken in exploring teacher language assessment literacy (LAL) has shown that in many educational systems across Europe, English Language Teachers (ELTs) are not well prepared to create quality assessment materials and procedures (Csepes, 2013; Tsagari, 2013; Vogt and Tsagari, 2014). This is at least partly because they are not sufficiently trained in the area of language testing and assessment. There is therefore an urgent need to develop an efficient, relevant and sustainable LAL training infrastructure for ELTs to help them attain relevant LAL levels and support the expansion and exchange of LAL expertise between European educational contexts and beyond in a creative and innovative way.

This poster will present preliminary results of a three-year long project entitled 'Teachers' Assessment Literacy Enhancement' (TALE). The project involved the collaboration of experts from five different European countries to create innovative LAL training materials and services delivered through a freely accessible online course via synchronous and asynchronous modes. The poster will focus on the evaluation and feedback from trial participants from the second phase of the project life and discuss the processes and lessons learnt. The poster highlights the complexities and challenges involved in the identification, design and piloting of online training materials intended to help teachers' enhance their LAL. These materials are intended be both contextually situated and suited to the needs of users across stakeholder boundaries, e.g. inside and outside the project countries. With this poster we aim at sharing the experiences gained from the life of the project so far and hope to receive constructive feedback and advice from the conference audience especially those with an interest in developing teachers' LAL.

2. Developing materials to enhance the assessment literacy of Parents and Educators of K-12 English language learners

10:45–12:00 · Mario Laserna—2nd floor—Calle del Saber

Ahyoung Alicia Kim, WIDA, University of Wisconsin-Madison, USA

Mark Chapman, WIDA, University of Wisconsin-Madison, USA

Carsten Wilmes, WIDA, University of Wisconsin-Madison, USA

In the United States, kindergarten to 12th grade (K-12) English language learners (ELLs) are required to take an annual English language proficiency exam as part of the federal government regulations. The purpose of this exam is to (1) ensure that ELLs are acquiring the type of English required for academic success in English-medium classrooms, and (2) identify ELLs who may be re-classified out of the legally-protected status. This poster presentation describes a large-scale language assessment, which evaluates K-12 ELLs' English proficiency in the four domains of speaking, listening, reading, and writing. As of the current academic year, the test has grown to be used by 38 states and assesses about two million ELLs a year. After the administration of the test, relevant stake-holders receive score reports.

It is important that the stakeholders, particularly parents and educators of K-12 ELLs, accurately understand these score reports. Parents and educators need to have a clear grasp of the information presented in the score reports to understand the student's English language development and to make sound educational decisions based on interpretations of the score data. Therefore, a number of materials and resources have been developed to enhance parents' and educators' understanding of the information presented in the score reports.

The purpose of this poster is to present the various resources created to enhance parents' and educators' assessment literacy; namely to support the interpretation and use of the score reports. These resources reflect the needs and suggestions provided by parents and educators, collected via survey and interviews during a two phase study. In Phase 1, over 1400 educators completed an online survey regarding their interpretation of the score reports; in Phase 2, 13 parents and 18 educators participated in interviews to recount how they interpret and use the score reports. Moreover, the score report resources incorporate test developers' expert knowledge and experience in communicating test scores to various stake-holders. The main score report resources that have been developed include reports translated into over 40 languages, a parent guide to help with understanding the data reported, an Interpretive Guide for administrators and educators that describes in detail how to interpret the test scores, and a range of webinars and workshops that are delivered across the states that use the assessment.

The level of technical content presented in these materials varies depending on the level of assessment literacy and needs of the different stakeholders. For example, the parent guides have been developed with the input of numerous parents and present a reader friendly and non-technical account of what the test scores mean. In contrast, the Interpretive Guide is written for state and district educational and assessment staff and is much more technical in its detailed descriptions of how the test scores may be used to make inferences about students' language development and growth at the school and district levels. The poster will describe the intended audiences and how the support materials address their individual needs, based on different skills, knowledge, and competencies of the various stakeholders.

3. Measuring Language Proficiency in K-12 US Bilingual Programs: Contesting the Monolingual Narrative of Bilingual Learners

10:45–12:00 · Mario Laserna—2nd floor—Calle del Saber

Mariana Castro, WIDA Consortium at UW-Madison, USA

Margo Gottlieb, WIDA Consortium at UW-Madison, USA

Mark Chapman, WIDA Consortium at UW-Madison, USA

Bilingual education programs in the US typically have the goal of helping their students become bilingual and biliterate, in addition to aiding their overall academic achievement (Crawford, 2004). The goal of many bilingual programs is that students are able to understand and produce orally and in written form, in two languages. Historically, however, the measurement of progress and achievement in these programs has focused on students' academic achievement as measured in English, and in some cases, on the rate of English language proficiency gain achieved by the students (Collier & Thomas, 2004)). However, little has been done to formally assess students' language proficiency over time in other languages. PODER is an assessment being developed at the Wisconsin Center for Education Research in partnership with the Center for Applied Linguistics to address this concern, with the goal of assessing Spanish language proficiency in academic contexts.

Spanish-English Bilingual programs in the US vary greatly in their student populations and pedagogic goals, presenting challenges to the development of an assessment that is relevant to all stakeholders. Bilingual programs may include newcomers to the US from Spanish speaking countries, students who speak Spanish at home in the US but may have never lived in a Spanish-speaking country, heritage speakers of Spanish, and students who speak English at home and who do not have additional exposure to Spanish except at school (Crawford, 2004; Yoon Kyong, Hutchinson & Winsler, 2015). Some bilingual students may be part of programs that only offer instruction in Spanish until children reach sufficient proficiency in English to transition into English-only program models. Other programs may offer bilingual programs as an approach to maintaining the home language of their students. Still, other programs may offer bilingual programs as a way to help students acquire Spanish as an additional language (Crawford, 2004; Yoon Kyong, Hutchinson & Winsler, 2015).

This diverse population, in terms of cultural and linguistic background and variety in programmatic approach mean that the experiences and knowledge brought by Spanish language educators and administrators are at least equally important in the test development process as that provided by the language assessment specialists. The expertise of Spanish language educators in terms of language pedagogy, sociocultural values, and local educational practice is essential to the development of a valid Spanish language assessment in the US K-12 bilingual education context. This poster will focus on how the development of PODER incorporated the input of both test development professionals, literate in a traditional understanding of how language tests are created, and of Spanish language educators who were integral to the test development process. It will outline the dynamic process of competency development that featured in the project as Spanish language educators and test developers worked collaboratively on test design and content development. The poster will present the feedback collected from Spanish language educators in Puerto Rico and three US states (IL, NM, and WI) on how they felt they benefited professionally from participating a test development project.

