The British Council Assessment Research Awards and Grants

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 2018 are:

- **Glyn Jones** (Lancaster University, UK, supervisor Dr Luke Harding)
- **Roman O. Lesnov** (Northern Arizona University, USA, supervisor Professor Joan M. Jamieson)
- **Olena Rossi** (Lancaster University, UK, supervisor Dr Tineke Brunfaut)
- **Sonca Vo** (Iowa State University, USA, supervisor Professor Gary Ockey)
- **Jorge Luis Beltran Zuniga** (Teachers College, Columbia University, USA, supervisor Professor James E. Purpura).

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 2018:

- **Sathena Chan, Daniel Lam and Tony Green** (CRELLA, University of Bedfordshire) for their project which will investigate the textual features and revising processes of EFL and L1 English writers in China in Aptis for Teens Writing Task 4.
- **Nivja de Jong and Jos Pacilly** (Leiden University Centre of Linguistics, Leiden University, Holland) for their project which will look at new techniques to measure fluency in speech automatically.
- **Ute Knoch and Catherine Elder** (University of Melbourne, Australia) for their project to investigate the discourse produced at score levels B2.2 to C2 on the Apts Advanced Writing Test.
- **Judit Kormos** (Department of Linguistics and English Language, Lancaster University) for her project which will explore time-extension and the second language reading performance of children with different first language literacy profiles.

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 Carol Spoettl**.

**Carol Spoettl** is a Senior Lecturer at the School of Education, University of Innsbruck and Head of the Language Testing Group Innsbruck (LTRGI). Carol has dedicated most of her professional life to the promotion of best practice in language assessment, both nationally and internationally. Most notably, she single-handedly initiated and steered the reform of the school-leaving examination in Austria, leading to the change of an entire country’s examination system. Carol has also contributed to the professional and academic community by serving on the Executive Committees of EALTA and the IATEFL TEASIG. She has published her work in many book chapters and presented it at numerous international conferences, including IATEFL, EALTA and EUROSLA.

Innovation in Assessment Prize

The Innovation in Assessment Prize is awarded to an individual or organisation for innovation in the field of language assessment. This year’s winner is **Tobias Haug**, Professor at the University of Applied Science of Special Needs Education (HH), for developing the Automatic Sign Language Recognition System for Swiss-German sign language vocabulary assessment.

Assessment Research Awards and Grants

Key dates for 2019

- Call for proposals: October 2018
- Closing date: 30 January 2019
- Winners announced: March 2019

www.britishcouncil.org/aptis
# TABLE OF CONTENTS

Welcome Message from the ILTA President ................................................................. 3

Welcome Message from the Chair of the LTRC 2018 Organising Committee ........ 5

Conference Organisation ................................................................................................. 7

ILTA 2018 ......................................................................................................................... 8

Award winners ................................................................................................................... 9

LTRC 2018 General Information ..................................................................................... 12

LTRC 2018 Main Conference Venue ............................................................................... 13

University of Auckland City Campus ........................................................................... 14

Auckland CBD Map .......................................................................................................... 15

Restaurants and cafés on campus .................................................................................. 16

Restaurants, cafés, and bars in the CBD and on the Waterfront ................................... 17

Explore Auckland ........................................................................................................... 18

LTRC 2018 Conference Overview .................................................................................. 19

LTRC 2018 Conference Schedule .................................................................................. 21

Pre-conference events ...................................................................................................... 33

Roundtable ....................................................................................................................... 33

Pre-conference workshops .............................................................................................. 34

LTRC 2018 Keynote Speakers ......................................................................................... 39

The Alan Davies Lecture ................................................................................................. 39

The Samuel J Messick Memorial Lecture ....................................................................... 41

The Distinguished Achievement Award Lecture ......................................................... 43
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symposia</td>
<td>45</td>
</tr>
<tr>
<td>Research Papers</td>
<td>59</td>
</tr>
<tr>
<td>Wednesday 4 July</td>
<td>59</td>
</tr>
<tr>
<td>Thursday 5 July</td>
<td>76</td>
</tr>
<tr>
<td>Friday 6 July</td>
<td>86</td>
</tr>
<tr>
<td>Demo Presentations</td>
<td>95</td>
</tr>
<tr>
<td>Posters</td>
<td>97</td>
</tr>
<tr>
<td>Work in Progress Session</td>
<td>109</td>
</tr>
<tr>
<td>Index of Presenters</td>
<td>122</td>
</tr>
</tbody>
</table>
Welcome Message from the ILTA President

Cathie Elder
University of Melbourne, Australia

As President of the International Language Testing Association (ILTA), I am delighted to welcome you to Aotearoa, Land of the Long White Cloud, the home of my forbears, and to the University of Auckland, where I spent four happy years teaching applied linguistics and doing the foundational work for the now flourishing Diagnostic English Language Needs Assessment program (DELNA). DELNA, along with The British Council, Cambridge Assessment English, ETS-TOEFL, Language Training and Testing Center (Taiwan), Language Learning and Testing Foundation, Occupational English Test, Paragon and Pearson, is one of the sponsors of this conference. On behalf of ILTA I would like to heartily thank all of our sponsors for their generous support.

This is the 40th Language Testing Research Colloquium (LTRC) and only the second time the conference has been held in the southern hemisphere. Those of us who live on this side of the world are used to travelling long distances to the LTRC, so we do understand and have sympathy for the jetlag that many of you “northerners” will be suffering. We trust that the stimulating program which John Read, this year’s Conference Chair, and his fellow committee members (Aek Phakiti, Judit McPherson, Karen Ashton, Martin East, Morena Botelho de Magalhães, Peter Gu, Peter Keegan, Rosemary Erlam, Stephanie Rummel, Susy Macqueen, and Ute Knoch) have put together will help to recharge you. A huge amount of thought and effort has gone into organizing the conference and we are immensely grateful to the team for their work and for the benefits this event promises for ILTA members.

This year’s LTRC is organized around the broad themes of Movement, Transition and Change. These themes truly reflect ILTA’s mission to bring together language testers from around the world to grapple with language assessment issues that arise in catering for global movements in population, in making decisions at entry and exit points in educational systems and society, and in monitoring or facilitating language development and innovation at both classroom and institutional levels.

Don’t miss the plenary talks from Pamela Moss, our ETS-sponsored Messick speaker, from Joe Lo Bianco, who will deliver the British Council-sponsored Alan Davies lecture and from Carolyn Turner, this year’s winner of the Cambridge/ILTA Distinguished Achievement Award. These talks offer very different takes on our roles and responsibilities as language testers in interpreting score meaning, in making policy decisions and in devising appropriate methodologies for our research. As well as choosing what entices you most from the rich array of conference papers on the program, please also make a point of supporting those attendees who have travelled from afar to get your feedback on their posters and work-in-progress presentations. One of the nicest things about the LTRC is its warm welcoming atmosphere and the sense of community that it engenders, so I invite you to carry on the tradition in whatever way you can.

If this is your first time at LTRC, do come along to the Language Learning and Testing Foundation-sponsored Newcomers’ Session on Tuesday 3 July at 4.00pm before the conference begins. You will find out more about ILTA and meet others like you who may be keen to exchange information, socialize and make the most of the conference experience. And if you are a graduate student, drop in to the student gathering organized by the ALTAANZ student representatives happening on Friday 6 July during the lunch hour. I look forward to catching up with both old and new friends at these events and at the Banquet on Friday night where we will have the chance to celebrate and honour the winners of our annual awards.
What of ILTA? We have held our annual jointly hosted symposia at AAAL and TESOL in Chicago and continued with our program of annual awards. As part of our mission to enhance language testing practice and build language assessment literacy around the world, we have solicited translations of the ILTA Code of Ethics into new languages, initiated a review of the ILTA Guidelines for Practice and of our Meetings and Workshop award and posted new resources on our ILTA Website. We have drafted guidelines for Special Interest Groups (SIGs) and have two new SIGs in the offing. We have also expanded our membership and facilitated, through our ILTA-funded awards program, the creation of a network of assessment professionals in Cuba, an email discussion group on language assessment in Turkey and the creation of a Latin American Association for Language Testing and Assessment (LAALTA). Over the past year, we have also been doing a lot of thinking about new directions for our organization. We have formed four ad hoc committees to review our membership, finance, communications and public engagement activities in advance of a more formal strategic planning session to be held in Atlanta in 2019. As President, I urge all ILTA members present to come to the ILTA Annual Business Meeting at lunchtime on Thursday 5 July to express your views on what we are thinking and doing, and to hear of our plans for future LTRCs. If you have not yet joined ILTA you can do so online at www.ildaonline.com/page/MembershipPlans or face-to-face at the Conference Reception desk, where you can say hi to Christine Kadi from our management company Nardone.

Finally, I recommend that you make the most of your stay in Auckland and brave the weather to enjoy the best of this lively and attractive city and its surrounds.

Cathie Elder

President, International Language Testing Association (ILTA)
Welcome Message from the Chair of the LTRC 2018 Organising Committee

Naumai, haere mai ki Tamaki Makaurau!
Come, welcome to Auckland!

It is my great pleasure to welcome you to this, the 40th Language Testing Research Colloquium, being held in New Zealand for the very first time and for only the second time in the Southern Hemisphere. Since it first ventured beyond North America in 2003, the conference has taken place in many parts of the world and given people in each country and region the opportunity to experience this annual gathering of specialists in language testing and assessment that so many of us cherish. I feel privileged to have been able to participate in more than half of the forty LTRCs to date, and I am delighted to be able to bring the conference here on this occasion, to make it possible for those in New Zealand, Australia and the wider Pacific region to take part.

We chose as the theme for this year’s LTRC “Language Testing and Assessment in Times of Movement, Transition and Change”. We saw this as capturing something of the spirit of the times we are living in, while at the same time offering a broad umbrella under which a whole range of contemporary work in the field could be presented. More particularly, it echoes the theme of Sociolinguistics Symposium 22 (SS22), the conference being held at the same venue a week earlier than ours, “Crossing Borders: South, North, East, West”. Increasingly, language testers need to address the ways in which our assessments can facilitate – but all too often impede or prevent – the passage of individuals and whole populations across boundaries in a geographical, social or educational sense.

It is quite a recent innovation for LTRC to have an opening session on Tuesday evening. On this occasion, we see the symposium organised by Luke Harding as representing a kind of conceptual bridge from SS22 to LTRC, bringing sociolinguistic concerns to the forefront on the topic of Transition, Mobility, Validity: English as a (Multi)lingua Franca Perspectives on Language Assessment. The discussant for the symposium, Joe Lo Bianco, will be back on Wednesday morning to deliver his Davies Lecture, sponsored by the British Council, on the role of language tests in national language policies. In the afternoon the sociolinguistic strand continues with two parallel symposia: Claudia Harsch’s one on The Role of Language Tests for Immigration, Integration and Participation, and the other chaired by Albert Weideman on Perspectives from the South on Academic Literacy Assessment.

At the heart of our field is the multi-faceted concept of validity, and our programme on Thursday morning is devoted to a re-consideration of the concept from various points of view, beginning with the symposium chaired by Spiros Papageorgiou on principles of test validation, and proceeding to the ETS-sponsored Messick Lecture by Pamela Moss on knowledge infrastructures in measurement.

In fact, we were delighted at the number of excellent proposals we received for symposia, and have been able to accommodate two more on Friday morning. One chaired by Han Yu focuses on the exciting new China’s Standards of English Language Ability project, which promises to transform English language education in China. The other symposium, organised by Bill Bonk and Ali van Moere, will offer insights on the longitudinal effects of taking a test more than once.
We will gather on Friday afternoon in the closing plenary session for the keynote lecture delivered by Carolyn Turner as the winner of the Cambridge/ILTA Distinguished Achievement Award. This is an opportunity to honour one of our most respected colleagues and to hear her reflections on the evolution of research methodology in our field.

Along the way we have scheduled 48 research papers and demos that came through our rigorous review procedure. It is particularly pleasing to see the number of graduate student papers that were accepted this time and that will keep the selection committee busy in their work of choosing the winner of the Lado Award for the best student paper. We have also continued the LTRC tradition of giving a prominent place to posters and work-in-progress presentations, in order to widen participation in the conference programme and to give opportunities for informal discussion and networking.

Needless to say, we have included all the other components of an LTRC programme: the pre-conference workshops, the social events, including the Banquet in the Fale Pasifika, the business meetings, and the welcome to newcomers. I particularly appreciated the initiative of MaryAnn Christison and Buzz Palmer in organising the Language Learning sponsored Roundtable on Connecting Language Testers and Specialists in L2 Pedagogy as the first event of the conference week.

No one can organise a conference like this singlehandedly, and I must acknowledge the multiple contributions by members of the organising committee. In many ways Ute Knoch has acted as a co-chair, by taking on the major responsibility of managing the review of proposals and putting together the conference schedule. She has also been one of the three ILTA officers, along with Cathie Elder and Jay Banerjee, who have been constant sources of advice and support all along the way. To a large degree we have been able to implement the kind of close cooperation between the LTRC organising committee and the ILTA Board which was envisaged by the provision for an LTRC Advisory Committee in the last major revision of the ILTA By-Laws in 2013. I have also benefitted greatly from the professional services provided by Donna Heavener, Michele Doyle and Christine Kadi at the ILTA Business Office and by our Event Manager at the University of Auckland, Stephanie Szmurlo.

We hope you gain as much pleasure and intellectual stimulation from participating in the colloquium as we have had in organising it. And be sure to enjoy as much of our beautiful city as you can.

John Read

on behalf of the LTRC 2018 Organising Committee
Conference Organisation

LTRC Organising Committee 2018

Chair
John Read, University of Auckland, New Zealand

Other Members of the Organising Committee

Karen Ashton (Massey University, New Zealand)
Morena Botelho de Magalhães (University of Auckland, New Zealand)
Martin East (University of Auckland, New Zealand)
Rosemary Erlam (University of Auckland, New Zealand)
Peter Gu (Victoria University, Wellington, New Zealand)
Ute Knoch (University of Melbourne, Australia)
Susy Macqueen (Australian National University, Australia)
Judit McPherson (University of Auckland, New Zealand)
Aek Phakiti (University of Sydney, Australia)
Stephanie Rummel (University of Auckland, New Zealand)

Student volunteers
Xiaohua Liu (coordinator)
Orlando Chaves Varon
Maria Guapacha Chamorro
Mandana Hazrat
Priscilla Shak
Miyoung Song
Maria Treadaway

University Event Manager
Stephanie Szmurlo (University of Auckland, New Zealand)

LTRC 2018 Programme Book Designer
Judit McPherson (University of Auckland, New Zealand)

Abstract Reviewers

Karen Ashton
Jayanti Banerjee
Khaled Barkaoui
David Beglar
William Bonk
Morena B. de Magalhães
Tineke Brunfaut
Yuko Butler
Nathan T. Carr
Micheline Chalhoub-Deville
Carol Chapelle
Lijing Cheng
Yeonsuk Cho
Ikkyu Choi
Martyn Clark
Christian Colby-Kelly
Deborah Crusan
Alister Cumming
Sara Cushing
Jee Wha Dakin
Larry Davis
John H.A.L. de Jong
Barbara Dobson
Dan Douglas
Martin East
Catherine Elder
Jason Fan
Timothy Farnsworth
Kellie Frost
Glenn Fulcher
Ardeshir Geranpayeh
April Ginther
Tony Green
Peter Gu
Luke Harding
Claudia Harsch
Ching-Ni Hsieh
Ari Huhta
Yo In’nami
Talia Isaacs
Eunice Jang
Yan Jin
Dorry Kenyon
Ute Knoch
Antony Kunnan
Yong-Won Lee
Constant Leung
Gad Lim
Sari Luoma
Susy Macqueen
Meg Malone
Lyn May
Tim McNamara
Fumiyo Nakatsuhara
Heike Neumann
Sally O’Hagan
Barry O’Sullivan
Gary John Ockey
Spiros Papageorgiou
Lia Plakans
India Plough
James Purpura
David D. Qian
Daniel Reed
Steven Ross
Stephanie Rummel
Shahrzad Saif
Nick Saville
Yasuyo Sawaki
Jonathan Schmidgall
Charles Stansfield
Masanori Suzuki
May Tan
Carolyn Turner
Alan Urmston
Alistair Van Moere
Erik Voss
Elvis Wagner
Yoshinori Watanabe
Mikyung Kim Wolf
Guoxing Yu
ILTA 2018

ILTA Executive Board 2018

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

Members at Large:
Benjamin Kremmel (University of Innsbruck, Austria)
Jin Yan (Shanghai Jiao Tong University, China)
Tineke Brunfaut (University of Lancaster, UK)
Claudia Harsch (University of Bremen, Germany)

ILTA Staff
Christine Kadi (Association Manager)
Michele Doyle (Association Assistant)

LTRC 2018 Chair:
John Read (University of Auckland, New Zealand)

ILTA Nominating Committee 2017
Chair: Lia Plakans (University of Iowa, USA)
Noriko Iwashita (The University of Queensland, Australia)
Heike Neumann (Concordia University, Canada)
Ikkyu Choi (Educational Testing Service, USA)

Award Committees

Cambridge/ILTA Distinguished Achievement Award
Chair: Claudia Harsch (University of Bremen, Germany)
Tony Green (University of Bedfordshire, UK)
Cathie Elder (University of Melbourne, Australia)
Nick Saville (Cambridge Assessment English UK)

ILTA Student Travel Awards
Chair: Micheline Chalhoub-Deville (The University of North Carolina, Greensboro, USA)
Tineke Brunfaut (Lancaster University, UK)
Benjamin Kremmel (University of Innsbruck, Austria)

ILTA Workshops and Meetings Award
Chair: Cathie Elder (University of Melbourne, Australia)
Ofra Inbar (Tel Aviv University, Israel)
John Pill (Lancaster University, UK)

Lado Award (to be awarded at LTRC)
Chair: Benjamin Kremmel (University of Innsbruck, Austria)
Beverly Baker (University of Ottawa, Canada)
Ikkyu Choi (Educational Testing Service, US)
Kellie Frost (University of Melbourne, Australia)

ILTA Best Article Award 2016
Chair: Alan Urmston (Hong Kong Polytechnic University, Hong Kong)
Luke Harding (Lancaster University, UK)
Beverly Baker (University of Ottawa, Canada)
Troy Cox (Brigham Young University, USA)
Jin Yan (Shanghai Jiao Tong University, China)
Award winners

**The Davies Lecture Award**
Joseph Lo Bianco (University of Melbourne, Australia)
Lecture title: *No Policy without Testing! How the Language of Policy Persuasion and Persuasive Language Help to Make Testing Count*