4. Stakeholder views on the assessment and certification of English language proficiency in global interuniversity collaboration

10:45–12:00 · Mario Laserna—2nd floor—Calle del Saber

Lut Baten, KULeuven, Belgium

Jan van Maele, KULeuven, Belgium

Yoennis Moreno, Universidad de Oriente, Cuba

Geisa Dàvila Pérez, Universidad de Oriente, Cuba

Frank van Splunder, University of Antwerp, Belgium

Autonomous learning means that students should have a 'capacity for detachment, critical reflection, decision-making, and independent action' (Little, 1991). It is a capacity taken for granted with PhD students applying for a scholarship in a global context. Moreover, in an academic environment, this capacity needs to be displayed in English. As an entry ticket to this world, candidates have to provide valid certificates proving their level of English. However, do these certificates also cover their autonomy (in English) in working, studying, networking at their host universities?

In the process of project design and implementation, stakeholder expectations need to be regularly consulted. Since 2004, Flanders (Belgium) and Cuba have cooperated in capacity building projects for research in human and natural sciences, engineering and technology in higher education, granting joint PhDs to Cuban students (http://www.vliruos.be), with the use of English as the lingua franca. This certification requirement has so far been an internal issue with many pitfalls and frustrations on both sides, as criteria were debatable and requirements of the Flemish host universities differing. In 2013, we started a transversal project at Universidad de Oriente (Santiago de Cuba) to find out how local tests may cross-fertilize with standardized international tests, and hopefully lead to the launch of an official language testing center by 2018, when the project ends. Recent political evolutions in Cuba have made this undertaking all the more adamant as the local situation now presses for more robust test validity and for assessment literacy from all stakeholders. As emphasized by different speakers at the recent Language Testing Literacy Symposium at Lancaster University (2016), this endeavor should include not only the testers and language instructors, but also the test users and university administrators.

Following Morris and Baddache's (2012) five-step approach to stakeholder engagement, this poster outlines the perspectives of different stakeholders at the Universidad de Oriente with respect to the assessment and certification of English language proficiency. Building on the stakeholder mapping exercise in Van Maele, Rodríguez González, Díaz Moreno, van Splunder and Baten (2015), in which we identified the most important stakeholder groups, we will now focus on the perspectives of the internal stakeholder groups, notably the language instructors, project leaders, PhD students, and university leadership on the Cuban side as well as the project leaders on the Flemish side. We will define the various engagement strategies and how to prepare for them, evaluate prior and on-going engagement actions, and report on the impact of the language assessment literacy trainings that have taken place.

5. An electronic pre-test for the EPPLE examination: Test design, preliminary results, technical challenges and mobile technology

10:45–12:00 · Mario Laserna—2nd floor—Calle del Saber

Douglas Altamiro Consolo, State University of Sao Paulo (UNESP), Brazil

Debora Mieko Aguena, State University of Sao Paulo (UNESP), Brazil

In this poster presentation we report on the design and implementation of an electronic pre-test for the EPPLE, a language proficiency examination for foreign language teachers. English language teachers' proficiency for teaching purposes has been a challenge for a number of non-native teachers worldwide, leading to claims for more successful actions towards teachers' language development and assessment that have motivated the EPPLE project. Due to the fact that not all EFL teachers may already be ready to take EPPLE and achieve successful results, pre-testing is seen as a guidance for candidates so as to motivate them to either take EPPLE or else to engage in actions for language development in order to probably be successful in EPPLE on a future attempt. It has been opted to pilot a first version of the EPPLE pre-test for oral skills only, focusing on both listening comprehension and speaking production. The test has been designed on Lingt Classroom (http://lingtlanguage.com) and it provides detailed instructions for candidates to take the test, and to submit their answers for correction and test results. It can be easily taken online and in mobile devices such as cell phone and tablets. Based on its preliminary results, investigations about the EPPLE pre-test proceed towards advances in its technical aspects, to make access and test-taking easier for candidates, as well as the implementation of an efficient process for test correction and the provision of feedback to candidates wishing to take the EPPLE examination.

6. Introduction of four-skilled college entrance exams: English teachers' perceptions and needs for assessment literacy among teachers

10:45–12:00 · Mario Laserna—2nd floor—Calle del Saber

Yasuko Okabe, Center for Entrance Examination Standardization, Japan

Mika Hama, Benesse Corporation, Japan

Yuko Umakoshi, Benesse Corporation, Japan

It is often mentioned that students in Japan emerge from six years of post-elementary English education with limited competence in productive skills. In an effort to change this, the Japanese government is moving toward replacing current college English entrance exams that only assess receptive skills with exams that focus on four language skills, hoping that the change will encourage English teachers to focus more on productive skill development in the classroom. Under the noted circumstances, we have investigated two questions: 1) whether with the change in the exams, teachers thought they needed to change the way they teach, and 2) what were teachers' perceptions regarding the government approved four-skilled exams, especially regarding the productive skill component. In order to address these questions, a questionnaire

was sent to 165 Japanese English-language teachers at the secondary level who participated in an event in which they took one of the new entrance exams, GTEC CBT, to familiarize themselves with the upcoming change in the entrance exams.

Of those 165 teachers, 106 teachers submitted their answers to the questionnaire. The data from the questionnaires showed 90% of the teachers indicated that they would change their teaching style; however, their comments revealed that they had numerous concerns regarding the change in the exams and the impact it would have on their teaching. A closer look at the comments revealed that the concerns stemmed primarily from a lack of knowledge of teaching productive skills. Providing support to ensure teachers successfully adjust to the change and meet government expectations is critical as many washback researchers have argued in the past (e.g., Cheng, 2008; Wall & Alderson, 1993; Watanabe, 2004).

Regarding the teachers' perceptions on one of the new exams, GTEC CBT, the results showed that 72% of teachers had positive impressions of the test itself. However, questions and concerns were also raised regarding productive skills assessments. Some teachers also mentioned that they acknowledge the need to assess student progress in their classrooms because of the change in the entrance exams, but are uncertain how to proceed. These questions and concerns point to a need for close communication between test developers and teachers and to the importance of supporting the development of assessment literacy among teachers (e.g., Fulcher, 2012; Malone, 2013; Taylor, 2009).

In this poster, the data addressing the two investigation questions mentioned above and the concerns teachers mentioned are shown. Teacher supporting activities, which were developed and provided by the creators of GTEC CBT based on teachers' concerns, are also mentioned.

7. Language assessment literacy of EFL teacher trainers

10:45–12:00 · Mario Laserna—2nd floor—Calle del Saber

Salomé Villa Larenas, Lancaster University, United Kingdom

Research on language assessment literacy (LAL) has so far primarily focused on the knowledge, skills and practices of language teachers, learners, and a range of stakeholders in the educational community (Taylor, 2013; Fulcher, 2012; Malone, 2013; Pill & Harding, 2012; Scarino, 2013; Tsagari, 2011). However, to date, to our knowledge, no studies have looked into detail in the LAL of teacher trainers, whose role it is to prepare pre-service teachers for key areas of the teaching profession. Unmistakably, the latter includes assessment of language learners to evaluate their progress and achievements in language proficiency development. At the same time, recent surveys have shown that teachers feel insufficiently trained in language assessment (see e.g. Vogt and Tsagari, 2014), and researchers have argued that one reason is limited or no focus on assessment issues in teacher education programmes (Brindley, 2001, Levy-Vered & Alhija, 2015; Lopez Mendoza & Bernal Arandia, 2009; Mosquera & Marcias, 2015; Volante & Fazio, 2007).

Foreign language teacher education programmes often have the dual role of providing training in (language) pedagogy as well as developing the foreign language proficiency of the teacher trainees. Thus, these programmes have scope for formal courses on language assessment knowledge and skills development as part of the pedagogy strand, and they also offer opportunities to model good practice in language assessment through the foreign language

acquisition courses as the latter will involve assessment of the pre-service teachers' language proficiency development. Indeed, recent research has emphasized the need to combine formal training with practical implementations to capitalize LAL development (Brunfaut & Harding, in press; Fulcher, 2012). Thus, those working in teacher education are ideally placed for the multi-dimensional LAL development of future teachers. At present, however, this potential seems underexploited. In addition, it is unclear to what extent teacher trainers themselves are assessment literate; although they may have gained some practical assessment expertise through extensive classroom teaching experience, they often have gone through similar teacher training programmes with little or no emphasis on LAL.

Based on these gaps, a study has been designed which aims to explore the LAL of language teacher trainers in terms of their theoretical knowledge and skills (assessment pedagogy) as well as their practical knowledge and skills and their application base (modelling assessment through language acquisition courses for the pre-service teachers). More concretely, the study will look into the language assessment literacy of English as a Foreign Language (EFL) teacher trainers in Chile, and the extent and nature of the LAL training they conduct with pre-service EFL teachers. The study involves a survey and interviews with Chilean language teacher educators in terms of their language assessment knowledge and practices, analyses of currently used instructional materials, and classroom observations.

This poster will present the findings from a pilot questionnaire which was administered to English teacher trainers, as well as from preliminary analyses of pre-service language assessment course materials at Chilean Universities. This will offer an initial look into the current state-of-affairs of LAL training in Chilean teacher education programmes.

8. Re-evaluating commonly held views of residual-based fit statistics in language assessment research: Rasch analysis of the CanTEST listening test

10:45–12:00 · Mario Laserna—2nd floor—Calle del Saber

Ángel Arias, University of Ottawa, Canada

In validation research involving Rasch modeling, a large degree of importance is placed on the objective measurement of individual abilities and item difficulties (Karabatsos, 2000). Thus, it is important to check the consistency of the observed data with the assumptions underlying the measurement model. In the context of Rasch measurement, it is only when the data conform to the model expectations that inferences can be validly made. Residual-based fit statistics (Smith, 1991; Wright, 1977; Wright & Panchapakesan, 1969; Wang & Chen, 2005) are usually used to check the extent to which the data fit the Rasch model. Although the use of residual-based fit statistics have been criticized regarding their instability and sensitivity to sample size (e.g., Karabatsos, 2000), they are widely use because they have shown to be useful. In fact, there are well informed broad guidelines for the use of these statistics (Wu & Adams, 2013) to determine data-model (e.g., item, person) fit.

In the language testing community, it has become common practice to use fixed rules of thumb or guidelines on numerical ranges for acceptable residual-based fit statistics to assess item and person fit (e.g., Aryadoust, Goh & Lee, 2011; Eckes, 2005; Hsieh, 2013; Winke, 2011). However, the use of fixed numerical ranges for misfit diagnosis across different testing situations varying in

sample size leads to both overdetection and underdetection of misfit. This study reports the results from a Rasch analysis that takes into account sample size for misfit diagnosis when considering mean square fit statistics. The data (n = 845; n = 1,548; n = 1,759) stemmed from three listening tests of the CanTEST and the focus of the analysis was to flag misfitting items for review purposes. The CanTEST is a standardized English proficiency test used to determine whether or not candidates meet the language requirements of Canadian postsecondary institutions and Canadian professional licensing associations.

The results suggest that adhering to sample-sensitive criteria for misfit diagnosis in Rasch measurement helps in signalling misfitting items that may not be flagged under traditionally fixed cutoff criteria. A post hoc content review of the misfitting items revealed problems that required item editing and revision. These results have practical implications for item review processes that use misfit information as an indicator for item fine-tuning. In addition, inferences about item quality and person standing on a given language construct are more precise and theoretically founded when misfit diagnosis is based on sample size.

It is worth remembering that fit statistics only provide one small piece of information in making judgements about the quality of an item. Other information such as test targeting, reliability, and construct representation need to be considered before deleting test items.

This work has fundamental implications for the language assessment literacy (LAL) development of professional language testers who need to be aware of knowledge of theory, principles and concepts (Taylor, 2013) relating to Rasch measurement to ensure the quality of test items and the integrity of language tests.

9. Establishing construct validity in the development of an oral placement test at an intensive English program

10:45–12:00 · Mario Laserna—2nd floor—Calle del Saber

Kristin Rock, University of Hawai'i at Mānoa, USA

Tasks that require learners to accomplish communicative goals in the target language have increasingly factored into language proficiency tests at the university-level to ascertain learners' readiness to engage in academic discourse (Norris, 2016). However, despite growing interest, more research is needed on the "development, use, and validation of task-based tests" (Norris, Brown, Hudson & Bonk, 2002, p.414). Factors that ought to be considered include the complexity of simulations of target L2 communicative tasks, the quality of the rating scale used to evaluate the samples, and the consistency of the rater(s), seeing as these factors influence the outcome and reliability of oral placement tests. Investigations of rater reliability have been well documented in previous literature (Kondo-Brown, 2002; Weigle, 1998), though practical applications of reliability calculations to task-based language assessments are difficult to find.