**Samuel J. Messick Memorial Lecture Award**
Pamela Moss (University of Michigan, USA)
Lecture title: *Evolving our Knowledge Infrastructures in Measurement/Recovering Messick’s Singerian Approach to Inquiry*

**Cambridge/ILTA Distinguished Achievement Award**
Carolyn Turner (McGill University, Canada)
Lecture title: *The Methodological Evolution in Language Testing/Assessment Research and the Role of Community: A Personal View*

**Robert Lado Memorial Award**
To be announced at the LTRC banquet

**ILTA Student Travel Awards 2018**
**Winners:**
Clarissa Lau (University of Toronto) supervised by Eunice Eunhee Jang
Simon Davidson (University of Melbourne) supervised by Cathie Elder & Ute Knoch

**Jacqueline Ross TOEFL Dissertation Award 2018**
**Winner:**
Bart Deygers, *Assessing High-Stakes Assumptions: A Longitudinal Mixed Methods Study of University Entrance Language Tests and of the Policy that Relies on Them*
Supervised by Kris Van den Branden and Co-supervisor Koen Van Gorp at The University of Leuven

**Finalists:**
Laura Ballard, *The Effects of Primacy on Rater Cognition: An Eye-Tracking Study*
Supervised by Paula Winke at Michigan State University
Supervised by Larry Mikulecky at Indiana University Bloomington
Caroline Clapham IELTS Masters Award, 2017
Martin Stark, Exploring the Relationship between Automated Analyses of Linguistic Features and Human Ratings of Test-Taker Performances on an ESL Writing Task
Supervised by Tineke Brunfaut at Lancaster University, UK

ILTA Workshops and Meetings Award 2018
Mario López-Gopar, Jamie L. Schissel, and Constant Leung
Establishing a Language Assessment Association in Mexico: Focusing on Local Language Learning and Assessment Practices Inclusive of Indigenous Languages and Communities

ILTA Best Article Award 2016
Winner:

Shortlisted articles:

TIRF 2016 Doctoral Dissertation Grant Awardees in Language Assessment
1. Heidi Liu Banerjee, Investigating the Construct of Topical Knowledge in a Scenario-based Assessment Designed to Simulate Real-Life Second Language Use
Supervised by James E. Purpura at Columbia University, USA

2. Alish Biler, The Role of Content Word Overlap in Second Language Reading
Supervised by Nina Moreno at the University of South Carolina, USA

Supervised by Karsten Klenkies at the University of Strathclyde, UK
Supervised by Ute Knoch, Neomy Storch at the University of Melbourne, Australia

5. Özgür Şahan, *The Impact of Rater Experience and Essay Quality on Rater Behavior and Scoring*
Supervised by Salim Razi at Çanakkale Onsekiz Mart University, Turkey

6. Friederike Sell, *Mental Effort in Pragmatic Competence: The Role of Working Memory in L2 Pragmatic Production*
Supervised by Klaus P. Schneider at the University of Bonn, Germany
LTRC 2018 General Information

Conference Venue - Owen G Glenn Building (OGGB), University of Auckland
12 Grafton Rd, Auckland Central
Owen G. Glenn Building is open from 7.00am until 6.30pm

Wifi
Wireless internet access is available for use during the conference.
NETWORK : UoA-Guest-WiFi
USERNAME: ltrc2018@auckland.nz
PASSWORD: HyygaYqh

Nearby Services
Banking and foreign exchange services: ASB Bank, Level 1, Owen G Glenn Building (OGGB) University of Auckland
Pharmacy & Post Shop: Level 1 Kate Edger Information Commons, Corner Alfred St & Symonds Street

Getting Around
There are many taxi companies to choose from in Auckland. We suggest:
- Blue Bubble Taxis
  Tel: 09 300 3000
  Website: www.cooptaxi.co.nz
The conference venue is in a prime location for public transportation. Find details at:
  www.at.govt.nz/bus-train-ferry
  Auckland Transport – download the AT mobile app on your smartphone for live updates

Medical and Illness
If you have an emergency, you can contact the police, paramedics and fire department by calling 111 from any landline or mobile phone.
The Conference Venue is: Owen G Glenn Building, 12 Grafton Road, Auckland Central.
If you require non-emergency medical attention during the conference, please inform the registration desk where a first aid kit will be available.

No Smoking
Smoking is not permitted anywhere within The University of Auckland campus.

Parking
The best place to park for the conference is the Owen G Glenn Building (OGGB).
Hours: Monday to Friday, 6.30am to 11.30pm, Saturday and Sunday, 7am to 11pm.

<table>
<thead>
<tr>
<th>Parking Option</th>
<th>Cost</th>
<th>Hours</th>
<th>Days</th>
</tr>
</thead>
<tbody>
<tr>
<td>Casual hourly parking - weekdays</td>
<td>$5 an hour</td>
<td>6.30am to 5pm</td>
<td>Monday to Friday</td>
</tr>
<tr>
<td></td>
<td>$25 for more</td>
<td>6.30am to 5pm</td>
<td>Monday to Friday</td>
</tr>
<tr>
<td></td>
<td>than 5 hours</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$6 flat rate</td>
<td>After 5pm</td>
<td>Daily</td>
</tr>
<tr>
<td>Casual daily parking - weekends</td>
<td>$6 flat rate</td>
<td>7am to 11pm</td>
<td>Saturday &amp; Sunday</td>
</tr>
</tbody>
</table>
Main Conference Venue

Level 0, Owen G Glenn Building

Use Central Elevators to Access Additional Rooms | Decima Glenn Level 3
University of Auckland City Campus

Workshops | Main Conference | Welcome
Owen G Glenn Building, Level 0

Conference Dinner
Fale Pasifika
Walking distances from the University of Auckland

<table>
<thead>
<tr>
<th>Outlet</th>
<th>Indicative walking distance and time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Revive Cafe</td>
<td>950 m (12 minutes)</td>
</tr>
<tr>
<td>Elliot Stables</td>
<td>1.1 km (15 minutes)</td>
</tr>
<tr>
<td>SkyCity</td>
<td>1.4 km (19 minutes)</td>
</tr>
<tr>
<td>Harbourside Ocean Bar Grill</td>
<td>1.7 km (21 minutes)</td>
</tr>
</tbody>
</table>
# Restaurants and cafés on campus

## Student Quad, 34 Princes St

<table>
<thead>
<tr>
<th>Outlet</th>
<th>Description</th>
<th>Hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barilla</td>
<td>Chinese dumplings</td>
<td>Mon-Fri 8.30am-4pm</td>
</tr>
<tr>
<td>Hello Chinese</td>
<td>Various Chinese dishes</td>
<td>Mon-Fri 9am-4pm</td>
</tr>
<tr>
<td>Hollywood Bakery</td>
<td>Café (sandwiches and baked goods)</td>
<td>Mon-Fri 8am-4pm</td>
</tr>
<tr>
<td>Jewel of India</td>
<td>Indian eatery</td>
<td>Mon-Fri 9am-5pm Sat 10am-3pm</td>
</tr>
<tr>
<td>Moustache Milk &amp; Cookie Bar</td>
<td>Cookie shop</td>
<td>Mon-Fri 11am-4pm</td>
</tr>
<tr>
<td>Relax Café</td>
<td>Café (sandwiches, baked goods, and cooked meals)</td>
<td>Mon-Fri 6am-5pm Sat 8am-2pm</td>
</tr>
<tr>
<td>Tank Juice Bar</td>
<td>Juices, smoothies, salads and wraps</td>
<td>Mon-Fri 8am-7pm Sat-Sun 9am-5pm</td>
</tr>
<tr>
<td>The Quad Café</td>
<td>Café (sandwiches, baked goods, and cooked meals)</td>
<td>Mon-Fri 8am-4pm</td>
</tr>
<tr>
<td>Uni Sushi</td>
<td>Sushi and other Japanese dishes</td>
<td>Mon-Fri 8.30am-4pm</td>
</tr>
</tbody>
</table>

## Kate Edger Information Commons, 11 Symonds St

<table>
<thead>
<tr>
<th>Outlet</th>
<th>Description</th>
<th>Hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indochine</td>
<td>Vietnamese street food</td>
<td>10am-5.30pm</td>
</tr>
<tr>
<td>Munchy Mart</td>
<td>Convenience store</td>
<td>Mon-Thu 8am-9pm Fri 8am-8pm Sat-Sun 10am-3pm</td>
</tr>
<tr>
<td>Shaky Isles Coffee Co</td>
<td>Licensed Café (sandwiches, baked goods, cooked meals, wine and beer)</td>
<td>Mon-Fri 8am-4pm</td>
</tr>
<tr>
<td>Strata Café</td>
<td>Café (sandwiches, baked goods, and cooked meals)</td>
<td>Mon-Fri 8am-4pm</td>
</tr>
<tr>
<td>The Taco Joint (by Mexicali)</td>
<td>Mexican food (tacos, burritos, nachos, salads, and quesadillas)</td>
<td>Mon-Fri 10am-5.30pm</td>
</tr>
</tbody>
</table>

## Arts Quad (formerly HSB Courtyard), 10 Symonds St

<table>
<thead>
<tr>
<th>Outlet</th>
<th>Description</th>
<th>Hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mojo Coffee</td>
<td>Café (baked goods)</td>
<td>Mon-Fri 8am-4pm</td>
</tr>
<tr>
<td>Ha!Poke</td>
<td>Healthy bowls (fruit bowls 8am-12pm; Asian bowls 12pm-6pm)</td>
<td>Mon-Fri 8am-6pm</td>
</tr>
</tbody>
</table>
Restaurants, cafés, and bars in the CBD and on the Waterfront

**CBD**

<table>
<thead>
<tr>
<th>Description</th>
<th>Address / website</th>
</tr>
</thead>
</table>
| Revive Café | Vegetarian/Vegan food outlet serving salads and a number of healthy hot dishes | 33 Lorne Street  
  www.revive.co.nz |
| Elliott Stables | A food complex with 12 eateries | 39 Elliott Street  
  www.elliottstables.co.nz |
| SkyCity | A food and entertainment complex with over 20 bars and restaurants | Corner of Victoria Street and Federal Street  
  https://www.skycityauckland.co.nz/ |

**Waterfront**

<table>
<thead>
<tr>
<th>Description</th>
<th>Address / website</th>
</tr>
</thead>
</table>
| Britomart | A variety of restaurants, cafés, and bars around Auckland’s main transport hub | Quay Street, Auckland City  
  www.britomart.org/eat-and-drink |
| The Viaduct | A variety of restaurants and bars on Auckland’s waterfront | Quay Street, Auckland City  
  www.viaduct.co.nz |
| Wynyard Quarter | Another precinct with bars, cafés, and restaurants on the waterfront | Quay Street, Auckland City  
  www.wynyard-quarter.co.nz |
| Harbourside Ocean Bar Grill | Sophisticated seafood dishes, cocktails and an all-season, panoramic deck  
  The venue for the wine tasting evening on Wednesday, July 4 | Quay Street, Auckland City  
  www.harbourside.co |
Explore Auckland

Walk

Viaduct & Britomart Precincts – Auckland’s urban revival scene, featuring the hottest restaurants, bars and cafes.

Wynyard Quarter – Auckland’s expanding urban revival scene, awash in development and more restaurants. Access via footbridge from the Viaduct.

Auckland Museum – Top of the Auckland Domain, featuring cultural, natural and human history exhibits.

Auckland Art Gallery – A short walk from the conference venue at the south-western corner of Albert Park. Classic and modern art from local and international artists.

Markets

- The French Market held at La Cigale (69 St Georges Bay Road, Parnell) takes place 8am to 1pm on Saturday & Sunday.

- Auckland Night Markets run every evening in different locations. For details visit http://aucklandnightmarkets.co.nz/

Bus (Journey Planner at http://www.at.govt.nz)

Mission Bay/Kohimarama/St. Heliers – Auckland’s “Eastern Bays” are popular hangouts, with cafes and beaches. Best at high tide with a southerly or SE wind direction.

Takapuna – Across the bridge on the “North Shore”, another popular hangout with a fantastic beach, waterfront town, and a freshwater lake. Best in a westerly or SW wind at any tide; lots of watersports gear for hire.

Ponsonby Road – Long strip of independent retailers and restaurants stretching between College Hill and Karangahape Road. Easy access using the Outer Link bus outside the conference venue (head uphill towards the city).

Dominion Road & Sandringham Road – Transport yourself to China, India, Korea, Thailand and everywhere in-between in the shops around Dominion & Valley Road, Dominion & Balmoral Road or Sandringham & Kitchener Road.

Mt. Eden – One of the best views of Auckland CBD can be found atop this volcanic cone, with a lovely village of restaurants, bars and cafes underneath.

Ferry (all boats leave from the Downtown Ferry Terminal)

Devonport/Stanley Bay/Cheltenham – Across the harbour from the CBD, these historic villages are well preserved. Nearby volcanic cones provide stunning views of the harbour and city. Reasonable $40 tour takes you everywhere.

Rangitoto Island – The youngest volcano in Auckland is preserved in its natural state. Walk up to the top for fantastic views and lava caves (bring a torch).

Tiritiri Matangi Island – A full-day trip, but a must-do if you want to meet NZ’s endemic birdlife. The island is a pest-free sanctuary and you’re guaranteed to meet many birds on many well-maintained bushwalking tracks.
# LTRC 2018 Conference Overview

## Monday July 2

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Venue</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.30 am – 5.00 pm</td>
<td><strong>Language Learning Roundtable</strong>: Connecting Language Testers and Specialists in L2 Pedagogy: Issues in Classroom-based Assessment</td>
<td>OGGB, Case Room 3 (260-055)</td>
</tr>
</tbody>
</table>
| 9.00 am – 4.00 pm | **Workshops**  
Workshop 1: A Learning-Oriented Assessment Approach in Classroom Contexts: An Expanded Conceptualization of Performances  
Workshop 2: Mixed Methods Research in Language Assessment Validation Studies  
Workshop 3: A Problem-Solving Approach to Exploratory andConfirmatory Factor Analysis and Structural Equation Modeling | OGGB, Seminar Room B (260-040B)  
OGGB, Seminar Room C (260-040C)  
OGGB, Computer Lab 12 (260-036)  |

## Tuesday July 3

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Venue</th>
</tr>
</thead>
</table>
| 9.00 am – 4.00 pm | **Workshops**  
Workshop 1: A Learning-Oriented Assessment Approach in Classroom Contexts: An Expanded Conceptualization of Performances  
Workshop 2: Mixed Methods Research in Language Assessment Validation Studies  
Workshop 3: A Problem-Solving Approach to Exploratory andConfirmatory Factor Analysis and Structural Equation Modeling | OGGB, Seminar Room B (260-040B)  
OGGB, Seminar Room C (260-040C)  
OGGB, Computer Lab 12 (260-036)  |
| 11.00 am – 3.30 pm | **ILTA Executive Board Meeting** | Arts 2 Building, Room 207-314 |
| 4.00 pm – 5.00 pm | **LRTC Newcomers’ Session** | OGGB Foyer, Level 0 |
| 5.00 pm – 7.00 pm | **Symposium 1**: Transition, Mobility, Validity: English as a (Multi)lingua Franca Perspectives on Language Assessment | OGGB 4 Lecture Theatre (260-073)  |
| 7.00 pm – 9.00 pm | **Opening reception – Welcome to LTRC 2018** | OGGB Foyer, Level 0  |

## Wednesday July 4

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Venue</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.30 – 9.00</td>
<td><strong>Official Opening</strong></td>
<td>OGGB 4 Lecture Theatre (260-073)</td>
</tr>
<tr>
<td>9.00 – 10.00</td>
<td><strong>Davies Lecture</strong>: No Policy Without Testing! How the Language of Policy Persuasion and Persuasive Language Help to Make Testing Count</td>
<td>OGGB 4 Lecture Theatre (260-073)</td>
</tr>
<tr>
<td>10.00 – 10.30</td>
<td><strong>Coffee Break</strong></td>
<td></td>
</tr>
<tr>
<td>10.30 – 11.00</td>
<td><strong>Parallel sessions</strong></td>
<td></td>
</tr>
<tr>
<td>11.05 – 11.35</td>
<td><strong>Parallel sessions</strong></td>
<td></td>
</tr>
<tr>
<td>11.40 – 12.10</td>
<td><strong>Parallel sessions</strong></td>
<td></td>
</tr>
</tbody>
</table>
| 12.10 – 1.30 | **Lunch break**  
**Language Assessment Quarterly (LAQ) Editorial Board Meeting**  
**ALTAANZ Annual General Meeting** | OGGB, Level 3, Decima Glenn Room (260-310)  
OGGB, Case Room 3 (260-055)  |
### Symposium 2: The Role of Language Tests for Immigration, Integration and Participation – From Critical Analyses to Inclusive Approaches

Venue: OGGB 4 Lecture Theatre (260-073)

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Venue</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.30 – 3.30</td>
<td>Symposium 2: The Role of Language Tests for Immigration, Integration and Participation – From Critical Analyses to Inclusive Approaches</td>
<td>OGGB 4 Lecture Theatre (260-073)</td>
</tr>
<tr>
<td>3.30 – 3.50</td>
<td>Coffee break</td>
<td></td>
</tr>
<tr>
<td>3.50 – 4.00</td>
<td>Group photo</td>
<td>OGGB Foyer</td>
</tr>
<tr>
<td>4.00 – 4.30</td>
<td>Parallel sessions</td>
<td></td>
</tr>
<tr>
<td>4.35 – 5.05</td>
<td>Parallel sessions</td>
<td></td>
</tr>
<tr>
<td>5.10 – 5.40</td>
<td>Parallel sessions</td>
<td></td>
</tr>
<tr>
<td>6.30 onwards</td>
<td>Wine Tasting</td>
<td>Harbourside Grill, City</td>
</tr>
</tbody>
</table>