Using FACETS analysis (Linacre, 1996) to examine rater judgments of oral performance, the current research reports on the development and incorporation of a task-based oral assessment into the placement procedures for a university-based Intensive English Program (IEP). Based on the scores assigned by three teacher raters to the performance of 60 university students from Japan and Korea on a series of communicative tasks, the relationships among the performance of examinees, raters, and tasks were first examined using FACETS analysis. Then, by calculating the

G-coefficients for various raters and tasks, the number of raters and tasks that would produce the most reliable scores for this university's IEP were identified. This information then allowed stakeholders to make an informed decision about the most efficient way to implement task-based language assessment at this IEP. This poster will present a real-world application of task-based language assessment, and results will be discussed in terms of the positive impact of the project on increasing the language assessment literacy of the administrators, teachers and learners at the IEP.

10. Automated scoring of speaking: Lessons from a real-world assessment

10:45–12:00 · Mario Laserna—2nd floor—Calle del Saber

Larry Davis, Educational Testing Service, USA

Su-Youn Yoon, Educational Testing Service, USA

Anastassia Loukina, Educational Testing Service, USA

Klaus Zechner, Educational Testing Service, USA

Automated scoring systems have the potential to assess constructed responses faster than human raters and at a lower cost. Accordingly, considerable research has been carried out on the technical development of automated scoring systems for productive language, and to a somewhat lesser extent, efforts have been made to identify validity issues for automated scoring as well. However, there are few accounts that describe the complexities of implementing automated scoring in an operational language assessment. In this presentation we describe an effort to implement automated scoring of speaking within the context of a low-stakes test of functional English for classroom teachers, which was aligned with a teacher-training curriculum. We outline the various practical and validity considerations encountered and our responses to these challenges, with the goal of sharing lessons learned.

In the development phase challenges revolved around issues of construct and feasibility. The assessment included speaking items that elicited highly constrained speech (e.g., read aloud) or predictable short answers; these items reflected the language targeted in the curriculum and were conducive to optimizing the automated speech recognition (ASR) module of the scoring system. But, this approach created technical challenges in measuring aspects of the construct. For example, measures of suprasegmental features tended to be unstable for short responses, so responses were concatenated to measure these features. A feasibility issue was that some test taker responses could not be scored by the automated system due to audio quality or other issues. In response, an automated filtering system was developed for identifying such responses for subsequent scoring by human raters; this hybrid approach combined the efficiencies of automated scoring with human expertise in resolving difficult cases. In the operational phase challenges included initially low numbers of test takers as well as differences in the language backgrounds of test takers who participated in the pilot vs. those taking the operational test. These issues delayed the training of the automated scoring system to score additional test forms, and so human scoring was used exclusively until enough data had been collected to ensure good performance of the automated system. Finally, an overall conceptual challenge was properly evaluating system performance in the context of an achievement test, where the typical

performance metric used for automated scoring, correlation with human scores, was less useful. So, in addition to correlation with human scores, other metrics such as absolute score difference and classification accuracy were also used.

To conclude the presentation we summarize the conceptual and practical issues that must be considered in the actual implementation of automated scoring and how we addressed these issues in our context. We argue that automated scoring must be considered as part of an entire assessment system, and share our "lessons learned" to illustrate some of the complexity inherent in such systems.

11. Why listening to the voices of stakeholders matters in defining LAL

10:45–12:00 · Mario Laserna—2nd floor—Calle del Saber

Dina Tsagari, University of Cyprus, Cyprus

Language assessment literacy – LAL (Malone, 2013; Inbar-Lourie, 2008; 2013; Taylor, 2009; 2013) has so far been defined as the competences required in language assessment by various stakeholder groups (e.g. developers, teachers, candidates, users, the public). Fulcher's (2012) definition of LAL included skills, knowledge and abilities, awareness of the theoretical basis for assessment, and awareness of "the role and impact of testing on society, institutions and individuals" (p. 125). Taylor (2013) presented a model for LAL where assessment needs are differentiated according to the requirements of various stakeholder groups and areas of literacy.

Against this background, the paper seeks to understand the needs and the ways two major stakeholder groups (teachers and students) conceptualize LAL within the ecology of a particular educational context. More specifically, the paper discusses the results of the first phase of a three-year European-funded project. This involved a needs analysis achieved through extensive consultation with English language teachers (ELTs) and their students aiming to ensure that the future online LTA training course envisaged for the purposes of the project would meet their needs. The study adopted an exploratory method design based on quantitative data collected via online questionnaires designed to investigate how teachers and students perceive and practice assessment. The paper will report on the results obtained from 404 ELTs and 909 students. Descriptive and inferential statistics have been used in order to examine the trends identified between teachers and students and also interrelationships between the LTA needs of ELTs and their students. The findings provide in-depth understanding of the notion of assessment concepts, practices and priorities among different stakeholder groups. This highlights the need to view LAL as a dynamic process of competency development based on consultation with various groups of stakeholders rather than a static skill-based construct.

12. The road to success: Advising English language education policy in Chile–Findings and recommendations from research on a national benchmarking test

10:45–12:00 · Mario Laserna—2nd floor—Calle del Saber

Graeme Harrison, Cambridge English Language Assessment, United Kingdom

Agnieszka Walczak, Cambridge English Language Assessment, United Kingdom

In Chile, from 2010 to 2014 a national English language testing programme was implemented called SIMCE Inglés. Students in 3rd grade of secondary school were evaluated biennially using a test of Reading and Listening based on the Cambridge English: Key examination, which measured from pre-A1 to B1 or above. In 2014, a total of around 150,000 students were evaluated across the state and private sectors.

In 2015, using the results of SIMCE Inglés, as well as a number of questionnaires given to stakeholders such as students, parents and teachers, and other contextual data, such as socio-economic profile and location of schools, Cambridge English undertook a research project to measure factors affecting English language attainment in Chile. The purpose of the study was to offer evidence-based recommendations for English language education policy. Three analytical techniques were employed to make sense of the data: descriptive statistics; multilevel modelling; and regression analysis.