### Thursday July 5

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Venue</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.45 – 10.45</td>
<td>Symposium 4: Re-conceptualizing, Challenging, and Expanding Principles of Test Validation</td>
<td>OGGB 4 Lecture Theatre (260-073)</td>
</tr>
<tr>
<td>10.45 – 11.05</td>
<td>Coffee break</td>
<td></td>
</tr>
<tr>
<td>11.10 – 12.10</td>
<td>Messick Lecture: Evolving our Knowledge Infrastructures in Measurement/Recovering Messick's Singerian Approach to Inquiry</td>
<td>OGGB 4 Lecture Theatre (260-073)</td>
</tr>
<tr>
<td>12.10 – 1.50</td>
<td>Lunch ILTA Annual Business Meeting</td>
<td>Case Room 3 (260-055)</td>
</tr>
<tr>
<td>1.50 – 2.20</td>
<td>Parallel sessions</td>
<td>OGGB Lobby, Level 0</td>
</tr>
<tr>
<td>2.20 – 3.30</td>
<td>Poster session</td>
<td>OGGB Lobby, Level 0</td>
</tr>
<tr>
<td>3.30 – 3.50</td>
<td>Coffee Break</td>
<td>OGGB Lobby, Level 0</td>
</tr>
<tr>
<td>3.50 – 4.20</td>
<td>Parallel sessions</td>
<td>OGGB Lobby, Level 0</td>
</tr>
<tr>
<td>4.25 – 4.55</td>
<td>Parallel sessions</td>
<td>OGGB Lobby, Level 0</td>
</tr>
<tr>
<td>5.00 – 5.30</td>
<td>Parallel sessions</td>
<td>OGGB Lobby, Level 0</td>
</tr>
<tr>
<td>5.45 – 7.00</td>
<td>Assessing Writing Editorial Board Meeting</td>
<td>Arts 2 Building, Room 207-314</td>
</tr>
</tbody>
</table>

### Friday July 6

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Venue</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.45 – 10.45</td>
<td>Symposium 5: China’s Standards of English Language Ability: Impetus for Change in Language Learning, Teaching and Assessment</td>
<td>OGGB 4 Lecture Theatre (260-073)</td>
</tr>
<tr>
<td>10.45 – 11.05</td>
<td>Coffee Break</td>
<td></td>
</tr>
<tr>
<td>11.05 – 12.20</td>
<td>Work-in-Progress Session</td>
<td>OGGB, Team Based Learning Lab (260-040)</td>
</tr>
<tr>
<td>12.20 – 2.00</td>
<td>Lunch Language Testing Editorial Board meeting</td>
<td>OGGB, Level 3, Decima Glenn Room (260-310)</td>
</tr>
<tr>
<td>1.00 – 2.00</td>
<td>Graduate Student Meeting</td>
<td>OGGB, Seminar Room C (260-040C)</td>
</tr>
<tr>
<td>2.00 – 2.30</td>
<td>Parallel sessions</td>
<td></td>
</tr>
<tr>
<td>2.35 – 3.05</td>
<td>Parallel sessions</td>
<td></td>
</tr>
<tr>
<td>3.10 – 3.40</td>
<td>Parallel sessions</td>
<td></td>
</tr>
<tr>
<td>3.40 – 4.00</td>
<td>Coffee Break</td>
<td></td>
</tr>
<tr>
<td>4.00 – 5.00</td>
<td>Cambridge/ILTA Distinguished Achievement Award Winner Lecture: The Methodological Evolution in Language Testing/Assessment Research and The Role of Community: A Personal View</td>
<td>OGGB 4 Lecture Theatre (260-073)</td>
</tr>
<tr>
<td>5.00 – 5.15</td>
<td>Closing ceremony</td>
<td>OGGB 4 Lecture Theatre (260-073)</td>
</tr>
<tr>
<td>6.30 – late</td>
<td>Banquet</td>
<td>Fale Pasifika, Wynyard Street</td>
</tr>
</tbody>
</table>
# LTRC 2018 Conference Schedule

## Monday, 2 July

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.30 – 5.00</td>
<td><strong>Language Learning Roundtable</strong>&lt;br&gt;Connecting Language Testers and Specialists in L2 Pedagogy: Issues in Classroom-Based Assessment&lt;br&gt;Convenors: MaryAnn Christison &amp; Adrian Palmer&lt;br&gt;Sponsored by <em>Language Learning</em>&lt;br&gt;Venue: OGGB, Case Room 3 (260-055)</td>
<td></td>
</tr>
<tr>
<td>9.00 – 4.00</td>
<td><strong>Workshop 1</strong>&lt;br&gt;A Learning-oriented Assessment Approach in Classroom Contexts: An Expanded Conceptualization of Performances&lt;br&gt;James E. Purpura &amp; Carolyn E. Turner&lt;br&gt;Venue: Seminar Room B (260-040B)</td>
<td><strong>Workshop 2</strong>&lt;br&gt;Mixed Methods Research in Language Assessment Validation Studies&lt;br&gt;Mehdi Riazi&lt;br&gt;Venue: Seminar Room C (260-040C)</td>
</tr>
<tr>
<td></td>
<td><strong>Workshop 3</strong>&lt;br&gt;A Problem-solving Approach to Exploratory and Confirmatory Factor Analysis and Structural Equation Modeling&lt;br&gt;Aek Phakiti&lt;br&gt;Venue: Computer Lab 12 (260-036)</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Coffee break sponsored by the Language Training and Testing Center (LTTC)</strong></td>
<td></td>
</tr>
</tbody>
</table>

## Tuesday, 3 July

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.00 – 4.00</td>
<td><strong>Workshop 1</strong>&lt;br&gt;A Learning-oriented Assessment Approach in Classroom Contexts: An Expanded Conceptualization of Performances&lt;br&gt;James E. Purpura &amp; Carolyn E. Turner&lt;br&gt;Venue: Seminar Room B (260-040B)</td>
<td><strong>Workshop 2</strong>&lt;br&gt;Mixed Methods Research in Language Assessment Validation Studies&lt;br&gt;Mehdi Riazi&lt;br&gt;Venue: Seminar Room C (260-040C)</td>
</tr>
<tr>
<td></td>
<td><strong>Workshop 3</strong>&lt;br&gt;A Problem-solving Approach to Exploratory and Confirmatory Factor Analysis and Structural Equation Modeling&lt;br&gt;Aek Phakiti&lt;br&gt;Venue: Computer Lab 12 (260-036)</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Coffee breaks sponsored by Pearson and by the Occupational English Test (OET)</strong></td>
<td></td>
</tr>
<tr>
<td>11.00 – 3.30</td>
<td><strong>ILTA Executive Board meeting</strong>&lt;br&gt;Venue: Arts 2 Building, Room 207-314</td>
<td></td>
</tr>
<tr>
<td>4.00 – 5.00</td>
<td><strong>LTRC Newcomers’ Session</strong>&lt;br&gt;Sponsored by the Language Learning and Testing Foundation (LLTF)&lt;br&gt;Venue: OGGB Foyer, Level 0</td>
<td></td>
</tr>
<tr>
<td>5.00 – 7.00</td>
<td><strong>Symposium 1</strong>&lt;br&gt;Transition, Mobility, Validity: English as a (Multi)lingua Franca Perspectives on Language Assessment&lt;br&gt;Luke Harding (Chair), Niina Hynninen, Andrew Sewell, Elana Shohamy, Constant Leung, Jennifer Jenkins, Joe Lo Bianco&lt;br&gt;Venue: Lecture Theatre OGGB 4 (260-073)</td>
<td></td>
</tr>
<tr>
<td>7.00 – 9.00</td>
<td><strong>Welcome Reception</strong>&lt;br&gt;Sponsored by Paragon Testing Enterprises and Cambridge Assessment English&lt;br&gt;Maori cultural performance by Auckland Girls’ Grammar School Kapa Haka Group&lt;br&gt;Venue: OGGB Foyer, Level 0</td>
<td></td>
</tr>
</tbody>
</table>
**Wednesday, 4 July**

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.30 – 9.00</td>
<td><strong>Official Opening</strong></td>
</tr>
<tr>
<td></td>
<td>Venue: Lecture Theatre OGGB 4 (260-073)</td>
</tr>
<tr>
<td>9.00 – 10.00</td>
<td><strong>Davies Lecture</strong></td>
</tr>
<tr>
<td></td>
<td>No Policy without Testing! How the Language of Policy Persuasion and</td>
</tr>
<tr>
<td></td>
<td>Persuasive Language Help to Make Testing Count</td>
</tr>
<tr>
<td></td>
<td>Joe Lo Bianco</td>
</tr>
<tr>
<td></td>
<td>Sponsored by the British Council</td>
</tr>
<tr>
<td></td>
<td>Venue: Lecture Theatre OGGB 4 (260-073)</td>
</tr>
<tr>
<td>10.00 – 10.30</td>
<td><strong>Coffee break</strong></td>
</tr>
<tr>
<td>10.30 – 11.00</td>
<td><strong>4 Parallel Sessions</strong></td>
</tr>
<tr>
<td>Learning-oriented Assessment Feedback for Interactional Competence: Developing a Checklist to Support Teachers and Learners</td>
<td>Lyn May, Fumiyo Nakatsuahara, Daniel Lam, Evelina Galaczi</td>
</tr>
<tr>
<td>Examining Academic Speaking Ability and its Cognitive Dimension through a Scenario-based Speaking Assessment</td>
<td>Yuna Patricia Seong</td>
</tr>
<tr>
<td>EMI Classroom Assessment Practices: A Survey of Taiwanese Tertiary Teachers</td>
<td>Jessica Wu, Naihsin Li</td>
</tr>
<tr>
<td>Measuring the Construct of Spoken Collocational Competence: Theoretical Underpinnings, Operational Measurement, and Implications</td>
<td>Jing Xu</td>
</tr>
<tr>
<td>Venue: Lecture Theatre OGGB 4 (260-073)</td>
<td>Venue: Case Room 2 (260-057)</td>
</tr>
<tr>
<td></td>
<td>Venue: Case Room 3 (260-055)</td>
</tr>
<tr>
<td></td>
<td>Venue: Case Room 4 (260-009)</td>
</tr>
<tr>
<td>11.05 – 11.35</td>
<td><strong>4 Parallel Sessions</strong></td>
</tr>
<tr>
<td>Assessing Interactional Competence in Contexts of Movement and Change</td>
<td>Jayanti Banerjee, Noriko Iwashita, India Plough</td>
</tr>
<tr>
<td>Validating the Retelling Task of Test for English Majors Band 4 Oral in China: Evidence from a Corpus-based Exploration</td>
<td>Xiaojoy Kong</td>
</tr>
<tr>
<td>Classroom Assessment Literacy for EFL Teachers: Conceptualisation, Operationalisation, and Validation</td>
<td>Peter Gu</td>
</tr>
<tr>
<td>Examining the Effect of Delivery Mode on Scores in the Speaking Domain of ACCESS for ELLs 2.0</td>
<td>Kyungwon Bishop, Shinhye Lee, Howard Gary Cook</td>
</tr>
<tr>
<td>Venue: Lecture Theatre OGGB 4 (260-073)</td>
<td>Venue: Case Room 2 (260-057)</td>
</tr>
<tr>
<td></td>
<td>Venue: Case Room 3 (260-055)</td>
</tr>
<tr>
<td></td>
<td>Venue: Case Room 4 (260-009)</td>
</tr>
</tbody>
</table>
### Wednesday 4 July cont’d

<table>
<thead>
<tr>
<th>Time</th>
<th>Session/Activity</th>
<th>Venue 1</th>
<th>Venue 2</th>
<th>Venue 3</th>
<th>Venue 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.40 – 12.10</td>
<td><strong>4 Parallel Sessions</strong></td>
<td>Re-conceptualizing the Construct of Computer-based Speaking Assessments: Use of Communicative Strategies in Paired Discussion</td>
<td>Development and Examination of a Scenario-based, Semi-direct Pragmatic Speaking Test for an Academic Target Language Use Domain</td>
<td>DEMO: Diagnosing Linguistic Problems in Chinese Students’ Academic Writing in English: Specifying the Target Constructs for Assessment Design</td>
<td>The Discourse of the IELTS Speaking Test: Interactional Design and Practice</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lin Zhang, Yan Jin</td>
<td>Fred S Tsutagawa</td>
<td>Qin Xie</td>
<td>Paul Seedhouse</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Venue: Lecture Theatre OGGB 4 (260-073)</td>
<td>Venue: Case Room 2 (260-057)</td>
<td>Venue: Case Room 3 (260-055)</td>
<td>Venue: Case Room 4 (260-009)</td>
</tr>
<tr>
<td>12.10 – 1.30</td>
<td>Lunch break</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Language Assessment Quarterly Editorial Board Meeting</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Venue: OGGB, Level 3, Decima Glenn Room (260-310)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>ALTAANZ Annual General Meeting</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Venue: Case Room 3 (260-055)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.30 – 3.30</td>
<td><strong>Symposium Sessions</strong></td>
<td><strong>Symposium 2</strong></td>
<td><strong>Symposium 3</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Venue: Lecture Theatre OGGB 4 (260-073)</td>
<td>Venue: Lecture Theatre OGGB 3 (260-092)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.30 – 3.50</td>
<td><strong>Coffee break</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.50 – 4.00</td>
<td><strong>Group photo</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Venue: OGGB Foyer, stairs between Level 0 and 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>4 Parallel Sessions</td>
<td>4 Parallel Sessions</td>
<td>4 Parallel Sessions</td>
<td>Venue: Lecture Theatre OGGB 4 (260-073)</td>
<td>Venue: Case Room 2 (260-057)</td>
</tr>
<tr>
<td>----------</td>
<td>--------------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------------</td>
<td>-----------------------------------------</td>
<td>---------------------------------</td>
</tr>
<tr>
<td>4.00 – 4.30</td>
<td>Testing Readiness for University Entry: An Investigation of the Relationship between Writing Proficiency Level and the Strategic Behaviours Involved in Integrated Listening- and Reading-to-write Performances</td>
<td>International Undergraduate Reading Demands: The Case for Disciplinary Literacy Perspectives in Language Test Development</td>
<td>Digging for Gold or Sticking to the Criteria - Teachers Training to be Raters of Texts</td>
<td>Sally O'Hagan, Ute Knoch, Michelle Czajkowski, Susy Macqueen, Craig Wattam</td>
<td>Elizabeth Jean Larson, Megan Vincett, Hyunah Kim, Bruce Russell, Chris Barron</td>
</tr>
<tr>
<td>4.35 – 5.05</td>
<td>Assessing Content for Enhancing Construct Representation in Source-based Writing Performance Assessment</td>
<td>Measuring Cohesion: Fair Partial-credit Scoring in Sentence-sequencing Reading Tasks</td>
<td>Consistency in Times of Change – Marrying Proficiency with Achievement</td>
<td>Yasuyo Sawaki</td>
<td>Richard David Spiby, Karen Dunn</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Venue: Lecture Theatre OGGB 4 (260-073)</td>
<td>Venue: Case Room 2 (260-057)</td>
</tr>
</tbody>
</table>
### Wednesday 4 July cont’d

<table>
<thead>
<tr>
<th>5.10 – 5.40</th>
<th><strong>4 Parallel Sessions</strong></th>
<th>An Investigation of How Test Takers Integrate Content from Reading and Listening Stimulus Materials into Performances on a TOEFL iBT Speaking Task Designed to Measure Readiness for Academic Entry</th>
<th>Designing a Diagnostic Assessment of EAP Reading: A Needs-driven Approach</th>
<th>Linguistic Dimensions of TOEFL iBT Essays Compared with Successful Student Disciplinary Writing</th>
<th>Developing a Validity Argument for a Diagnostic Computer-adaptive Vocabulary Knowledge Profiler</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Kellie Frost, Joshua Clothier, Gillian Wigglesworth</td>
<td>Xiaohua Liu</td>
<td>Eric Friginal, Sara Cushing, Scott Grapin, Margaret Malone</td>
<td>Benjamin Kremmel</td>
<td>Venue: Lecture Theatre OGGB 4 (260-073)</td>
</tr>
</tbody>
</table>

| 6.30 onwards | **Wine Tasting** | Harbourside Grill, Ferry Building, Quay Street, City |

### Thursday 5 July

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>10.45 – 11.05</td>
<td><strong>Coffee break</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11.10 – 12.10</td>
<td><strong>Messick Lecture</strong></td>
<td>Evolving our Knowledge Infrastructures in Measurement / Recovering Messick’s Singerian Approach to Inquiry</td>
<td>Pamela Moss</td>
</tr>
<tr>
<td>12.10 – 1.50</td>
<td><strong>Lunch break</strong></td>
<td>ILTA ABM</td>
<td>Venue: OGGB, Level 0, Case Room 3 (260-055)</td>
</tr>
<tr>
<td>Time</td>
<td>Session</td>
<td>Title</td>
<td>Speakers</td>
</tr>
<tr>
<td>------</td>
<td>---------</td>
<td>-------</td>
<td>----------</td>
</tr>
<tr>
<td>1.50 – 2.20</td>
<td>4 Parallel Sessions</td>
<td>DEMO: The Duolingo English Test</td>
<td>Burr Settles, Cynthia M. Berger, Masato Hagiwara</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cloze Test Item Validation Study of L2 and L1 Performance Using Mechanical Turk</td>
<td>Jonathan William Trace</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pressure for High-Stakes Tests to Contribute towards Positive Washback: The Case of a Locally Developed Test</td>
<td>Carol Dabarera</td>
</tr>
<tr>
<td></td>
<td></td>
<td>An Investigation of Strategy Use on Expeditious and Careful Reading Tasks</td>
<td>Limei Zhang</td>
</tr>
<tr>
<td>2.20 – 3.30</td>
<td>Poster Session</td>
<td>Determining Seat Time for a Shortened Section of a High-Stakes Standardized Test</td>
<td>L. D. Nicolas May, Patrick M. McLain, Stephen P. O’Connell, Gad Lim</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Towards Improved Clinical Communication: An Overview of the Revised Version of the Occupational English Test (OET)</td>
<td>Gad S Lim, Brigita Seguis, Barbara Ying Zhang</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Language Assessment Education in a Globalized Era—A Cross-Cultural Comparative Study on Language Testing and Assessment Courses</td>
<td>Wenjing Yao</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A Re-envisioning and Re-validation of an Assessment Framework for a Large-scale Language Proficiency Test: Participation and Impact of Stakeholders</td>
<td>Margo Gottlieb, Mariana Castro, Tim Boals</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Assessing Comprehension of Conversational Implicature: Comparing Two Item Formats</td>
<td>Stephen P. O’Connell</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mixed-methods Development of a Test-taker-oriented Writing Rating Scale</td>
<td>Beverly Baker, Maryam Homayounzadeh, Angel Arias</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Using Response Time Data to Understand Test Performance for Adolescent Learners of English as a Foreign Language</td>
<td>Venessa Manna, Hanwook Yoo</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Assessment of Medical Terminology</td>
<td>Simon David Boynton, Laura Wakeland, Michelle Reyes Raquel</td>
</tr>
</tbody>
</table>
### Poster session cont’d

**2.20 – 3.30**

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.20 – 3.30</td>
<td><strong>Using Survival Analysis to Investigate Test Performance for Learners of English as a Foreign Language</strong>&lt;br&gt;Lora F Monfils, Venessa F Manna</td>
</tr>
<tr>
<td></td>
<td><strong>Faking Low Proficiency for Legal Advantage: Is ‘Standardized’ Testing Possible?</strong>&lt;br&gt;Margaret van Naerssen</td>
</tr>
<tr>
<td></td>
<td><strong>Self-Assessment (SA) of Foreign Language (FL) Literacy Abilities among Adolescent Chinese Learners of English</strong>&lt;br&gt;Huan Liu, Cindy Brantmeier</td>
</tr>
<tr>
<td></td>
<td><strong>Developing and Validating Tasks to Assess English Core Competencies and Tracking Students’ Development in Senior High School</strong>&lt;br&gt;Shaoqian Luo, Yumei Zhang</td>
</tr>
<tr>
<td></td>
<td><strong>Linking the Global Test of English Communication (GTEC) to CEFR Levels</strong>&lt;br&gt;Masashi Negishi, Yoji Kudo, Yasuko Okabe, Yuko Kashimada, Mika Hama, Yuko Umakoshi</td>
</tr>
<tr>
<td></td>
<td><strong>The Test-takers’ Perception: An Investigation of the Source-Text Influence on Integrated Writing Performance</strong>&lt;br&gt;Youxia Sun, Xiaoxian Guan</td>
</tr>
<tr>
<td></td>
<td><strong>Strengthening Beginning Chinese Heritage Language Teachers’ Assessment Literacy: Conceptions and Practices</strong>&lt;br&gt;Jiani Yun, Mary Hill, Martin East</td>
</tr>
<tr>
<td></td>
<td><strong>Working towards an English Language Test for Young EFL Learners in Shanghai: The Role of Stakeholders</strong>&lt;br&gt;Jufang Kong, Mingwei Pan, Jason Fan</td>
</tr>
</tbody>
</table>

**Venue:** Lobby, OGGB Level 0  

### 3.30 – 3.50 Coffee break

### 3.50 – 4.20 Coffee break

### 4 Parallel Sessions

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.50 – 4.20</td>
<td><strong>How Young Learners’ Interest Can Be Used to Support a Learner-oriented Assessment Approach</strong>&lt;br&gt;Megan Vincett, Clarissa Lau</td>
</tr>
<tr>
<td></td>
<td><strong>Classroom Assessment for Engaging Chinese First-year University Students: A Didactic-experiential Continuum</strong>&lt;br&gt;Jiming Zhou</td>
</tr>
<tr>
<td></td>
<td><strong>Can the Strength of a Speaker’s Accent Be Reliably Measured?</strong>&lt;br&gt;Gary John Ockey</td>
</tr>
<tr>
<td></td>
<td><strong>Creating a Validity Argument for an Existing Testing Program for University Admissions and Professional Certification</strong>&lt;br&gt;Angel Arias, Beverly Baker, Carol Chapelle</td>
</tr>
</tbody>
</table>

**Venue:** Lecture Theatre OGGB 4 (260-073)  
**Venue:** Case Room 2 (260-057)  
**Venue:** Case Room 3 (260-055)  
**Venue:** Case Room 4 (260-009)
## Thursday 5 July cont’d

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Title</th>
<th>Speakers</th>
<th>Venue</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.25 – 4.55</td>
<td><strong>4 Parallel Sessions</strong></td>
<td>The Effects of Administration and Response Modes on Grades 1-2 Students’ Academic Writing Performance</td>
<td>Ahyoung Alicia Kim, Shinhye Lee, Mark Chapman, Carsten Wilmes</td>
<td>Venue: Lecture Theatre OGGB 4 (260-073)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Analysis of Written and Oral Comments Within Peer Feedback Processes from the Perspective of Recipients</td>
<td>Qiyun Zhu</td>
<td>Venue: Case Room 2 (260-057)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Developmental Patterns of Young EFL Students’ Oral Narratives</td>
<td>Ching-Ni Hsieh, Yun Wang, Florencia Tolentino</td>
<td>Venue: Case Room 3 (260-055)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Evaluating Validity Claims for Existing Diagnostic Assessments: Consequences Revisited</td>
<td>Yong-Won Lee</td>
<td>Venue: Case Room 4 (260-009)</td>
</tr>
<tr>
<td>5.00 – 5.30</td>
<td><strong>4 Parallel Sessions</strong></td>
<td>Young Korean EFL Learners’ Reading and Test-taking Strategies in a Paper and Computer-based Reading Comprehension Test</td>
<td>Soohye Yeom, Henry Jun</td>
<td>Venue: Lecture Theatre OGGB 4 (260-073)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DEMO: Tuning into the Listeners’ Brain: Providing Neurocognitive Evidence for Validity through the Near-Infrared Spectroscopy (fNIRS) Technology</td>
<td>Vahid Aryadoust, Chiao-Yi Wu</td>
<td>Venue: Case Room 2 (260-057)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A Comparison of ESL Students’ Performance in Face-to-face and Virtual-world Group Oral Tests</td>
<td>Jayoung Song</td>
<td>Venue: Case Room 3 (260-055)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standards of Language Assessment for Placement Purposes: A Survey of University-level English Placement Tests</td>
<td>Jason Fan, Yan Jin</td>
<td>Venue: Case Room 4 (260-009)</td>
</tr>
<tr>
<td>5.45 – 7.00</td>
<td><strong>Assessing Writing Editorial Board Meeting</strong></td>
<td></td>
<td></td>
<td>Venue: Arts 2 Building, Room 207-314</td>
</tr>
</tbody>
</table>

## Friday 6 July

<table>
<thead>
<tr>
<th>Time</th>
<th>Symposium Sessions</th>
<th>Title</th>
<th>Speakers</th>
<th>Venue</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.45 – 10.45</td>
<td><strong>Symposium Sessions</strong></td>
<td>Symposium 5: China’s Standards of English Language Ability: Impetus for Change in Language Learning, Teaching and Assessment</td>
<td>Han Yu, Jianda Liu, Lianzhen He, Wenzia Zhang, Sha Wu, Barry O’Sullivan, Nick Saville, Yan Jin</td>
<td>Venue: Lecture Theatre OGGB 4 (260-073)</td>
</tr>
</tbody>
</table>
**Friday 6 July cont’d**

<table>
<thead>
<tr>
<th>10.45 – 11.05</th>
<th>Coffee break</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.05 – 12.20</td>
<td><strong>Work-in-Progress Session</strong></td>
</tr>
<tr>
<td></td>
<td>See the guidelines on page 109</td>
</tr>
</tbody>
</table>

<p>| 1. <strong>An Exploration of Scenario-based Assessment of Speaking for Argument-building</strong>&lt;br&gt;Jorge Luis Beltran Zuniga |
| 2. <strong>Leveraging Eye Tracking for Gathering Validity Evidence Based on Response Processes</strong>&lt;br&gt;Ruslan Suvorov |
| 3. <strong>Investigating Picture-based Prompt Comparability in an In-house Writing Test</strong>&lt;br&gt;Yali Shi |
| 4. <strong>Towards Change in Language Assessment Literacy Development: Designing an Instrument to Assess the LAL of EFL Teacher Educators</strong>&lt;br&gt;Salome Villa Larenas |
| 5. <strong>The Impact of Raters’ Use of an Analytic Rating Scale on Language Advising</strong>&lt;br&gt;Stephanie Rummel, Morena Botelho de Magalhães, Judit McPherson |
| 6. <strong>Using Virtual Reality (VR) for Language Assessment</strong>&lt;br&gt;Jayoung Song, Wei-Li Hsu |
| 7. <strong>Development of a Framework for the Assessment of Interactional Competence: A Case of Spanish Language Learners in Trinidad &amp; Tobago</strong>&lt;br&gt;Romulo Guedez-Fernandez |
| 8. <strong>Searching for the Yardstick: Measuring the &quot;Readiness&quot; to Incorporate Sources in L2 Academic Writing</strong>&lt;br&gt;Izumi Watanabe-Kim |
| 9. <strong>An Investigation into the Cognitive Validity of the Speaking Section of the Vietnam’s Standardized Test of English Proficiency</strong>&lt;br&gt;Huu Nguyen |
| 10. <strong>Considerations for Assessment in the Multicultural Tertiary EFL Context: The Case of Arab EFL Students</strong>&lt;br&gt;Özlem Yağcı Çolakoğlu, Merve Selçuk |
| 11. <strong>Enacting the Assessment Use Argument for a Large-scale Academic English Language Proficiency Test: The Question of Construct.</strong>&lt;br&gt;Rosalie Grant, Mark Chapman |
| 12. <strong>Revision of a Pronunciation Rating Scale for a University-based English Oral Assessment: What Do Raters Look for?</strong>&lt;br&gt;Li Liu |
| 13. <strong>Rater-Cognition in Progress: An Institutional Change Study</strong>&lt;br&gt;Sukran Saygi |
| 14. <strong>Investigating the Design of Assessments of Writing Ability for Transition from Education to Work</strong>&lt;br&gt;Alan Urmston, Michelle Reyes Raquel, Felicia Fang |</p>
<table>
<thead>
<tr>
<th>Time</th>
<th>Session Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.05 – 12.20</td>
<td>Work-in-Progress Session cont’d</td>
</tr>
<tr>
<td></td>
<td><strong>15. Assessing the Strategic Competence of Young Language Learners for Oral Communication: The Development of a Self-assessment Scale</strong></td>
</tr>
<tr>
<td></td>
<td>Jun Wang</td>
</tr>
<tr>
<td></td>
<td><strong>16. Development and Validation of a Diagnostic Rating Scale for an EFL University Writing Classroom</strong></td>
</tr>
<tr>
<td></td>
<td>Apichat Khamboonruang</td>
</tr>
<tr>
<td></td>
<td><strong>17. Diagnosing Listening and Reading Skills in EFL Context: Cross-mode Stability and Variability</strong></td>
</tr>
<tr>
<td></td>
<td>Lidi Xiong, Shangchao Min, Min Wang</td>
</tr>
<tr>
<td></td>
<td><strong>18. Integrating Language Proficiency into Academic Readiness Screening</strong></td>
</tr>
<tr>
<td></td>
<td>Peppi Taalas, Anna Grönlund</td>
</tr>
<tr>
<td></td>
<td><strong>19. Investigating the Spoken Production of Young Learners Aged 13-15</strong></td>
</tr>
<tr>
<td></td>
<td>Gwendydd Caudwell</td>
</tr>
<tr>
<td></td>
<td><strong>20. Assessing Machine Translation in Use: Trustworthiness and Safety of Google Translate Translations in Obtaining Surgical Consent</strong></td>
</tr>
<tr>
<td></td>
<td>Susy Macqueen, Christine Phillips, Tina Campbell, Zhengdao Ye, Johanna Rendle-Short, Katrina Anderson, Diana Slade</td>
</tr>
<tr>
<td></td>
<td>Venue: OGGB, Team Based Learning Lab 260-040</td>
</tr>
<tr>
<td>12.20 – 2.00</td>
<td>Lunch break</td>
</tr>
<tr>
<td></td>
<td><strong>Language Testing Editorial Board Meeting</strong></td>
</tr>
<tr>
<td></td>
<td>Venue: OGGB, Level 3, Decima Glenn Room (260-310)</td>
</tr>
<tr>
<td>1.00 – 2.00</td>
<td><strong>Graduate Student Meeting</strong></td>
</tr>
<tr>
<td></td>
<td>Venue: OGGB, Seminar Room C (260-040C)</td>
</tr>
<tr>
<td>2.00 – 2.30</td>
<td>4 Parallel Sessions</td>
</tr>
<tr>
<td></td>
<td><strong>Developing a Social Justice Questionnaire for Language Assessment</strong></td>
</tr>
<tr>
<td></td>
<td>Matthew Wallace</td>
</tr>
<tr>
<td></td>
<td><strong>Lecturer English Proficiency Assessment: What Norms?</strong></td>
</tr>
<tr>
<td></td>
<td>Slobodanka Dimova</td>
</tr>
<tr>
<td></td>
<td><strong>The Effect of Multilingual Assessment on the Science Achievement of Pupils</strong></td>
</tr>
<tr>
<td></td>
<td>Fauve De Backer</td>
</tr>
<tr>
<td></td>
<td><strong>The Effect of Nonverbal Communication on Listening Item Type Performance</strong></td>
</tr>
<tr>
<td></td>
<td>Aaron Olaf Batty</td>
</tr>
<tr>
<td></td>
<td>Venue: Lecture Theatre OGGB 5 (260-051)</td>
</tr>
<tr>
<td></td>
<td>Venue: Case Room 2 (260-057)</td>
</tr>
<tr>
<td></td>
<td>Venue: Case Room 3 (260-055)</td>
</tr>
<tr>
<td></td>
<td>Venue: Case Room 4 (260-009)</td>
</tr>
</tbody>
</table>
## Friday 6 July cont’d

### 2.35 – 3.05

<table>
<thead>
<tr>
<th>4 Parallel Sessions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language and Citizenship: A2 Tests and Low-educated Learners</td>
</tr>
<tr>
<td>The Domain Expert Perspective on Workplace Readiness: Investigating the Standards Set on the Writing Component of an English Language Proficiency Test for Health Professionals</td>
</tr>
<tr>
<td>Profiling Learners Using Latent Variable Mixture Modelling: Motivation in an EMI Context</td>
</tr>
<tr>
<td>Looking into Listening: Establishing the Cognitive Validity of the Aptis Listening Test</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Speakers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bart Deygers</td>
</tr>
<tr>
<td>Simon John Davidson</td>
</tr>
<tr>
<td>Karen Dunn, Barry O’Sullivan</td>
</tr>
<tr>
<td>Kathrin Eberharter, Franz Holzknecht, Benjamin Kremmel, Gareth McCray, Matthias Zehentner, Eva Konrad, Carol Spöttl</td>
</tr>
</tbody>
</table>

| Venue: Lecture Theatre OGGB 5 (260-051) |
| Venue: Case Room 2 (260-057) |
| Venue: Case Room 3 (260-055) |
| Venue: Case Room 4 (260-009) |

### 3.10 – 3.40

<table>
<thead>
<tr>
<th>4 Parallel Sessions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Fairness: Examining Differential Functioning of the Reading Comprehension Section of the GSEE across Graduation Time in China</td>
</tr>
<tr>
<td>The Effect of Classroom-based Assessments on Medical Students’ Learning: A Case Study of EMP Assessments in a University Classroom</td>
</tr>
<tr>
<td>Community-based Participatory Research Applied to Language Assessment</td>
</tr>
<tr>
<td>The Impact of Connected Speech on Item Difficulty in Second Language Listening Assessment</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Speakers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shangchao Min, Lianzhen He</td>
</tr>
<tr>
<td>Lan Luo</td>
</tr>
<tr>
<td>Lia Plakans</td>
</tr>
<tr>
<td>Jonathan William Trace, James Dean Brown</td>
</tr>
</tbody>
</table>

<p>| Venue: Lecture Theatre OGGB 5 (260-051) |
| Venue: Case Room 2 (260-057) |
| Venue: Case Room 3 (260-055) |
| Venue: Case Room 4 (260-009) |</p>
<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.40 – 4.00</td>
<td>Coffee break</td>
</tr>
<tr>
<td>4.00 – 5.00</td>
<td><strong>Distinguished Achievement Award Winner Lecture</strong></td>
</tr>
<tr>
<td></td>
<td>The Methodological Evolution in Language Testing/Assessment Research and the Role of Community: A Personal View</td>
</tr>
<tr>
<td></td>
<td>Carolyn E. Turner</td>
</tr>
<tr>
<td></td>
<td>Sponsored by Cambridge Assessment English and ILTA</td>
</tr>
<tr>
<td></td>
<td>Venue: Lecture Theatre OGGB 4 (260-073)</td>
</tr>
<tr>
<td>5.00 – 5.15</td>
<td><strong>Closing ceremony</strong></td>
</tr>
<tr>
<td></td>
<td>Venue: Lecture Theatre OGGB 4 (260-073)</td>
</tr>
<tr>
<td>6.30 – late</td>
<td><strong>Banquet</strong></td>
</tr>
<tr>
<td></td>
<td>Venue: Fale Pasifika, Wynyard Street</td>
</tr>
</tbody>
</table>
Pre-conference events

Roundtable
Sponsored by Language Learning

Monday, July 2
8:30 am – 5:00 pm
Venue: OGGB, Case Room 3 (260-055)

Connecting Language Testers and Specialists in L2 Pedagogy: Issues in Classroom-based Assessment

Organisers: MaryAnn Christison and Adrian Palmer

The purpose of the Roundtable is to (1) target some of the issues in classroom-based assessment from the perspectives and experiences of nine specialists in L2 testing and (2) provide an opportunity for the exchange of information related to these issues. The Roundtable is intended to provoke discussion among L2 testers who are currently working with classroom teachers. It is also intended to jumpstart discussions between L2 testers and specialists in L2 pedagogy, including experienced L2 teachers and teacher educators. The presenters are specialists in L2 testing who are working with L2 teachers, either pre-service teachers or practicing teachers, and the six papers reflect their current work as they engage with L2 teachers on issues concerning classroom-based assessment. The specialists in L2 testing are working on a broad range of topics from justification to language proficiency frameworks and in diverse contexts to include public schools, higher education, formal courses, and informal workshops. The participant audience will be individuals who are experienced L2 teachers and L2 teacher educators, as well as other language testers, so that we might have a rich and informed discussion and receive critical and informed responses to the work of the roundtable presenters.

Paper 1
Lyle Bachman and Barbara Damböck
A new approach to classroom-based language assessment

Paper 2
Adrian Palmer, Daniel Dixon, and MaryAnn Christison
Teaching language test design and development: The use of software and discovery learning

Paper 3
Gary John Ockey
The alignment of teaching and assessing oral communication skills

Paper 4
James Purpura
Learning-oriented language assessment: A framework for considering the nexus of instruction, learning, and assessment in classroom contexts

Paper 5
John Read and Mark Dawson-Smith
Adapting an international proficiency framework for a national context and the consequences for classroom teachers
Pre-conference workshops
Monday, July 2 and Tuesday, July 3
9:00 am – 4:00 pm

WORKSHOP 1
Venue: OGGB, Seminar Room B (260-040B)

A Learning-oriented Assessment Approach in Classroom Contexts: An Expanded Conceptualization of Performance

Leaders: James E. Purpura and Carolyn E. Turner

The goal of this workshop is to introduce participants to a learning-oriented approach to assessment (LOA) so that this approach can serve as a conceptual framework for the design and validation of summative assessments, as well as a frame of reference for examining assessments embedded in instruction and learning, and at times mediated through classroom discourse.