Findings were along expected lines but interesting nonetheless. A small but significant improvement in performance in most school contexts between the 2012 and 2014 test administrations was found, suggesting that recent Ministry of Education initiatives may have positively impacted on English language education. Furthermore, various factors were found to be associated with positive performance on the test. Among those were: the importance of student exposure to English, both inside and outside the classroom, as well as teacher-related factors such as level of qualification, English language proficiency, and use of resources.

13. The effects of timed and untimed testing conditions on the item difficulty parameters of Spanish reading proficiency test

10:45–12:00 · Mario Laserna—2nd floor—Calle del Saber

Troy Cox, Brigham Young University, USA

Gregory Thompson, Brigham Young University, USA

Clifford and Cox (2013) define proficient reading as "the active, automatic, far-transfer process of using one's internalized language and culture expectancy system to efficiently comprehend an authentic text for the purpose for which it was written" (p. 50). Test developers using this definition need to determine how to operationalize the "automatic" nature of proficient reading. One possibility is to investigate the effect timing has on the item difficulty of the passages used in a reading proficiency test.

In this study, a Spanish reading proficiency test was developed and administered in both untimed and timed testing conditions. The test consisted of 28 passages representing three different

ACTFL proficiency levels (Intermediate, Advanced, and Superior) with a single question for each passage. In the untimed testing condition, examinees could take as much time as they needed to complete the test. In the timed testing condition, for each passage and question examinees were give one minute to complete the Intermediate, two minutes for the Advanced, and four minutes the Superior. If the examinee failed to answer the question within the allotted time, it was marked wrong and the examinee was presented with the next question. To ensure the examinees were doing their best, this test was administered at the same time examinees (n» 1,000) were taking a challenge exam to test out of and receive college credit for lower division Spanish classes. We will present the effect that the timing condition had on the item difficulties and how examinees of different reading proficiency levels performed on items/passages at the three different ACTFL levels. Implications for test developers will then be discussed.

14. Post-enrollment English language assessment and support for international graduate students: Is it worth the effort?

10:45–12:00 · Mario Laserna—2nd floor—Calle del Saber

Sara T. Cushing, Georgia State University, USA

Lisa Armistead, Georgia State University, USA

David Chiesa, Georgia State University, USA

The growing number of international graduate students at US and other English-medium universities has put pressure on institutions to identify and support students who are at risk linguistically (Murray, 2016), even if they have met the minimum admission standards on international tests such as TOEFL or IELTS. While many universities have implemented in-house post-enrollment testing and programs to support students at risk, there is little empirical evidence demonstrating that such programs are worth the resources required to administer and score in-house tests and to staff language support courses that may not count towards students' degrees. In the context of limited resources and increased administrative pressure to retain and graduate students in a timely fashion, such evidence is critical to justifying these activities. In this paper we use data from five years of admission at a large public university to explore the effectiveness of post-admission testing and placement into English language support courses. Our research questions were: (a) what is the relationship between scores on international standardized tests (TOEFL and IELTS) and an in-house test of reading, writing, speaking and listening; (b) what is the relationship between scores on these tests and student success data (specifically, GPAs after the first and second semesters, and retention beyond the first year); (c) what are the effects of taking recommended language support courses on these outcome measures? We examined test scores and performance indicators from 831 international graduate students who entered the university between 2012 and 2015. Preliminary results show that (a) the correlations between TOEFL/IELTS and our in-house test are robust (.r = .72, $p < .01$ with IELTS; r = .70, p <. 01 with TOEFL); (b) the correlation between TOEFL and GPAs at the end of the first ($r = .14$, $p < .01$) but not the second ($r = .10$, $p > .05$) semester was small but statistically significant, whereas the correlations between our in-house test and GPAs were higher for the first ($r = .21$, $p < .01$) and second ($r = .23$, $p < .01$) semester GPAs; and (c) for those students who took recommended language support courses (99% of those referred), correlations between

the in-house test and GPAs were non-significant, suggesting that the courses were effective in reducing or eliminating language-related challenges to student success. We supplement these quantitative results with interview data from graduate directors and students regarding their experiences with post-admission assessment and coursework. We will discuss policy implications for post-admission testing and language support for international graduate students.

15. English language assessment in Colombia: A teacher's perspective

10:45–12:00 · Mario Laserna—2nd floor—Calle del Saber

Patricia Lorena Bustos González, Universidad de los Andes, Colombia

This study aimed at examining how a sample of Colombian English language teachers currently use language assessment within the classroom, specifically, on two major aspects: the purposes of English language assessment, and English language assessment practices (traditional assessment, alternatives in assessment, scoring and feedback). The sample corresponds to 32 Colombian English language teachers who answered a 30 question online. The questionnaire was divided into the aspects mentioned above. The intention was to include teachers who worked at any educational level (primary or secondary school, university (undergraduate), university (postgraduate) or at a language school).

The results suggest that teachers in most of the cases develop assessment instruments by using their own knowledge and experience, in order to obtain information on the students' progress, to provide them with feedback during the learning process, and to determine their final grades. With respect to assessment practices, teachers make use of a wide range of tasks in order to assess their students in the classroom. The traditional types of assessment like written and oral tests remain as the most preferred options by teachers, with other innovative instruments. With regard to the alternatives in assessment, teachers claim to use most of them, such as portfolios, projects, self- and peer-assessment, in different ways and times during the teaching process.

Based on the results provided by this research, it can be said that teachers have clear understanding of what assessment means and what it entails for stakeholders. However, an aspect that this research evidenced is that training in this field, in undergraduate and graduate programs is also needed. Most teachers make use of their intuition and experience to create tests. Item writing techniques and LAL training is a must. If teachers are able to understand the importance of assessment in the teaching practice, they will also comprehend the great impact of good assessment instruments will have in teaching. A further important aspect of this study is that many of the previous research has focused on one skill or area of assessment. In this particular research, its nature was to cover a wide range of topics from the point of view of the teachers. Finally, the findings may serve as an evaluation to report on important decisions or strategies that can be made by curriculum designers, educational administrators, policy makers, and teacher trainers to develop insights into ways of assessment practices that can be improved in Colombia and in Latin America.

16. Using an argument-based framework for language program evaluation: A case study

10:45–12:00 · Mario Laserna—2nd floor—Calle del Saber

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

Mónica Cárdenas-Claros, Pontificia Universidad Católica de Valparaíso, Chile

Katherine Rick, Lincoln College International, Saudi Arabia

Language program evaluation can be a challenging enterprise for evaluators because of differences in evaluation contexts, criteria, levels, outcomes, types of available data, and stakeholders (Norris, 2016). An increased blending of technologies in language instruction makes such evaluation even more challenging due to the wide range of factors involved, varying from micro-level (e.g., students' adoption of specific tools) to macro-level aspects (e.g., university infrastructure and policies regarding the use of technology). To help evaluators address these challenges, we have adopted an argument-based framework (Chapelle, Enright, & Jamieson, 2008; Kane, 2006) for evaluating language programs at three levels of focus: micro (classroom), meso (department), and macro (institution). The proposed framework comprises four stages (i.e., planning an argument-based evaluation, gathering the evidence, presenting an argument, and appraising the argument) and can be used to evaluate the considerations of sustainability, purpose, appropriateness, and multimodality (Gruba & Hinkelman, 2012) in language programs.

In this presentation, we will demonstrate the application of an argument-based framework for a meso-level evaluation of blended learning in the English Language Institute (ELI) at a university in the Pacific region. Conducted in Spring 2015, this evaluation project covered 13 sections of ELI courses (131 students) that were taught by ten instructors. Focusing on sustainability of blended learning in the ELI, this evaluation entailed the development of an argument consisting of five inferences (i.e., domain definition, evaluation, explanation, utilization, and ramification) and underlying warrants and assumptions about blended learning in the ELI. To garner evidence supporting the inferences and assumptions in the argument, we conducted semi-structured interviews with two ELI administrators and nine ELI instructors, administered an anonymous online survey (completed by 34 ELI students), and analyzed the ELI documentation. The evidence was subsequently presented as backing for assumptions underlying each of the inferences in the argument. The final stage of this evaluation project involved an appraisal of the argument by (a) assessing the strengths of claims and collected evidence and (b) providing potential rebuttals that could weaken the claims in the argument.

The results of this case study indicate that an argument-based framework provides a flexible and rigorous heuristic for language program evaluation. In particular, this framework offered us (i.e., evaluators) and stakeholders a mechanism to create a coherent narrative that specified the inferences, warrants, and assumptions underlying and guiding the meso-level program evaluation focusing on sustainability of blended learning in the ELI. While the scope of this project did not allow for gathering evidence to support the assumptions associated with the ramification inference, we were able to support assumptions for all other inferences and appraise the strength of claims outlined in our argument. We will conclude our presentation with a discussion of considerations that need to be made when using an argument-based framework for language program evaluation and propose directions for future work in this area.

17. Developing a professional knowledge of language assessment test for EFL teachers

10:45–12:00 · Mario Laserna—2nd floor—Calle del Saber

Hossein Farhady, Yeditepe University, Turkey

Kobra Tavassoli, Karaj Branch, Islamic Azad University, Iran

Within recent developments in education, learners are expected to become more self-regulated and autonomous, and teachers need to move away from traditional transmit of information approach to providing a context where learners would possess, transform, and apply the information to real-life contexts. Conforming to such changes, EFL teachers are expected to move from using traditional testing procedures to applying effective techniques of assessment to enhance learning. Within such a demanding context, teachers need to be provided with opportunities to improve different aspects of their professional knowledge to meet the new requirements.

Knowledge of language assessment as one of the critical parts of teachers' professional knowledge, at its basic level is known as 'assessment literacy' and defined as "having the capacity to ask and answer critical questions about the purpose for assessment, about the fitness of the tool being used, about testing conditions, and about what is going to happen on the basis of the results" (Inbar-Lourie, 2008, p.389). However, in order to help teachers to improve their professional knowledge of language assessment (PKLA), the first step is to have valid and reliable information on their present status regarding the minimum amount of PKLA.

To address the issue, this study was designed to examine EFL teachers' PKLA in a two-phase project. In the first phase, a needs assessment was conducted to explore EFL teachers' perception of the importance of PKLA using a modified version of the needs questionnaire developed by Fulcher (2012) supplemented by some open-ended questions. The findings revealed that the majority of the 246 participants considered major topics in language assessment as either essential or important to be included in language assessment courses. In the second phase, a PKLA test was developed using the information obtained from the needs assessment phase. More specifically, the topics which were identified as essential or important by a significant majority of the participants were selected and checked with the topics in the available language assessment textbooks. The test was developed in six parts each focusing on one major area of language assessment with closed-item formats including 'matching', 'ordering', and 'multiple-choice'. After having the test reviewed by experts in the field, it was revised and then piloted with 50 EFL teachers with diverse ages and experiences. The test was finalized after making necessary revisions based on the results of piloting, and administered to a group of more than 100 EFL teachers. The preliminary analysis of the data revealed that contrary to teachers' claims in the first phase of the study, the majority of them had low levels of PKLA. Furthermore, investigating the relationship between teachers' PKLA scores and certain factors such as their age, gender, experience, and academic degree led to some meaningful findings that will be explained in detail along with implications and applications of the study to teacher education programs.

18. Assessment of ESL sociopragmatics for informing instruction in an academic context: From Australia to Canada

10:45–12:00 · Mario Laserna—2nd floor—Calle del Saber

Valeria Kolesova, University of Ottawa, Canada

This mixed methods study aimed to provide some validity evidence for the use of the ESL sociopragmatics test developed by Roever, Elder, and Fraser (2014) for formative purposes. The test developers recommend further validation of the tool, originally developed for the Australian context. In this study, the test items were used to reveal areas of weakness in sociopragmatic knowledge in a group of learners of an academically oriented English Intensive Program in Canada. Analysis of the test scores revealed a lack of knowledge of norms of appropriateness and politeness in English, which was further targeted with an instructional unit informed by the items of the test. Two weeks after the instructional unit was delivered, the participants were asked to complete a follow-up questionnaire. The questionnaire results provided insight into the participants' perceptions of usefulness of the instructional unit. The learners found explicit instruction on ESL sociopragmatics useful for their language learning experience as well as day-to-day interactions in English. Particularly, they claimed to feel more confident communicating in English after receiving explicit instruction on ESL sociopragmatics. They were able to use information from the lesson in situations such as talking to their language instructors, communicating with university personnel, and participating in service encounter interactions. Therefore, the test proved to have potential for developing instructional materials in an academic context. Based on the findings of the study, suggestions on incorporating sociopragmatic competence into the institution's EAP curriculum were made.

INDEX OF PRESENTERS

Congratulations to LTRC 2017 Bogotá from the Assessment and Evaluation Language Resource Center (AELRC) at Georgetown University!