L2 assessments have come in many forms over the years, and have been used by L2 teachers to obtain information on what students have learned, what remains to be learned, and what instruction needs to transpire. Teachers have also dealt with the need to align assessments with large-scale tests external to the classroom (Pellegrino et al., 2001). This has brought about some confusion as classroom-based assessment (CBA) is contextually bound and at times socially constructed, involving several stakeholders. Evidence of learning (or lack thereof) is collected through a repertoire of methods, and used to monitor learning (and teaching) on an ongoing basis. In other words, assessment is more a resource for learning than a means of judgment. While CBA can overlap with the methods used in standardized tests in many ways, most CBAs are assessment without tests, where teachers are concerned with integrating instruction and assessment to support learning. External standardized tests, on the other hand, are not designed to promote such integration.

Such circumstances have contributed to a growing interest by testers in the importance and uniqueness of the classroom as the nexus for teaching, learning and assessment. However, only in the late 1990s, did researchers start studying the role of assessment in classroom settings (Turner, 2012), and only more recently, was CBA considered a different paradigm of assessment (James, 2006; Moss, 1996), with a need for a unified system across external tests, curriculum and pedagogy (Cumming, 2009; Hudson, 2012).

Within this setting, Purpura and Turner (forthcoming) have proposed a learning-oriented assessment approach to assessment in which multiple dimensions of the assessment context converge to address synergies among assessment, instruction, and learning; that is, where performance moderators are taken into consideration as they are related to instruction and learning in both construct-relevant and irrelevant ways. How the LOA framework can be used in the construction and interpretation of assessments will be the focus of this workshop.

Day 1:
- Explore why a broader conceptualization of the assessment process is needed.
- Situate LOA within the assessment literature.
- Describe the dimensions of the LOA framework using a classroom test (contextual, elicitation & evidence, proficiency, socio-cognitive, instructional, affective, social-interactional).
- Analyze the design and results of an online test (the wetlands experience), in which instruction and learning have been embedded into the assessment. (To gain first-hand experience with this type of assessment, participants will be asked to take this test before the workshop.)
- Discuss how assessments can be constructed from a learning orientation, and examine examples from a variety of learner groups.

Day 2:
- Explore LOA as an organizing framework for examining assessments embedded in instruction and learning, and at times, mediated through classroom discourse.
- Use a modified conversation-analytic approach to examine assessment episodes in transcript data and examine how the framework can highlight different dimensions of the data, providing insights for how assessment relates to instruction and learning.
- Examine examples from varying age groups
- Work on adapting/designing assessments for your specific contexts.
James E. Purpura is Professor of Linguistics and Education at Teachers College, Columbia University, where he teaches courses in L2 assessment and research methods. His research focuses on grammar and pragmatics assessment, cognition and L2 assessment, learning-oriented assessment (LOA), scenario-based assessment (SBA), and L2 test validation. In addition to numerous articles and chapters, Jim has written Strategy use and L2 test performance (1999, CUP), Assessing grammar (2004, CUP), and is currently co-authoring a book on LOA in L2 classrooms (Routledge) and a book on the collection of the works of Lyle Bachman. Jim is the co-editor of Language Assessment Quarterly, co-editor of the series, New Perspectives in Language Assessment (Routledge), and co-editor of the series, Language Assessment at ETS: Innovation and Validation (Routledge). He served as President of ILTA (2007 & 2008), as member of the TOEFL Committee of Examiners, and currently serves on the U.S. Defense Language Testing Advisory Panel. He was recently a Fulbright Scholar in Siena, Italy, where he worked on an LOA approach to scenario-based assessments. Jim is co-founder of LOA Solutions Inc., which provides a platform for Internet and mobile phone delivery of assessments that enable rapid iteration and artificial intelligence to improve design, authenticity, efficiency, and feedback.

Carolyn E. Turner has been Associate Professor of Second Language Education in the Department of Integrated Studies in Education at McGill University. Her courses focused on classroom assessment, language testing and research methods. In December 2015 she retired from McGill. She currently continues to be active in research and service to her professional community. Her research examines language testing/assessment in educational and healthcare settings. More recently, her focus has also been on learning-oriented assessment (LOA) in classroom contexts. She is co-authoring a book with James Purpura on LOA where they examine the potential of learning as integrated into assessments and how assessments serve learning when embedded in teaching. Carolyn served as President of ILTA (2009 & 2010); was Associate Editor of the journal of Language Assessment Quarterly from its inception up until 2012, remains on its Editorial Advisory Board; and was a founding member of the Canadian Association for Language Assessment/Association canadienne d’évaluation des langues. She has worked with organizations concerning high-stakes testing issues: for example, the TOEFL Committee of Examiners (ETS); the Test Advisory Panel for Paragon Testing Enterprises; International Civil Aviation Organization (ICAO) Aviation English Language Test Service (AELTS), and the Quebec Ministry of Education. Her publications are in journals such as Language Testing, Language Assessment Quarterly, TESOL Quarterly, Canadian Modern Language Review and Health Communication, and chapters in edited collections. In addition, Carolyn has valued supervising graduate students and mentoring young faculty. She is encouraged by their increasing involvement and presence in the international language testing/assessment community.
Mixed Methods Research in Language Assessment Validation Studies

Leader: Mehdi Riazi

Mixed-methods research (MMR) is based on the premise that quantitative and qualitative methods can be combined in the collecting, analysing and presenting of research data to make more comprehensive inferences. The application of MMR to the investigation of research problems is increasingly gaining prominence and utility in social science and educational research. This development acts as a counterpoint to the dominant polarisation, until relatively recently, of adopting an ‘either/or’ perspective (i.e. either quantitative or qualitative) in research methodology. What is clear is that the appeal of MMR is growing and proving valuable for a wide range of researchers in a variety of academic disciplines including applied linguistics in general and language assessment in particular.

While the development and use of MMR in social and educational research has been growing considerably, it has been subject to a substantial amount of misinterpretation and unsystematic use and the rationale for choosing a particular type of mixed-method design partly as a result of its own recent process of development and partly because researchers are sometimes not clear as why, when, and how the two methodologies might be mixed. This workshop intends to provide a space for participants’ awareness of the criteria for choosing a particular purpose and design and to explicate the use of MMR in language assessment validation studies and especially the correspondence between argument-based approach to test validation and MMR.

Participants are encouraged to familiarise themselves with MMR literature before attending the workshop to develop a common background on MMR. They may read all or one of the following sample sources. In the workshop, we will discuss how a test validation study might be designed using an argument-based approach and how MMR may be used to collect and analyse data and provide evidence for each of the inferences in the argument-based framework to test validation. As a case in point, we will discuss Johnson and Riazi, A.M. (2017) and Han and Riazi (2017). Participants are also encouraged to discuss their own projects (if they are involved in any) in the workshop for discussion and reflection.

The workshop will end with some recommendations for using MMR more systematically in language assessment validation studies.

Pre-workshop readings
Mehdi Riazi is professor of applied linguistics in the Department of Linguistics, Macquarie University, Sydney, Australia. He received his PhD in 1995 from the University of Toronto, Canada. He worked at Shiraz University for 14 years and joined Macquarie University in Australia in 2009. Currently, he convenes and lectures on two postgraduate units of Research Methods and Language Assessment at Macquarie University. He also supervises both masters and doctoral students. He has successfully supervised to completion 20 doctoral students and 48 master’s theses, and currently supervises 6 PhD candidates. His areas of interest include language testing and assessment, academic communication, second language reading and writing, and research methodology. He was the principal investigator of three research projects funded by IELTS, Pearson, and Educational Testing Service (TOEFL-iBT) on test validation. Reports of these projects were presented at EALTA (Assessing Writing-SIG) and LTRC conferences in 2014 and were then published in Papers in Language Testing and Assessment (PLTA), Assessing Writing, and an edited volume. His recent monographs on research methodology are: The Routledge Encyclopedia of Research Methods in Applied Linguistics (https://www.routledge.com/The-Routledge-Encyclopedia-of-Research-Methods-in-Applied-Linguistics/Riazi/p/book/9780415816274) published by Routledge in 2016 and Mixed Methods Research in Language Teaching and Learning (https://www.equinoxpub.com/home/mixed-methods-research-in-language-teaching-and-learning-a-mehdi-riazi/) published by Equinox in 2017.

WORKSHOP 3
Venue: OGGB, Computer Lab 12 (260-036)

A Problem-solving Approach to Exploratory and Confirmatory Factor Analysis and Structural Equation Modeling

Leader: Aek Phakiti

Overview
This two-day workshop aims to provide an introduction to exploratory and confirmatory factor analysis (EFA and CFA, respectively) and structural equation modeling (SEM) for language assessment research. EFA is a multivariate statistical technique for exploring the factors that account for the variation in participants’ responses to research instruments, such as Likert-type scale questionnaires. Common factors are often discussed as latent variables. Unlike EFA, CFA is typically used to confirm a common factor that underlies an ability or psychological facet, such as motivation or anxiety. CFA is also used as a measurement model in SEM, which establishes relationships (linear or causal; direct or indirect) among latent variables. This workshop stresses the importance of conceptual understandings and interpretations of analysis of outputs, rather than statistical formulae. In this workshop, Statistical Package for Social Sciences (SPSS) and EQS 6.3 will be used. Participants will have an opportunity to practise each analysis after a step-by-step instruction.

Target participants
The target participants are graduate students in language testing and assessment or early career researchers who are interested to know more about EFA, CFA and SEM. Although this workshop is introductory in nature, participants should have a good knowledge of basic statistics, such as descriptive statistics, normal distribution, probability and statistical significance, correlation, regression and reliability analysis. The reason for this is that EFA, CFA and SEM rely on an application of these statistics and terms linked to such statistics (e.g., univariate normality and variance and co-variance) may be cross-referenced.

Day 1
Morning: Day 1 of this workshop provides an overview of important aspects, considerations and practical guidelines for conducting EFA. This session also includes Principal Component Analysis (PCA). Participants will be given a step-by-step
instruction of how to perform EFA using SPSS and what to look for and how to interpret SPSS outputs. Participants will have an opportunity to perform EFA on their own.

Afternoon: Examples of language testing and assessment research that employs EFA or PCA will be discussed and critiqued. Suggestions of how to report EFA will be provided. In the remaining hours of Day 1 (approximately 1.5 to 2 hours), CFA and SEM will be introduced. CFA will be compared to EFA. CFA takes a confirmatory approach to common factors that underlie an ability or psychological facet, such as motivation or anxiety. Typically, researchers develop a hypothesised CFA model, which is then tested using an empirical data set. SEM will be discussed in relation to correlation, multiple regression and path analysis. Steps for CFA and how to read CFA models will be presented.

Day 2

Morning: This workshop will first revisit what was covered and arising issues from Day 1. Then, it will focus on the methodological basics of CFA and SEM techniques. Practical guidelines for how to perform CFA and SEM will be discussed. EQS program will be introduced to participants as it will be used to perform CFA and SEM. Participants will have an opportunity to familiarise with EQS program and have some initial diagram drawing practice prior to CFA. Participants will be provided with a step-by-step instruction of how to perform CFA and how to interpret EQS outputs, as well as revise and retest an initial CFA (if required). Participants will be advised on how to edit and prepare a CFA model for research reports.

Afternoon: In the remaining hours of Day 2, SEM will be discussed further in terms of how to apply it for testing complex models in language testing and assessment. CFAs are typically used as measurement models in SEM. The main focus of this workshop is only on testing and interpreting a relationship among latent and observed variables, and the role of non-random errors on parameter estimates. More sophisticated SEM analyses (e.g., simultaneous group comparisons) will be introduced but cannot be performed in this workshop. Examples of language testing and assessment research that employs CFA and SEM will be discussed and critiqued. Finally, how to report CFA and SEM will be recommended.

Readings

Two texts will be provided about one to two months prior to the workshop.

Prior to Day 1:

Prior to Day 2:

Aek Phakiti is an Associate Professor in TESOL and the Postgraduate Coursework Coordinator at the Sydney School of Education and Social Work at the University of Sydney. His research focuses on, language assessment, second language acquisition, and research methods in applied linguistics. He was the Vice President of ALTAANZ (Association of Language Testing and Assessment for Australia and New Zealand, 2015-2017) and Associate Editor of Language Assessment Quarterly (2012-2017). Aek is author and editor of numerous books in applied linguistics, including Language Testing and Assessment: Theory to Practice (Bloomsbury, 2018), The Palgrave Handbook of Applied Linguistics Research Methodology (Palgrave, forthcoming, 2018), Quantitative Methods for Second Language Research: A Problem-solving Approach (Routledge, 2018), Research Methods in Applied Linguistics: A Practical Resource (Bloomsbury, 2015); Experimental Research Methods in Language Learning (Bloomsbury, 2014), Teaching Academic Writing (University of Michigan Press, 2009), and Strategic Competence and EFL Reading Test Performance (Peter Lang, 2007). His research has been published in numerous peer-reviewed journals, including Language Testing, Language Learning, Language Assessment Quarterly, and TESOL Quarterly.
No Policy without Testing! How the Language of Policy Persuasion and Persuasive Language Help to Make Testing Count

In policy making discussions testing has a particular kind of presence and power because of what tests produce, or are assumed to produce, i.e., empirical statistically presented evidence of phenomena, characterized by objectivity, reliability, and validity. Policy making is a special field of talk that involves decision making about the disbursement of public resources, which, as classically understood in liberal democratic states, are collected from taxpayers on the assumption that they in turn receive representation. In effect this holds that citizens become tax payers, they surrender their hard earned income, for which in exchange they are entitled to participation in the use and deployment of that income forgone. The rallying phrase accompanying this contract of governance is No Taxation without Representation, attributed to John Otis in 1761, expressing the resentment of American colonists against being taxed by a British parliament to which they did not send representatives. This talk will connect the promise of language testing, and its role (assumed and actual) in decision making in various domains in which it is applied, with the work and ‘talk’ of language testing. I will reflect on the contribution made by Alan Davies to generating both an ethical discourse of testing, and a participatory mode of policy making, to the contract of governance. Testing is seen, in this view, as a way to make information “honest”.

Joseph Lo Bianco
University of Melbourne
Joseph Lo Bianco is Professor of Language and Literacy Education in the Melbourne Graduate School of Education, at the University of Melbourne, Australia and a past president of the Australian Academy of the Humanities. He specialises in language policy studies, bilingualism and intercultural education and research and action on peace and conflict in multi-ethnic settings. Since 2012 he has directed a multi-country project on language policy and social cohesion in conflict affected settings in SE Asia for UNICEF and conducted large scale policy workshops for high level policy officials across Asia, under the auspices of UNESCO. He has an extensive list of publications with a strong recent focus on social cohesion, peace and conflict mitigation in multi-ethnic settings. As author of the Australian National Policy on Languages in 1987 he was entrusted with its initial implementation, under which he assisted Professor Tim McNamara to establish The Language Testing Research Centre at the University of Melbourne, which began a long friendship and collaboration with Alan Davies, the founding Director of the Centre.
Evolving our Knowledge Infrastructures in Measurement / Recovering Messick’s Singerian Approach to Inquiry

There is a growing call in science and technology studies for critical attention to the infrastructures through which we generate knowledge about the human and natural worlds. Such Knowledge Infrastructures (KIs) entail networks of people, practices, norms, material and conceptual artifacts, standards, technologies, institutions, and the relationships among them. They refer to prior work that does not need to be reconsidered at the start of a new venture. In other words, they reflect what we take for granted as they ways things are done. [Edwards et al., 2013; Slota & Bowker, 2017; adapted from Moss & Lagoze, 2018.]

Our canonical understandings of validity theory in measurement, where Messick’s work has been seminal, are part of a robust KI supporting test development and evaluation and, in turn, action-orienting inferences about test takers, where other KIs implicated in local contexts come into play. A recent multi-disciplinary conference on KIs asked: (1) How are KIs changing? (2) How do KIs reinforce or redistribute authority, influence, and power? (3) How can we best study, know, and imagine today’s (and tomorrow’s) KIs? [Edwards et al., 2013.]

With his argument for a “Singerian approach to inquiry,” Messick (1989) has also given us tools to address questions like these. He called on us to evaluate one approach to inquiry in terms of others “to elucidate or disrupt” the focal inquiry processes. With Messick’s argument as inspiration and resource, I explore answers to the critical questions raised by KI scholars, focusing in particular on how our KIs in measurement might evolve to better support interpretation and use of test scores in local contexts.


**Pamela Moss** is a Professor of Education at the University of Michigan’s School of Education. Her work lies at the intersections of educational assessment, philosophy of social science, and interpretive or qualitative research methods. Her approach to the study of assessment engages the critical potential of dialogue across research discourses--educational measurement, hermeneutics and critical theory, and sociocultural studies--sometimes to complement, sometimes to challenge established theory and practice in assessment. Her current research agenda focuses on validity theory in educational assessment, assessment as a social practice, and approaches to methodological pluralism in research policy and practice. Sam Messick’s (1989) call for a “Singerian mode of inquiry”, where one method of inquiry is critically evaluated in terms another, provided an early inspiration for this multi-methodological agenda.

Moss is an elected member of the US National Academy of Education and a fellow of the American Educational Research Association (AERA). She was a member of the Joint Committee revising the 1999 AERA, APA, NCME Standards for Educational and Psychological Testing, of the US National Research Council’s Committee on Assessment and Teacher Quality, and chair of AERA’s Task Force on Standards for Reporting on Empirical Social Science Research in AERA Publications. She was co-founder (with Mark Wilson and Paul DeBoeck) of the journal, *Measurement: Interdisciplinary Research and Perspectives*. She recently completed a chapter on “Engaging Methodological Pluralism” (2016) with Ed Haertel, for AERA’s 5th Handbook of Research on Teaching.
The Methodological Evolution in Language Testing/Assessment Research and the Role of Community: A Personal View

As we look upon the diverse methods, methodology and study design employed in language testing/assessment research currently, one might think it was always that way. This would be far from the truth. What is intriguing about our present research context is the way we got here. I would argue that our present repertoire of research methods evolved due to many factors (e.g., the trajectory of research issues and questions which demanded new ways to explore and investigate phenomena as in other disciplines in the social sciences), but the main contributing factor I would like to focus on is the role of community and the different generations represented within it. I’m not referring to any scholarly community; I’m referring to our language assessment/testing community as manifested in our annual meeting of LTRC. As a developing researcher, I was initially influenced by the warm welcome I received at my first LTRC back in 1989 in San Antonio, Texas where I presented my PhD work. I also couldn’t help but notice the passionate and colorful debates taking place surrounding paper presentations (including my own which used structural equation modeling employing the computer program LISREL, a bit new in our field’s research at the time). The LTRC debates would frequently include the methods used and study design. What was refreshing and valuable, however, was that the dialogue would spill out after the presentations into other spaces where we had fun together – coffee breaks, meals, late evening events - and this is where much exploring, modeling and learning continued in terms of diverse methodologies for our research questions. As a growing scholar initially educated in quantitative methods but feeling the need to add a qualitative dimension and later a combined analysis (MMR), I was motivated by this sense of community and encouraged by the importance placed upon the research design of a study.
In this talk, I would like to build around the theme of the methodological evolution in language testing/assessment research, but explore it specifically through a community perspective and the impact of this perspective. I will draw on my own experiences in several instances, and demonstrate the questioning, but tolerant, approach to new ways of addressing research questions as exemplified in our community’s methodological trajectory. In my view, it is part of our collective identity and a positive force for novice and veteran researchers alike. I will argue how this is not always the case in other disciplines, and how this approach has served our community well in making advances in the issues at hand.