The AELRC provides **research, resources, workshops** and **online courses** to support foreign language educators' capacity to engage in useful language assessment and program evaluation. Visit us online to learn more!

https://aelrc.georgetown.edu/

@AELRCDC

facebook.com/AELRCDC/

AELRC
Assessment and Evaluation Language Resource Center

The AELRC is a Title VI Language Resource Center funded by the U.S. Department of Education

Interested in an MA or PhD in language testing and assessment?

Lancaster University

Think Lancaster University

Lancaster University has a world-wide reputation for excellence in research on language testing and assessment. Situated within the Department of Linguistics and English Language, our Language Testing Research Group (LTRG) includes Tineke Brunfaut, Luke Harding and John Pill, as well as honorary members Charles Alderson and Dianne Wall.

Lancaster offers a range of postgraduate degrees delivered on-campus or by distance, including our flagship MA in Language Testing.

- **MA in Language Testing (distance)**
  Unique part-time, online Masters programme in Language Testing
- **PhD in Linguistics by Research Only**
  Study full- or part-time, at Lancaster or off-site
- **PhD in Applied Linguistics by Thesis and Coursework**
  Study full- or part-time, at Lancaster or off-site with residential visits

We also offer an annual summer school every July-August: Language Testing at Lancaster

Other postgraduate degrees where language testing and assessment modules are taught include:

- MA TESOL (distance)
- MA in Applied Linguistics and TESOL
- MA in Language and Linguistics

3rd in the UK for Student Satisfaction
Complete University Guide 2017

15th in the World for Linguistics (academic reputation)
QS World University Rankings 2017

5th in UK for Linguistics
Times and Sunday Times Good University Guide 2017

95% of students in work/study 6 months after graduating
Unistats

Further information:
www.lancaster.ac.uk/linguistics/study
wp.lancs.ac.uk/ltrg/

Contact: postgraduatelinguistics@lancaster.ac.uk | Tel: +44 (0)1524 593028 | LU_LanguageTesting

BRITISH COUNCIL Aptis
Forward thinking English testing

The British Council Assessment Research Awards and Grants Results for 2017

The British Council Assessment Research Awards and Grants recognise achievement and innovation within the field of language assessment and form part of the British Council's extensive support of research activities across the world.

Assessment Research Awards

These awards are designed to assist research students in their studies or in presenting their work at an international conference. The maximum award given is £2,500.
Awardees for 2017 are:

Alun Evan Meredydd Roger (University of Bedfordshire, UK, supervisor Dr. Fumiyo Nakatsuhara)

Carolyn Westbrook (University of Bedfordshire, UK, supervisor Professor Anthony Green)

Qie Han (Teachers' College, Colombia University, USA, supervisor Professor James E Purpura)

Salomé Villa Larenas (Lancaster University, UK, supervisor Dr. Tineke Brunfaut)

Assessment Research Grants

This grant scheme is designed to support projects that are directly focused on Aptis, the British Council's English assessment tool. The maximum grant given is £17,500.

The following people have been awarded grants for 2017:

Trisevgeni Liontou (University of Athens, Greece) for her project which will apply automated analyses techniques to investigate discourse features in the Aptis for Teens writing test. Evidence will be taken from lower secondary EFL students.

Nathaniel Owen (Open University, UK) for his project to explore rater behaviour with test-taker responses in Aptis writing.

Okim Kang (Northern Arizona University, USA) for his project which will look at linguistic features and automatic scoring of Aptis speaking performances.

Carol Spöttl, Eva Konrad, Franz Holzknecht, Matthias Zehentner (Language Testing Research Group, University of Innsbruck) for their project which will look at assessing writing at lower levels. They will explore task development locally and internationally, and the opportunities presented by the extended CEFR descriptors.

Azlin Zaiti Zainal, Ng Lee Luan, Tony Green (Faculty of Languages and Linguistics, University of Malaya, Malaysia) for their project which will explore the impact of the ProELT 1 training programme and Aptis on Malaysian English teachers' classroom practice.

International Assessment Award

This award recognises an individual working for the promotion of excellence in language assessment internationally. This year's award is presented to Dr. Rukmini Banerji.

Rukmini Banerji is the CEO of Pratham, one of the largest education NGOs in the world. She is credited with transforming the public discourse on assessments in education in India by linking literacy and numeracy assessments to accountability of the public funded primary education system of the country, the largest in the world.

She has been the driving force behind the Annual Status of Education Report, or ASER (the acronym means "impact" in Hindi) since 2005. She was directly involved in the design and implementation of ASER, the largest annual study ever done by Indian citizens to monitor the status of elementary education in the country. ASER surveys about 600,000 children in more than 16,000 villages in 570 rural districts of India. The participation included over 25,000 to 30,000 volunteers and 500 partner organizations and educational institutions.

Initially trained as an economist (St. Stephen's College, Delhi University and Delhi School of Economics) Rukmini was a Rhodes Scholar at Oxford University between 1981 and 1983. In 1991, she completed her PhD at the University of Chicago's Education department. For the next several years, she worked in Chicago, first at the Population Research Center of the University of Chicago and later as a Program Officer at the Spencer Foundation.

Rukmini has published extensively on issues in education in both Hindi and English. She is also an enthusiastic writer of imaginative and much loved children's stories, published by Pratham Books.

Assessment Research Awards and Grants Key dates for 2018

Call for proposals:	**October 2017**
Closing date:	**30 January 2018**
Winners announced:	**March 2018**

www.britishcouncil.org/aptis

Canada's leader in English language testing

Paragon emphasizes research-lead assessment design with the aim of being a world-class participant in the field of language testing.