Carolyn E. Turner has been Associate Professor of Second Language Education in the Department of Integrated Studies in Education at McGill University. Her courses focused on classroom assessment, language testing and research methods. In December 2015 she retired from McGill. She currently continues to be active in research and service to her professional community. Her research examines language testing/assessment in educational and healthcare settings. More recently, her focus has also been on learning-oriented assessment (LOA) in classroom contexts. She is co-authoring a book with James Purpura on LOA where they examine the potential of learning as integrated into assessments and how assessments serve learning when embedded in teaching. Carolyn served as President of ILTA (2009 & 2010); was Associate Editor of the journal of Language Assessment Quarterly from its inception up until 2012, remains on its Editorial Advisory Board; and was a founding member of the Canadian Association for Language Assessment/Association canadienne d’évaluation des langues. She has worked with organizations concerning high-stakes testing issues: for example, the TOEFL Committee of Examiners (ETS); the Test Advisory Panel for Paragon Testing Enterprises; International Civil Aviation Organization (ICAO) Aviation English Language Test Service (AELTS), and the Quebec Ministry of Education. Her publications are in journals such as Language Testing, Language Assessment Quarterly, TESOL Quarterly, Canadian Modern Language Review and Health Communication, and chapters in edited collections. In addition, Carolyn has valued supervising graduate students and mentoring young faculty. She is encouraged by their increasing involvement and presence in the international language testing/assessment community.
Symposia

Symposium 1
Tuesday, July 3
5:00pm - 7:00pm
Venue: Lecture Theatre OGGB 4 (260-073)

Transition, Mobility, Validity: English as a (Multi)Lingua Franca Perspectives on Language Assessment

Chair: Luke Harding, Lancaster University, UK
Discussant: Joe Lo Bianco, University of Melbourne, Australia

Participants:
Niina Hynninen, University of Helsinki, Finland
Andrew Sewell, Lingnan University, Hong Kong
Elana Shohamy, Tel Aviv University, Israel
Constant Leung, King’s College London, UK
Jennifer Jenkins, Southampton University, UK

Summary

English as a Lingua Franca (ELF) communication represents one of the most significant challenges to language testing and assessment since the communicative revolution. Within a time of increasing mobility and transition across borders, ELF provides a means of theorising and describing language use among interactants who do not share the same linguacultural background. Having begun as an attempt to codify an international variety, ELF research has since given way to a view of language which places the dynamic and fluid nature of communicative interaction to the fore, seeking to understand how successful communication takes place in such environments. For language testing, an ELF perspective destabilises the place of the native speaker and the notion of assessing against a “stable variety” (Harding & McNamara, 2018; Jenkins & Leung, 2014; McNamara & Shohamy, 2017), while at the same time posing opportunities for reconceptualising and expanding language constructs, particularly in language for specific purposes contexts.

There has been little uptake in the language testing community of the principles of ELF in the design of instruments despite compelling arguments that target language use domains are increasingly rich sites of ELF communication. One reason for this resistance is that there is little consensus on how ELF might be defined and operationalised for assessment purposes. A related reason is that implementing ELF might weaken the validity of instruments. However, while the language testing community has been resisting ELF, research on ELF communication has expanded to provide a body of empirical work which sheds light on domains of interest to many language assessment developers.

This symposium draws on ELF research to pose four critiques of current language assessment practices, while also proposing innovative directions for exploring the interface between ELF and language assessment:

The first paper draws on the concept of “language regulation” in ELF encounters, which takes into account interactants’ perspectives of whose norms hold authority in a given communicative setting. Language regulation will be discussed in relation to questions of language norms and ELF competence in language assessment contexts.

The second paper focuses on the context of business English, drawing data from a corpus of online newspaper comments to illustrate stances towards ELF communication in the business world. These views will be juxtaposed with the comments of speaking examiners on an international English language exam to contrast different priorities.

The third paper considers ELF as a phenomenon within the broader scope of multilingualism. This paper will critique different approaches to assessing ELF from a validity perspective, and outline directions for the development of future ELF/multilingual assessments.

The final paper argues that the sociolinguistic reality of English as a (multi)lingua franca in higher education contexts renders current practices of assessing language for academic “readiness” unfit for purpose. The paper will critique test preparation practices for those same tests, and propose a new conceptualisation of what “readiness” entails in academic contexts.
Following the paper presentations, a discussant will comment, and the symposium will end with a block of time for discussion of the emerging themes and questions.

**Individual abstracts**

1. **Language Regulation and Assessment of English for Lingua Franca Use**  
Niina Hynninen, University of Helsinki, Finland

This paper discusses the concept of language regulation and its relevance to English language assessment. Language regulation refers to practices through which language users monitor, intervene in and manage their own or others’ language use (e.g. Hynninen & Solin 2017). In English as a lingua franca (ELF) research, it is oftentimes emphasised that ELF interactions are fluid and variable, and that speakers tend to orient to the demands of the situation rather than pre-existing standards. At the same time, it cannot be assumed that speakers can use language any way they like – they are bound to be constrained in some ways. By focusing on language regulation, it is possible to investigate what these ways might be; for instance, who is allocated the role of authority in language, or what is the relevance of norms that have more permanence and scope beyond a particular interactional situation. A focus on regulation can thus help to pinpoint who language users in ELF encounters turn to as authority in English, and what they consider to be ‘good’, ‘acceptable’ or ‘functional’ language use in different settings and for different genres. From the perspective of language assessment, this information may be used in deciding what kind of, and whose, language use might work as a model for assessing competence in ELF.

2. **English is a Lingua Franca: Business English and its Implications for Language Testing in an Age of Mobility**  
Andrew Sewell, Lingnan University, Hong Kong

It is frequently contended that a) most interactions in English now take place without native speakers being present, and b) a lingua franca perspective involves focusing on these non-native uses of English. However, the deeper implications of this perspective are frequently neglected. In this presentation I will argue that the lingua franca perspective destabilises a number of largely unquestioned assumptions and boundaries – including, ironically, those between ‘lingua franca’ and ‘non-lingua franca’ communication. English is not merely used as a lingua franca, it is a lingua franca for all of its users. I illustrate this claim by using data from a corpus of online newspaper comments on the subject of lingua franca communication in the business world. The data invites us to consider, among other things, that communication between ‘same-language’ or ‘native’ speakers can be as problematic as between ‘non-native’ speakers, that some of the most important factors affecting workplace communication have little to do with ‘language’ per se, and that communicating across linguistic and cultural boundaries may be one of the most crucial and yet neglected skills. I conclude by considering some current practices in the testing of spoken language, referring to another type of ‘comment’ in the form of examiner commentaries on candidates’ performance in international English tests.

3. **ELF and Multilingualism - in Search of Valid Assessment**  
Elana Shohamy, Tel Aviv University, Israel

Major developments have been taking place in the definition of languages in the past decade as the concept of multilingualism emerged as a dominating paradigm grounded in education, sociolinguistics and psycholinguistics. Topics such as ‘full language repertoire’, translanguaging, multimodality and L-1 as a resource are currently incorporated and legitimized while defying traditional views of language proficiency as monolingual, native-like, standard and pure. The notion of ELF (English as a lingua franca) (Jenkins, 2012; Seidlhofer, 2011 and others) is a derivative of multilingualism as it refers to English language variation which demonstrates that the goal of reaching ‘native’ English proficiency is not achievable by most learners as there is use of L-1 and other. Language testing has only begun addressing the issue of ELF and of multilingualism and innovative tests are being experimented with (Lopez, Turkan, and Guzman-Orth, 2017; Shohamy, 2011), not without major difficulties and obstacles. The concept of multilingualism and ELF being varied, fluid, dynamic and unpredictable poses major challenges for test developers seeking construct-valid tests which are often stable and fixed. This paper will describe the construct of ELF for test developments and will critique it in terms of construct validity. A number of proposals for the developments of appropriate ELF and multilingual tests will be introduced and evaluated.
4. Farewell to the Phantom of Standardisation
Constant Leung, King’s College London, United Kingdom
Jennifer Jenkins, Southampton University, United Kingdom

In line with the ‘transition’ aspect of this year’s Colloquium and, in particular, with ‘notions of “readiness” for entrance’, our paper will focus on university English language entry tests. We will argue that the ways in which English is used have changed dramatically over recent decades, and concomitantly, the use of English at university has become so diverse, that it is no longer possible to identify a standard or standards to which all candidates could—or should—defer. The only way forward, we believe, is to abandon conventional notions of ‘a universal standard’ as a gatekeeping chimera, and to focus instead on the individual local context and what is ‘standardly expected’ in that respect. In terms of university English language entry tests, this means considering the ways in which English is actually used in each individual setting. In many English-medium universities this is most likely to involve the use of English as a (multi)lingua franca, with students communicating across their diverse Englishes and making use of their multilingual repertoires, i.e. translanguaging. Bearing all this in mind, we argue, the use of standardised tests and preparation for such tests, does little to prepare students for ‘readiness for entrance’ into the university of their choice. We will consider alternative conceptualisations of ‘readiness’ that we believe to have far more potential to serve the intended assessment purpose.

Symposium 2
Wednesday, July 4
1:30pm - 3:30pm
Venue: Lecture Theatre OGGB 4 (260-073)

The Role of Language Tests for Immigration, Integration and Participation – From Critical Analyses to Inclusive Approaches

Chair: Claudia Harsch, University of Bremen, Germany
Discussant: Ofra Inbar-Lourie, Tel Aviv University, Israel

Participants:
Luke Harding, University of Lancaster, UK
Kamran Khan, University of Lleida, Spain
Cathie Elder, University of Melbourne, Australia
Ute Knoch, University of Melbourne, Australia
Owen Harradine, University of Melbourne, Australia
Lorenzo Rocca, Università per Stranieri, Perugia, Italy
Peppi Taalas, University of Jyväskylä, SIMHE, Finland
Anika Müller-Karabil, University of Bremen, Germany

Summary

The use of language tests for political purposes, as gatekeeping devices to limit access of certain groups to participation, integration and society, has long been criticized (e.g. Davies, 1997; Shohamy, 2001). Yet language testers and researchers are not always in a position to influence policy and practices. This symposium strives at bringing together critical contributions from around the globe that illustrate how policies and practices can be analyzed, underlying political agendas revealed, and inappropriate language test usage addressed. At the same time, the symposium aims at showcasing appropriate language test tools and assessment usage for inclusive, supportive purposes that open access to participation, particularly for displaced and vulnerable groups of migrants. We will open the symposium with a broad perspective on the contribution of security studies to uncover confluations of language testing with border security concerns. Two case studies will illustrate the intertwinedness of language tests and security concerns, a critical perspective that is much needed to better understand the contributions language testing research can make towards revealing, informing and influencing political agendas that may otherwise lead to inappropriate language test usage.
The second paper highlights the influence community activism can exert on national policies aiming at tightening security. A proposed change in Australia to tighten language requirements as prerequisite for citizenship tests was faced by severe community opposition. While the proposed changes allegedly served to promote integration and to tighten national security, a thematic analysis of data from a Senate Enquiry and ensuing presentations and debates revealed the potentially divisive impact of the proposed language testing policy and showed a high level of community awareness of issues such as fairness and equal access. This contribution illuminates the values underpinning community activism, and highlights the need for researchers to better understand such values, so that cases of inappropriate language test use can be counteracted.

The third contribution examines the role of the Council of Europe, the continent’s leading human rights organization, in its endeavor of supporting the linguistic integration of migrants. This paper shows how the Council and its language policy units can exert beneficial influence on theoretical aspects such as ethical recommendations for appropriate uses of the CEFR, as well as on practical aspects such as the development of a language evaluation toolkit for volunteers working with displaced vulnerable groups of migrants. The toolkit will be showcased, as will be a recommendation on ethical test use for policy makers by the Council’s Language Policy Unit.

Staying in the realm of linguistic integration, the final paper introduces two programs in Finland and Germany to prepare refugees for entering higher education. The two programs will be outlined in their language, content and assessment approaches, with a particular focus on inclusive test use. Furthermore, the paper will present first results from the evaluation research accompanying both programs in order to monitor whether the needs of this particular group of migrants are met.

Individual abstracts

1. **(In)security, Language Testing and “Managed” Mobility**
   Luke Harding, University of Lancaster, UK and Kamran Khan, University of Lleida, Spain

   Within the context of massive global movements of people for professional, educational, family or humanitarian purposes, the role of language tests as gatekeeping devices for immigration and citizenship has come under considerable scrutiny. Language assessment scholars, however, are yet to engage meaningfully with how language testing connects with broader processes of security, as theorised, for example by Huysmans (2014). Specifically, advances in security studies allow us to better understanding how language testing may be conflated with national/border security concerns both at the level of discourse and in practice.

   This paper will begin by making a case for the relevance of security studies in understanding language testing for immigration and citizenship purposes. Then, two cases will illustrate the deep connections between security and language testing. The first will discuss policy surrounding “secure English language tests” (SELTs) in the United Kingdom with reference to how testing organisations must align with the government’s national security framework through testing practices. The second will draw on qualitative data to demonstrate how tests in immigration contexts become forms of social sorting and surveillance that are perceived as distributing risk to particular sections of the population, justified by discourses around security.

   Through these cases, we will show how test use may be predicated on border control, with tests used as crucial tools to make migrants visible in the eyes of the state. Finally, we will discuss a research agenda for exploring language testing policy and practice which includes an orientation towards security.


2. **Resisting Changes to English Testing Requirements for Australian Citizenship: Insights from a Senate Enquiry**
   Cathie Elder, Ute Knoch, and Owen Harradine - University of Melbourne, Australia

   The role of language tests as political tools for limiting access of particular groups to prized forms of social participation and advancement has long been acknowledged (e.g., Davies 1997, McNamara & Roever 2006, Shohamy 2001). Efforts to resist such political manipulation have however received limited attention in our field.

   The present study focuses on a case of successful opposition by professional and community groups to a recent attempt by the Australian government to strengthen current requirements for Australian citizenship, including a proposal to introduce a requirement of IELTS Level 6 or equivalent as prerequisite for sitting the already mandatory citizenship test. The rationale for the new legislation was framed in terms of promoting social integration and tightening national security.

   Data from 635 submissions to a Senate Committee Enquiry and subsequent dialogue with participating politicians were thematically analyzed to uncover the values underlying the multiple and mainly oppositional arguments put
forward in response to the government proposal. The analysis revealed ways of thinking about language and social integration that were at odds with the government agenda, including the widely-held view that rather than promote integration, the new English requirement would unfairly disenfranchise and alienate certain groups of applicants including refugees from low literacy backgrounds.

We argue that better understanding among language testers of the values underpinning community arguments for and against the use of language tests in particular contexts is not only a key component of language assessment literacy but also critical to any advocacy work that language testers may engage in.

3. Integrating Vulnerable Groups of Migrants: Between Formal Tests and Formative Evaluation
Lorenzo Rocca, Università per Stranieri, Perugia, Italy

This contribution focuses on the central role of the Council of Europe (CoE) on the linguistic integration of vulnerable groups of migrants into Europe. Two aspects will be highlighted: First, the CoE’s influence in the critical reflection on the ethical use of language tests and the CEFR within the integration process of migrants will be illustrated by the CoE Recommendation 2034 on Integration tests: helping or hindering integration? (Parliamentary assembly: 2014). The second aspect considers the CoE’s facilitating role in supporting volunteers working with refugees. Here, a Language Evaluation Toolkit with 57 instruments available in 7 languages was designed in 2017. It aims at supporting volunteers working with refugees with tools for language teaching and formative evaluation.

As chair of the Language and Migration Special Interest Group (LAMI SIG) of the Association of Language Testers in Europe (ALTE) and member of Linguistic Integration of Adults Migrants (LIAM) in the CoE, I will focus on the ethical use of test results, as well as on the language and Knowledge of Society (KoS) requirements in the European countries (crucial themes in the LAMI Booklet, Language tests for access, integration and citizenship: an outline for policymakers: 2016). I will argue for the pressing need to provide different forms of non-formal evaluations, especially related to refugees. This approach is operationalized in the Toolkit which accounts for migrants’ plurilingual repertoire and their literacy profiles. I will explain the Toolkit’s rationale, structure and main contents, an instrument supporting volunteers in analyzing and supporting refugees’ linguistic needs.

4. Integrating Refugees into Higher Education In Europe – The Role of Language and Other Challenges
Peppi Taalas et al., University of Jyväskylä, SIMHE, Finland; Claudia Harsch and Anika Müller-Karabil, both University of Bremen, Germany

Over the past years, a large number of refugees have come to Europe, posing new challenges on educational integration efforts and language learning and assessment (e.g. Beacco et al., 2017). This joint presentation shares insights into two projects (University of Bremen, Germany / University of Jyväskylä, Finland), both of them investigating the journey of refugees into national Higher Education (HE) and the role language plays for it.

The project in Germany accompanies about 250 refugees during a language programme preparing them for HE and their first year in academia. Following a longitudinal mixed-methods approach, the study aims at evaluating a) participant’s individual language learning pathways during the programme via online learning diaries, summative and formative assessment data and interviews, and b) the effectiveness of the entire programme as perceived by participants, language tutors and teachers, and academic teachers via interviews and questionnaires.

The project in Finland focuses on developing an educational model to facilitate immigrants’ paths into academia, integrating university-level language education, content studies and civic and working life skills studies. Through instruments like an entry level language analysis scheme, individual study plans for language and content learning and personal guidance, the project partners aim to ensure that participants’ needs are met. The performance of the developed model will be investigated through impact evaluation.

This presentation will outline similarities and differences between the two projects concerning settings, approaches and research goals. It will shed light on challenges and present first research results regarding the effectiveness of the programmes.
Transformation and Transition: Four Perspectives from the South on Academic Literacy Assessment in Times of Change

Chair: Albert Weideman, University of the Free State, South Africa
Discussant: John Read, University of Auckland, New Zealand

Participants:
Alan Cliff, University of Cape Town, South Africa
Kabelo Sebolai, Stellenbosch University, South Africa
Mehdi Riazi, Macquarie University, Australia
Cassi Liardet, Macquarie University, Australia
Laura Drennan, University of the Free State, South Africa

Summary

While massification has affected universities globally, there are significant challenges for assessing and developing academic literacy at universities in the southern hemisphere. These may well have implications for other contexts. This symposium will highlight those challenges, not only in respect of prospective and new enrolments in universities, but also for the language demands of postgraduate study. At both levels, these demands call for appropriate assessments of language ability that are sensitive to the transformatory needs of both students and lecturers. They arise, furthermore, in a context of transition, where universities are changing specifically in respect of diversity in intake, and the subsequent challenge to provide relevant language support.

This symposium will therefore focus on issues of assessing academic literacy before entry to university, or after enrolling for the first time, or again just prior to making the transition to postgraduate study. A common thread in all four papers is that of making assessment not only relevant to, but useful for planning instruction, and aligning the language instruction with language development. At the same time, the papers each highlight from a different perspective the theme of the symposium, namely transformation, transition and change. They contribute singly and jointly to the goal of the symposium, which is to debate the challenges of assessing academic literacy at various levels and in different but comparable institutional settings. What is more, the presentations have cross-cutting sub-themes, as will be explained below. They have another common element: all refer to empirical studies, and the data yielded by such investigation, to illustrate their arguments.

Paper 1 has as its main issue the challenge of teaching academic literacy if one takes the kind of information yielded by academic literacy assessment seriously, while Paper 2 will make a case for generic assessments of academic literacy providing the basis of discipline-specific testing. Both therefore present us with views on the debate about generic versus field-specific tests, and on how one could tap into the diagnostic and other information to be found in generic tests of academic language ability for the purpose of teaching and language development. Both refer to the language assessment of undergraduate students, taking as their reference point pre-enrolment language tests, that might be used for access decisions, or post-enrolment tests, that are generally employed for placement purposes.

Paper 3 has as its focus how changes in institutional systems affect language assessment at university, and the accompanying need to assess language readiness. Paper 4 illustrates that careful assessment design could give us insight into the processes of first gathering and processing information, before being able to present what has been processed as new information. Taking the postgraduate level as their context, both therefore deal with preparedness to present information (mainly in writing).

There are other sub-themes that bring Papers 2 and 4 together, specifically how students can be managed and supported in transitional situations. Paper 1 and Paper 3, in turn, highlight the instructional challenges and opportunities for teachers who utilise the results of academic literacy assessments.

Keywords: academic literacy; pre-admission assessments; post-admission assessments; generic tests of language ability; discipline-specific tests of language ability; English-medium universities
Individual abstracts

1. A National Assessment of Academic Literacy and its Transformatory Role in the Development of Disciplinary Teaching
Alan Cliff, University of Cape Town, South Africa

For approximately 10 years now, the standardised National Benchmark Test in Academic Literacy (NBT AL) has been implemented in South African higher education at a significant number of institutions and for various purposes, which range from the selection of applicants for high-demand programmes to the diagnostic assessment of already-selected students’ generic academic literacy. This presentation focuses on the latter, offering an analysis of students’ academic literacy and its implications for disciplinary teaching. Past research has mostly been student-facing, focusing on the implications of scores for the placement of ‘at risk’ students in support programmes. This presentation is focused on the implications of student NBT AL scores for a transformatory teaching approach, i.e. a lecturer-facing focus. Recent research in South Africa has been aimed at engaging disciplinary academics, who teach at various levels and in a variety of disciplinary contexts, in the meaning and significance of the concept and construct of academic literacy. In essence, this research has challenged academics through two questions: (1) to what extent does disciplinary teaching involve the teaching of contextualised academic literacy; and (2) what does it take to develop students’ disciplinary academic literacy levels so that their learning is deepened by integrated, holistic teaching? The presentation offers some perspectives on how student-facing research data in these contexts might be used to raise lecturers’ reflective teaching practice approaches. Concluding comments will focus on a discussion of the affordances and challenges of using these kinds of student analytics for the improvement of teaching.

2. Generic Academic Literacy Testing: A Logical Precursor for Discipline-Specific Language Teaching and Assessment
Kabelo Sebolai, Stellenbosch University, South Africa

For more than a decade to date, South African universities have, in addition to Grade 12 results, been using tests of academic literacy for the admission and placement of students. Performance on these tests is intended to be indicative of generic levels of academic language ability among arriving students. The ultimate purpose of these tests, however, is to generate micro-level performance data that should inform academic language teaching and learning where necessary. The tests are, in this sense, designed to enable test users to know exactly which areas of language ability might need extra support or reinforcement for a particular group of students to boost their ability to cope with the discourse demands of university education and eventually, their chance of success at university study. The way that this support is currently theorised in South Africa is predominantly disciplinary context driven, a situation which has not left generic language testing and teaching unchallenged. Arguments have been advanced, in other words, that effective academic language support should be discipline specific because language ability is employed differently in different contexts. The aim of this paper is to demonstrate that generic academic language testing may be an ideal foundation for discipline specific language support and testing. It will use micro level performance on a test of academic literacy by students in different faculties at a South African university to show how it can guide discipline specific language instruction and assessment.

3. Reading-to-write: Assessing Writing Ability for Diagnostic Purposes in Higher Degrees Programs
Mehdi Riazi and Cassi Liardet, Macquarie University, Australia

The transition from coursework study to research is a challenge for students. Though coursework programs may include some research activities, the leap to higher degree research programs focussing on knowledge production through intensive research and writing activities is substantial. Students’ readiness for such a transition is judged based on their completion of prior degrees, as well as their language proficiency level. However, the use of conventional language proficiency tests at transition points (entry into program and post-entry language assessment) is controversial. This paper will refer to the way that an Australian university has dealt with these demands by first describing the system adopted for postgraduate study. There is currently no credit-based language program for higher degree research (HDR) students at Macquarie University. To ensure that these students’ academic and professional communication abilities contribute to their successful completion of their program is a major challenge. In 2017, a pilot initiative called the “Research Scholar Program” invited HDR entry students to sit for a diagnostic writing task and to then be coached to upskill their academic communication based on their performance on this task. In this paper, we will explain the design of the diagnostic reading-to-write task with reference to a sample of reflections on the task collected through post-task interviews. Challenges facing the definition of the diagnostic construct, the design of the
task as well as students’ perceptions and reflections will be discussed, as well as how those considerations might influence subsequent assessment and intervention designs.

4. **Measuring Students’ Academic Writing Readiness: A Discipline-Specific Assessment Design**
   Laura Dreenan, University of the Free State, South Africa

The massification of higher education in South Africa – unavoidable in light of correcting past injustices - may have contributed to what many perceive as a steady decline in the academic literacy and language proficiency levels of students being admitted to postgraduate studies, particularly in respect of writing ability. This calls for the development of language assessments that measure students’ ‘readiness’ to produce the kinds of writing required in particular subject areas at postgraduate level. Central to the conceptual understanding of a field is the ability to construct disciplinary arguments in community-specific ways. In response to this need, a discipline-specific test was designed to assess senior students’ readiness to write, or, more broadly, to present new academic information in a number of different ways, by means of an Assessment of Preparedness to Produce Multimodal Information (APPMI) as they transition from undergraduate to postgraduate-level studies. The paper discusses the theoretical justification for the construct underpinning the test design, as well as the selection and design of task types that measure students’ ability to gather and process information – key stages preceding the production of information. The pilot test results serve to support the argument that the test is based on a theoretically defensible construct and that the various subtests have the potential to deliver reliable results and valid interpretations. The test results will further inform the development of feasible and effective pedagogical resources in the form of academic writing interventions geared towards addressing weaknesses identified by the assessment.

### Symposium 4
**Thursday, 5 July 2018**
8:45am-10:45am
**Venue:** Lecture Theatre OGGB 4 (260-073)

**Re-conceptualizing, Challenging, and Expanding Principles of Test Validation**

**Chair:** Spiros Papageorgiou, Educational Testing Service, USA

**Participants:**
- Michael T. Kane, Educational Testing Service, USA
- Richard J. Tannenbaum, Educational Testing Service, USA
- Ikkyu Choi, Educational Testing Service, USA
- Jonathan Schmidgall, Educational Testing Service, USA

**Summary**

Argument-based approaches to validation have been influential in the field of language testing and have guided the development, revision, and evaluation of language tests (Chapelle, Enright, & Jamieson, 2008) or have informed validity frameworks that have become influential in language testing (Bachman, 2005; Bachman & Palmer, 2010; Fulcher & Davidson, 2009). Such approaches to test validation draw on Toulmin’s argument structure (Toulmin, 2003), which consists of making and supporting claims based on data (i.e., information or facts on which the claim is based). The Assessment Use Argument (AUA) framework (Bachman, 2005; Bachman & Palmer, 2010) has received particular attention in the language testing field, as it expands upon the basic foundations of earlier argument-based approaches to validation, in particular Kane (1992, 2006), by including criteria for the utility of assessment scores and the intended consequences as a result of using such assessments (Tannenbaum & Cho, 2014).

The argument-based approach facilitated an expansive and coherent approach to validation in which traditional concepts (e.g., reliability, validity, fairness) are concretely elaborated while aspects of their essential interdependence (e.g., reliability is a prerequisite to validity) are specified. Although initially conceived as a framework for justifying the use of a particular assessment for a particular purpose, the comprehensiveness of argument-based frameworks, such as the AUA, may facilitate more wide-ranging discussions of the issues associated with important topics in language assessment. Such issues are the focus of this symposium, which brings together four papers that touch upon the
theme of conceptualizing, challenging, and expanding principles of argument-based approaches to test validation. Each presenter brings a unique perspective to the symposium, through years of engagement with argument-based validation, either from a theoretical standpoint or a practical one. This combination of theoretical and practical perspectives is then summarized by the moderator with implications for LTRC attendees. More specifically, the first paper, given by a leading figure in test validation theory, presents an innovative way of thinking about test validation through the metaphorical lens of distinct properties of language. The second presenter focuses on what one could argue is the most important aspect of a language assessment for test users, that is, score reporting and in particular how reports can convey useful information about test performance and support intended interpretations and uses. The third presenter deals with the topic of evaluating validity arguments using statistical models and challenges current practices that distort such an evaluation, through illustrative examples. The fourth paper centers around the challenges for developing an AUA for language assessments in workplace contexts, and highlights the unique issues for construct definition for English in the workplace. Finally, the symposium moderator will summarize the main points raised by all papers and open the floor for discussion regarding the most important takeaways from the symposium. Taken together, these presentations provide suggestions for conceptualizing and implementing validity arguments based on the opportunities and challenges inherent in defining the interpretation and use of the scores and in developing the relevant evidence.

Individual abstracts

1: Evaluating Assessments in Terms of Syntax, Semantics, and Pragmatics
Michael T. Kane, Educational Testing Service, USA

The scores generated by assessment systems characterize individuals along certain dimensions and are intended to be used in certain ways. The scores are applied to particular objects of measurement (e.g., persons, groups of people, people at particular times or in particular contexts); the stipulated objects of measurement define the syntax of the scores as attributes of particular kinds of objects of measurement, and generalizability coefficients evaluate the extent to which the scores satisfy these syntactical requirements. The scores are interpreted in terms of their implications (e.g., claims about dispositions or constructs, predictions of future performances); the claims supported by the scores define the semantics of the scores, and the evaluation of these claims is a main focus of most validity arguments. Just as generalizability is part of validity, syntax is part of semantics. Assuming that we have confidence in the interpretations, the scores can be used to make decisions of various kinds (e.g., selection, placement, instructional) and these decisions are evaluated in terms of how well the scores work for the specific decisions to be made; the details of the decision process, its relationship to the assessment, and the impact of the context and irrelevant personal characteristics can be used to define the pragmatics of the score interpretation and use, and these factors are evaluated in terms of threats to the interpretation and use. Looking at validity arguments from this kind of linguistic perspective can provide fresh insights into the structure of validity arguments and a better understanding of the relationships between different kinds of evidence in validity arguments.

2: Validity Aspects of Score Reporting
Richard J. Tannenbaum, Educational Testing Service, USA

Scores and their intended meaning are communicated to stakeholders through some form of a score report, whether that report is paper-based or digital, static or dynamic. Stakeholders rely on the information presented in the report to guide their subsequent policies, recommendations, plans, and actions. The likelihood that such score-based decisions are reasonable and appropriate is dependent, in part, on the relevance and accuracy of the reported scores (and related information) with regards to the constructs being measured and the students’ standing on those constructs, and the ability of the stakeholders to comprehend the reported information in the way intended. As one considers that validity is inextricably tied to evidence supporting score interpretation and use, it is clear that we need to conceptualize and construct score reports purposefully to facilitate intended and desired interpretation and use. In this talk, I will discuss some of the key concepts and practices that are intended to support the validity (meaningfulness) of score reports.

3: Challenges in Evaluating Validity Arguments with Statistical Models
Ikkyu Choi, Educational Testing Service, USA

Statistical models are often used to evaluate validity arguments based on data. However, the nature of statistical models we use is not always well-aligned to the data we have access to. This misalignment brings about unique
challenges in properly evaluating validity arguments using statistical models. In this talk, I will present three such challenges that are particularly relevant to the validation of interpretations and uses of language assessments. Few statistical models can directly incorporate the precision of scores on language assessments, and, as a result, output from statistical models often presents a misleading picture of available information. In addition, sometimes data that are crucial for evaluating specific validity arguments are difficult to collect or do not even exist, limiting the credibility of model-based inferences we make. Lastly, the grain size of a statistical model is dependent upon the available data and validation task at hand, and therefore, it is difficult to establish a general statement and/or consensus about the usefulness of a model for a validation task. Each of these challenges will be illustrated with a case study that highlights how the challenges can distort the evaluation of validity arguments. I will conclude with potential ways to address these challenges through careful data collection design and model evaluation.

4: Challenges in Defining the Domain of Workplace English as Part of a Validity Argument
Jonathan Schmidgall, Educational Testing Service, USA

Many language assessment researchers advocate an interactionalist approach to construct definition in which context is part of the construct (e.g., Chalhoub-Deville, 2003; Chapelle, 1998). Although proficiency in English for academic purposes has been carefully considered from an interactionalist perspective (e.g., Xi, 2015), English for the workplace poses distinct and complex challenges for construct definition. Given that the workplace or occupational domain is among the most heterogeneous that can be defined, one fundamental issue is construct and domain specificity (e.g., English for the workplace vs. English for tourism). A second fundamental issue is how one incorporates the notion of repertoires (e.g., via paradigms such as English as a Lingua Franca and World Englishes). How one articulates and reconciles these issues has strong implications for whether and how much context is represented in the construct, as well as the implied goal of attainment (e.g., native-like proficiency, comprehensibility to native/local community, comprehensibility to a broader community). This presentation focuses on how these fundamental issues may impact construct definition for English in the workplace, and uses Bachman and Palmer’s (2010) assessment use argument framework to highlight aspects of test design and validation that may be of interest or concern to test developers and researchers focused on this domain.

Symposium 5
Friday, July 6
8:45am – 10:45am
Venue: Lecture Theatre OGGB 4 (260-073)

China’s Standards of English Language Ability: Impetus for Change in Language Learning, Teaching and Assessment

Chair: Han Yu, National Education Examinations Authority, China
Jianda Liu, Guangdong University of Foreign Language Studies, China
Discussant: Yan Jin, Shanghai Jiao Tong University, China
Participants:
Lianzhen He, Zhejiang University, China
Wenxia Zhang, Tsinghua University, China
Sha Wu, National Education Examinations Authority, China
Barry O’Sullivan, British Council, UK
Nick Saville, Cambridge Assessment English, UK

Summary

China’s role on the world stage has seen a shift in its position as a significant global diplomatic power. This has been most publically seen in the recently launched One Belt One Road Initiative, which has been described by the UK’s Guardian newspaper (May 12, 2017) as “an immensely ambitious development campaign through which China wants
to boost trade and stimulate economic growth across Asia and beyond.” To support China’s development in the new era with talents of high quality, important changes have been implemented across the education system in China. Especially in foreign language education, great attention is given to the introduction of formative assessment and a shift away from teaching to the test to the cultivation of core communicative competencies.

In 2014, the State Council of China decreed that a new assessment system of foreign languages should be constructed to generate a comprehensive reform in foreign language education in China. As an important part of the proposed system, the development of the China’s standards of foreign languages was commenced in 2014, and since English was the foreign language with the largest population of learners in China, the China’s Standards of English Language Ability (CSE) became the initial focus.

The CSE, specifically related to the Chinese EFL context, is expected to provide a set of transparent and consistent standards of English proficiency to enhance the communication between English teaching, learning and assessment. To be released by the Ministry of Education in early 2018, the CSE is hoped to set an example for the development of standards for other foreign languages.

This symposium aims to present the background, theory, and empirical studies relating to the development of the CSE, together with the implications and applications of the CSE both inside and outside China. Four presentations are prepared to cover the topic from the following perspectives.

1. the background, the rationale, and the theoretical framework of the CSE;
2. the development process of the CSE exemplified by a case study of listening;
3. the potential impact of the CSE in the Chinese education system;
4. a set of procedures to be used when claiming a link to the CSE by local and international tests.

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

**Individual abstracts**

1. **China’s Standards of English Language Ability: A Tale of the New Era about English Language Education in China**
   Liu Jianda, Guangdong University of Foreign Language Studies, China

   The influence of language frameworks and standards on curricula and examination reform has been acknowledged in different areas around the world (see North, 2014). China, which boasts the largest population of EFL learners, is not immune to the sweeping impact of such educational initiatives. With an increasing awareness of the different functions that standards serve in education reform, language educators, teaching practitioners, as well as policymakers in China reached a consensus that a set of unified language standards were urgently needed to describe learners’ performance and streamline their competence across different educational stages and different regions in the Chinese EFL context.