Recent research at Paragon includes:

- Li, Z., Banerjee, J. & Zumbo, B. (2017). Response time data as validity evidence: Has it lived up to its promise, and if not, what would it take to do so. In Zumbo, B. and Hubley, A. (Eds.). Understanding and Investigating Response Processes in Validation Research (pp. 159-177), Cham, Switzerland: Springer Press.
- Li, Z., Chen, M., & Banerjee, J. (March, 2017). A corpus-based study on the gender differences in writing performances. Paper presented at the annual conference of the American Association for Applied Linguistics (AAAL), Seattle, WA.
- Li, Z. & Lodge, S. (May, 2017). Disciplinary differences in university lecture slides as a part of classroom discourse: findings from corpus-based analysis and multimodal analysis. Paper presented at the annual conference of the Canadian Association for Applied Linguistics (CAAL), Toronto, ON.
- Stone, J., & Zumbo, B. D. (2016). Validity as a pragmatist project: A global concern with local application. In V. Aryadoust & J. Fox (Eds.), Trends in Language Assessment Research and Practice: The View from the Middle East and the Pacific Rim (pp. 555–573). Cambridge, UK: Cambridge Scholars Publishing, UK.
- Volkov, A., Li, Z. & Banerjee, J. (May, 2017). Speaking proficiency lies in the eye of beholder? A comparison of rating and justifications from university instructors and experienced raters. Paper presented at the Canadian Association of Language Assessment (CALA) Summer Symposium, Toronto, ON

For more information about Paragon's research, visit:

paragontesting.ca/research

TESTING ENTERPRISES Paragon | UBC

a subsidiary of
THE UNIVERSITY OF BRITISH COLUMBIA

LTRC 2017

Cambridge Michigan Language Assessments.
More than 60 years of successful English language testing worldwide.

CaMLA Working Papers

Investigating Lexico-grammatical Complexity as Construct Validity Evidence for the ECPE Writing Tasks: A Multidimensional Analysis.

Yan, X. & Staples, S. (2016).

CaMLA Working Papers. Ann Arbor, MI: Cambridge Michigan Language Assessments

The CaMLA Speaking Test: Face-to-Face vs. Audio Delivery.

Porter-Szucs, I., Macknish, C., & DeCicco, B. (in press).

CaMLA Working Papers. Ann Arbor, MI: Cambridge Michigan Language Assessments

CaMLA Research Publications (Non-Working Papers)

Linking the Common European Framework of Reference and the Michigan English Language Assessment Battery.

CaMLA (2017).

Ann Arbor, MI: Cambridge Michigan Language Assessments

Publications Using CaMLA Data

What's in a Topic? Exploring the Interaction Between Test-taker Age and Item Content in High-Stakes Testing.

Banerjee, J. & Papageorgiou, S. (2016).

International Journal of Listening, 30(1-2), 8-24.

Working with sparse data in rated language tests: Generalizability theory applications.

Lin, C. (2016).

Language Testing, 34(2), 271-289.

Fairness and Bias in Language Assessment.

Verhelst, N., Banerjee, J., & McLain, P. (2016). In D. Tsagari & J. Banerjee (Eds.).

Contemporary Second Language Assessment (243-260). London: Bloomsbury Academic.

Research and Internship Opportunities

Visit the CaMLA website to learn more about summer internships in language assessment. Funding is available on a limited basis for research related to CaMLA language tests. Grant funding is awarded under the Spaan Research Grant Program; project reports are published as part of the CaMLA Working Papers.

CaMLA

CAMBRIDGE ENGLISH
Language Assessment
Part of the University of Cambridge

UNIVERSITY OF MICHIGAN

CambridgeMichigan.org

CAMBRIDGE UNIVERSITY PRESS

CAMBRIDGE ENGLISH
Language Assessment
Part of the University of Cambridge

Research for All

Studies in Language Testing

An indispensable resource for anyone interested in new developments and research in language testing

To find out more about our full list of publications:

cambridge.org/elt/silt
cambridgeenglish.org/silt

Where to next...?

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND

LTRC 2018

Testing and assessment in times of movement, transition, and change

Auckland – New Zealand
2 – 6 July 2018

- **Enjoy the contemporary and stylish venue**

- **All the professional rewards an LTRC can offer**

- **Explore one of the most popular tourist destinations in the world**

See you there next year!

ETS TOEFL

Quality Beyond Measure.

More than four decades of research supports the validity of the *TOEFL*® Family of Assessments

Read the latest journal articles and reports on the *TOEFL*® Family of Assessments.

A Study of the Use of *TOEFL iBT*® Test Speaking and Listening Scores for International Teaching Assistant Screening. (2016). Wagner, E. *TOEFL iBT Research Report No. 27*. Princeton, NJ: ETS.

Chinese Users' Perceptions of the Use of Automated Scoring for a Speaking Practice Test. (2016). Xi, X., Schmidgall, J., & Wang, Y. In G. Yu & Y. Jin (Eds.), *Assessing Chinese Learners of English: Language Constructs, Consequences and Conundrums* (pp. 150–175). New York, NY: Palgrave McMillan.

Construct Validity in *TOEFL iBT*® Speaking Tasks: Insights From Natural Language Processing. (2016). Kyle, K., Crossley, S. A., & McNamara, D. S. *Language Testing*, *33*(3), 319–340.

Examining Content Representativeness of a Young Learner Language Assessment: EFL Teachers' Perspectives. (2016). Hsieh, C. In M. Nikolov (Ed.), *Assessing Young Learners of English: Global and Local Perspectives* (pp. 93–108). Cham, Switzerland: Springer.

Exploring the Relationship of Organization and Connection with Scores in Integrated Writing Assessment. (2017). Plakans, L., & Gebril, A. *Assessing Writing*, *32*, 98–112.

Is Writing Performance Related to Keyboard Type? An Investigation From the Examinees' Perspectives on the *TOEFL iBT*®. (2017). Ling, G. *Language Assessment Quarterly*, *14*(1), 36–53.

Making a Validity Argument for Using the *TOEFL Junior*® Standard Test as a Measure of Progress for Young Language Learners. (2017). Gu, L., Lockwood, J. R., & Powers, D. E. In M. K. Wolf & Y. G. Butler (Eds.), *English Language Proficiency Assessments for Young Learners* (pp. 153–170). New York, NY: Routledge.

Setting Language Proficiency Score Requirements for English-as-a-Second-Language Placement Decisions in Secondary Education. (2016). Baron, P. A., & Papageorgiou, S. *Research Report No. RR-16-17*. Princeton, NJ: ETS.

Using the Common European Framework of Reference to Facilitate Score Interpretation for Young Learners' English Language Proficiency Assessments. (2017). Papageorgiou, S., & Baron, P. A. In M. K. Wolf & Y. G. Butler (Eds.), *English Language Proficiency Assessments for Young Learners* (pp. 136–152). New York, NY: Routledge.

What and When Second-language Learners Revise When Responding to Timed Writing Tasks on the Computer: The Roles of Task Type, Second-language Proficiency, and Keyboarding Skills. (2016). Barkaoui, K. *The Modern Language Journal*, *100*(1), 320–340.

ETS — *Measuring the Power of Learning.*®

www.ets.org/toefl/research