   In response to the State Council’s document, the Ministry of Education initiated a nationwide project to develop an English language ability scale, known as the China’s Standards of English Language Ability (CSE). This project set out to define and describe the English proficiency of the English learners in China, to provide references and guidelines for English learning, teaching and assessment, and to enrich the existing body of language proficiency scales for alignment initiatives on a global basis in the future.

   This paper first introduces the socio-political and educational background and rationale of the CSE, followed by an explanation of the theoretical framework underlying its development. Various considerations concerning the development of the CSE are then presented. It ends with an elaboration of the construction processes and relevant frameworks.

   Reference

2. **The Development of the CSE—a Case Study of Listening**
   He Lianzhen, Zhejiang University, China

   Taking a use-oriented approach based on Bachman’s (1990) communicative language ability (CLA) model, the China’s Standards of English Language Ability defines language ability as the ability to use language to create and interpret meaning. The CSE includes eight language ability components which are in practice in the Chinese context of teaching, learning and assessment, namely listening, speaking, reading, writing, translation, interpreting, grammar and pragmatics. Following the format of “can-do statement” for descriptors used in the CEFR, the CSE adopts a three-element model for skills descriptors, namely, performance, criteria and conditions, with the first two as essential
elements. Adhering to the four principles of being scientific, specific, practical and operational, the development of the CSE went through the stages of descriptor collecting, sorting, classifying and validating with the help of an online platform, involving over 160,000 teachers, students, professionals, etc. in 28 provinces, municipalities, and regions. As a result, the CSE will include around 100 scales.

In this paper, following an overview of the CSE, a case study of the development of the listening scale will be presented, which will include the definition of listening construct, the development of an operational descriptive scheme of listening ability, the collection of descriptors by means of documentation and sampling, the classifying, scaling and validating of the descriptors. Finally, salient features of the different levels of the listening scales will be presented in such aspects as speech rate, phonological features, topics, cognitive processing and conditions.

Reference

3. The Potential Impact of the CSE on EFL Education in China: Issues and Challenges
Zhang Wenxia, Tsinghua University, China
Wu Sha, National Education Examinations Authority, China

The China’s Standards of English Language Ability is the first English proficiency scale in China to cover the full range of EFL learners. It aims to improve the reliability and validity of English tests in China, to facilitate the comparison of test results, and to promote meaningful communication between teaching, learning and assessment across different educational stages. Thus, its impact is likely to spread gradually to reach a broad group of stakeholders, e.g. teachers, students, examiners, test providers, policy makers, users at workplaces, etc. To achieve the objectives of the CSE, follow-up research is clearly needed to track its real effects and to guide further initiatives to ensure a positive impact. In this paper, we first describe the educational context of China to identify the key stakeholders for an initial impact study. Based on a survey launched before the CSE’s official release, we discuss the impact of the CSE anticipated by the stakeholders concerned, and what form or forms this might take. We also look into the issues and challenges the CSE may face during its application. It is envisaged that this impact study will be an on-going and dynamic process, and we conclude with a tentative framework for follow-up research.

4. Principles and Procedures Underpinning the Linking of Tests to the CSE
Barry O’Sullivan, British Council, UK
Nick Saville, Cambridge Assessment English, UK

The introduction of the Common European Framework of Reference for Languages (CEFR) in 2001 has had a significant impact on language learning and assessment in Europe and beyond. The publication of the Draft Linking Manual in 2003 (final version in 2009) saw the emergence of interest in the process of establishing evidence in support of any claim of a link between curricula, syllabi, learning materials or tests to the CEFR. Despite criticisms of its failure to highlight the need for an explicit theoretical foundation in the form of an appropriate validation model, the focus of the Manual on the systematic and structured gathering of evidence in support of any linking claim is widely accepted to have positively influenced learning and assessment practice in Europe and beyond.

In this paper, we set out the theoretical and philosophical rationale that supports the approach adapted by the CSE implementation team when deciding how local and international test developers should demonstrate an empirical link between their test and the CSE. The theoretical underpinnings of the CSE (Liu, this symposium) suggested that the socio-cognitive model of test development and validation was the most appropriate model to be used in the conceptualization of the phases of any linking process. The paper summarises the discussions held by the project steering and working groups which led to the proposal of a four-phase linking model: Construct Phase, Measurement Phase, Criterion Phase, and Decision Phase. These four phases will be presented and preliminary linking claims for two major international examinations discussed along with broader implications for the future international impact of the CSE that emerged during the project.
Symposium 6
Friday, July 6
8:45am – 10:45am
Venue: Lecture Theatre OGGB 5 (260-051)

Repeated Test-taking and Longitudinal Analysis of L2 Test Score Data

Chair: William J Bonk, Pearson, USA
      Alistair Van Moere, Metametrics, USA
Discussant: Tony Green, University of Bedfordshire, United Kingdom

Participants:
Yeonsuk Cho, Educational Testing Service, USA
Ian Blood, Educational Testing Service, USA
Sean Hanlon, Metametrics, USA

Summary

The measurement and tracking of learning gains over time using standardized assessment data is an under-researched area in second language testing. As Barkaoui (2013) points out, many validation studies examine factors that contribute to variability in single assessment instances, but few investigate variability longitudinally by examining score changes over time. Yet this area requires more attention if we are to understand not only sources of score variability, but also learners’ development of language skills over time.

Various researchers have investigated repeated test-taking under pre/post conditions. Ling, Powers and Adler (2014) analyzed score data from 111 learners who took two TOEFL iBT practice tests with English programs in the intervening 6-month period. They determined that learners improved their proficiency at least moderately, and that more hours spent studying was associated with greater score gains. Learners made gains differentially on the skills emphasized during their course of study – those based in China gained more in receptive skills such as reading, while those studying abroad in the US increased their speaking skills more. Sawaki (2017) investigated whether instructional feedback on reading tasks would improve TOEFL iBT reading scores, but although the 193 participants increased scores on average, there were no effects associated with amount of feedback they received during intervention. However, familiarity with the test itself was identified as a predictor positively related to test score gain.

Other repeated measures studies did not involve experimental pre/post conditions. Green (2005) compared the writing scores of 15,380 candidates who each took IELTS twice over approximately three years. He found that writing scores on test occasion 1 was a better predictor of writing score on test occasion 2 than was the interval of time between the two tests; learners with lower scores on occasion 1 made more rapid gains than learners with higher scores on occasion 1. Similarly, Zhang (2008) investigated scores of 12,385 candidates who repeated TOEFL IBT at least once within a 30-day period. Scores improved slightly on the second testing occasion; reading scores improved the most, while speaking scores improved the least.

These studies highlight the challenge of understanding which factors lead to score changes, as well as their magnitudes. Practice effects, complexity-accuracy-fluency trade-offs, differential growth in skills, and time elapsed between tests may all play a role. Furthermore, if exams are used to quantify learning gains over a course of study, as, for example in the case of APTIS (British Council) or Progress (Pearson), then apparent proficiency gains must be teased apart from measurement error.

This symposium further addresses patterns of change on test performance, and individual differences in change patterns over time. These questions can best be answered when at least three repeated measures are available for each learner. When such data is available multilevel modelling can be employed, which allows for within-person dependency. Thus, rather than treating each learner’s data as an independent snapshot, a learner’s repeated measures are nested within that individual. This method can be used to investigate patterns in longitudinal data over sustained periods of time.
1. A Bayesian Approach to Improving Measurement Precision over Multiple Test Occasions

Alistair Van Moere and Sean Hanlon, Metametrics, USA

In language assessment and in educational measurement more broadly there is a tendency to interpret scores from single-administration tests as accurate indicators of a latent trait (e.g. reading ability). Even in contexts where learners receive multiple formative assessments throughout the year, estimates of student ability are determined based on the most recent assessment. However, in other fields measures are often viewed as estimates, which are refined as more data become available; for example, US Bureau of Labor Statistics corrects its published employment data three times as new data becomes available. Applying a similar approach, this paper demonstrates a technique that incorporates prior test scores with current scores for learners who re-test periodically, in order to arrive at an estimate closer to the learners’ true score. Approximately 2,400 learners from two separate studies (EFL and native speaker) were tested for reading proficiency between 3 and 5 times each, over a one- to two-year period, on a multiple-choice reading test which reported reading ability in Lexile® measures. Applying Bayes theorem, prior scores and the most recent test score were combined with uncertainty parameters (i.e. measurement error) to produce new estimates of student ability. This is advantageous as prior administration data is re-used rather than discarded, and the impact of a single measurement occasion is reduced. This Bayesian approach is discussed in the context of low-stakes tests designed to measure proficiency gains over time, as well as frequently-taken high-stakes tests that allow candidates to report their highest score on job or university applications.

2. Effects of Multiple Repeated Tests and Skill Trade-offs On an Automatically Scored Speaking Test

William J Bonk, Pearson, USA

In assessment of speaking proficiency, candidates are unlikely to take face-to-face interview-style speaking tests repeatedly because they are costly and complex to administer. But, fully automated speaking tests with large item banks can be made available to learners on-demand for frequent self-evaluation. Data were obtained from 220 adult learners who had unlimited access to Versant English, a computerized testing system for speaking proficiency. Each learner took at least 15 tests per year for between one and three years, with an average of approximately 40 tests over two years. A subset of 31 learners additionally took tests in “bursts,” for example, 10 tests or more within 10 consecutive days. Multilevel modelling was applied to answer the questions: (i) How much do candidates gain in speaking scores over 1-3 years? (ii) How much do they gain over a ten-day time period? (iii) Do candidates exhibit evidence of trading off among competing enabling skills (e.g. increasing fluency at the expense of accuracy)? Learners who took many tests frequently over 10 days made score gains approaching those of adults during one year of working in an English-speaking environment (approximately one-third of a CEFR band). It is hypothesised that repeated practice and feedback on a task leads to enhanced performance, but that this is a short-term benefit which is not likely to be sustained over the long term. Moreover, there was evidence of cognitive trade-offs in repeated speaking performances, confirming findings of other authors (Skehan, 1998). The data are discussed in the context of other repeated test-taking longitudinal studies.

3. Assessing Young EFL Learners’ Language Proficiency: How Much Score Change Should You Expect?

Yeonsuk Cho and Ian Blood, Educational Testing Service, USA

In this study, we examined how much change in TOEFL® Primary™ listening and reading scores can be expected in relation to the time interval between test administrations. The TOEFL Primary tests are standardized English-language proficiency tests for young learners (8 years of age and older) who are learning English as a foreign language (EFL). The test records of 5,213 test takers in Japan and Turkey who repeated the tests were analyzed to examine score change as a function of time interval. The effect of time on score change was analyzed with a multilevel modeling approach, allowing both initial scores and rate of change among individual test takers to vary. In addition, we examined the effects of test-taker age and TOEFL Primary test level (i.e., Step 1 vs. Step 2) on score change. Separate analyses were conducted by country for ease of interpretation, because the countries represented in the study sample were quite different in terms of the amount of instruction that learners receive and the English-learning goals in their respective curricula. Results showed that the rate of change was positive and statistically different from zero, indicating that test scores increase gradually over time. However, the rate of change differed between the two countries. Furthermore, results indicated that repeaters’ test scores increased with their age and with the length of time between test administrations. Findings from the current study provide empirical evidence for schools to refer to when determining the timing of re-administration of the TOEFL Primary tests to their learners.
Learning-oriented Assessment Feedback for Interactional Competence: Developing a Checklist to Support Teachers and Learners

Lyn May, Queensland University of Technology, Australia
Fumiyo Nakatsuhara, CRELLA, University of Bedfordshire, United Kingdom
Daniel Lam, CRELLA, University of Bedfordshire, UK
Evelina Galaczi, Cambridge Assessment English, UK

Learning Oriented Assessment (LOA) foregrounds learners, learning processes, and learning outcomes (Purpura & Turner, 2014). Appropriate and timely feedback which learners can also feedforward is essential to support both current and future student learning (Carless, 2007). Although the challenge is to connect large-scale examinations with meaningful learning opportunities in language classrooms where test preparation is undertaken, recent advances in technology in the field of LOA have made it possible for teachers in test preparation classes to provide more informed and individualised feedback in a speedy and less labour-intensive manner (Jones & Saville, 2016).

However, in the case of interactional competence (IC), this potential seems yet to be fully realised. Existing test preparation materials and online LOA resources provide little guidance for teachers to feedback on learners’ interactional skills compared to other speaking sub-skills. This seems to relate to the somewhat ‘evasive’ definition of IC which is hard to pin down, despite a burgeoning body of research on assessing IC.

This research was motivated by the need for the teaching and learning of IC to benefit from technological advances in the field of LOA, and therefore offers implications in relation to the LTRC sub-theme of ‘changes in language learning and assessment’, specifically in terms of facilitating the use of technology in the teaching and assessment of IC.

This project developed practical tools to support the classroom assessment of learners’ IC and provide learning-oriented feedback in the context of Cambridge English: First. To develop a checklist, accompanying descriptions and recommendations for teachers to use in providing feedback on learners’ interactional skills, 72 stimulated verbal reports were elicited from six trained examiners who were also experienced teachers. They produced verbal reports on 12 paired interactions with high, mid, and low IC scores. The examiners were asked to comment on features of the interaction that influenced their rating of candidates’ interactional competence and, based on the features of the performance they noted, provide feedback to candidates. The verbal reports were thematically analysed using NVivo12 to inform a draft checklist and materials, which were then trialled by four experienced teachers in order to further refine these resources. The final product comprises (1) a full IC checklist with 8 main categories and over 50 sub-categories, accompanying detailed description of each area and feedback to learners, and (2) a concise version of the IC checklist with fewer categories and ‘bite-sized’ feedback to learners.

We discuss some aspects of operationalising these tools to support teachers and students, such as a) the ways in which educational technology can be used to select and deliver the relevant feedback electronically, and b) the complementary roles of the full and concise versions for teaching and real-time assessment to cater for teachers’ and students’ needs. It is hoped that the easy-to-use but empirically-based tools can be useful LOA resources which help teachers to more deeply understand the construct of IC, and to integrate learning and assessment in meaningful and innovative ways.
Examining Academic Speaking Ability and its Cognitive Dimension through a Scenario-based Speaking Assessment

Yuna Patricia Seong, Teacher College, Columbia University, USA

From the skills-and-elements approach to the more recent models of communicative language ability, the construct of second language (L2) proficiency has evolved and broadened over the past few decades (Bachman, 2010; Purpura, 2016). In line with this conceptual change and fueled by advances in digital technology, scenario-based assessment (SBA) has been utilized and examined by language testers (e.g., Sabatini, O’Reilly, Weeks, & Steinberg, 2016) seeking current and innovative assessment practices that can reflect the modern-day language use while appropriately measuring language proficiency that accounts for linguistic as well as non-linguistic factors (e.g., topical knowledge or strategy use).

The purpose of this study was to examine academic speaking ability and its cognitive dimension using an online Scenario-based Academic English Speaking Test (SBAEST) to better capture a broadened construct of academic speaking ability that includes cognitive thinking processes. L2 academic speaking ability not only calls for students’ communication of disciplinary knowledge, but it also involves integrated use of metacognitive and cognitive strategy use in order to successfully process, synthesize, and utilize the information (Chamot & O’Malley, 2004; Zweir, 2008) to perform complex academic speaking tasks such as participating in a class discussion or giving a presentation. SBA allows test takers to demonstrate their language proficiency in a meaningful and goal-oriented context by performing thematically related tasks sequenced in a way that simulates real life problems. The computer-based SBAEST for the current study was designed to replicate real-life academic speaking demands. Students must listen to audio and video materials on a given topic and orally respond to questions or share their opinions by summarizing and synthesizing the information from the materials. In addition to the speaking tasks, the test included strategy tasks designed to specifically elicit students’ use of cognitive strategies (e.g., planning, predicting, and recalling key points).

This study examined the nature of academic speaking ability and its cognitive dimension by studying the takers’ performance on the strategy tasks and its relationship with speaking performance. Thirty-two high-intermediate to advanced ESL learners took the test, and the test results were analyzed using many-facet Rasch measurement (MFRM). Students’ responses on the strategy tasks were scored based on qualitative examination of the patterns or characteristics indicative of effective or ineffective strategy use. Relationship between the students’ performance on strategy tasks and speaking tasks was analyzed using correlational analyses. Results indicated that strategy use is an integral component of the academic speaking ability construct, and some of the interesting patterns and characteristics of cognitive strategy use were reviewed in relation to the test takers’ speaking performance. Further suggestions on improving test design and implications of using SBA for assessing academic speaking ability will be discussed. Assessments such as the SBAEST may have implications for the diagnosis and assessment of student learning needs, specifically with respect to distinguishing between linguistic and cognitive aspects of academic speaking ability.

EMI Classroom Assessment Practices: A Survey of Taiwanese Tertiary Teachers

Jessica Wu, The Language Training & Testing Center, Taiwan
Naihsin Li, The Language Training & Testing Center, Taiwan

While assessment is part of teaching, it is often problematic in EMI (English-Medium Instruction) for those who teach non-native speakers of English via the medium of English (Dearden, 2014). Aiming to help EMI teachers to meet their assessment needs, the EMI Classroom Assessment Practices (ECAP) questionnaire was developed and administered to 40 EMI instructors teaching at the tertiary level in Taiwan.

The questionnaire surveyed university teachers regarding their assessment practices in EMI, and their self-perceived skills in these practices. It also investigated the effect of instruction medium on the assessment practices among the teachers who had conducted a similar course using Chinese as the medium of instruction. The responses were
analyzed to identify teachers’ needs for training and support in EMI classroom assessment and to explore the role of students’ L1 (Chinese) and L2 (English) in the EMI assessment. Furthermore, the relationships between questionnaire responses and teacher-/course-related variables were examined. Results showed that the teachers have, in general, practiced various assessment methods to evaluate student progress. However, the practices were mediated by some teacher-/course-related variables, such as discipline, experience of staying in an English-speaking country, and course size. Instruction medium also showed an effect. In particular, the following learning-oriented assessment practices were less often performed in EMI classrooms: making appropriate alignments between assessment and instruction goals; communicating assessment results to students and providing feedback; and, assessing students’ higher-level cognitive abilities. A high correlation was found between the assessment techniques that were less practiced and those that the teachers felt less skilled in, signaling the possibility that the assessment practices in the EMI classroom were dominated by the teachers’ familiarity with certain kinds of assessment techniques, rather than by instructional objectives. In addition, it was found that teachers perceived a critical gap in their skills in designing test items that assess high-level cognitive abilities, despite the fact that they frequently do so. In terms of students’ L1 and the role of English in the EMI classroom, almost all EMI teachers were aware of students’ difficulties with English and used simple English when presenting the assessment task. However, it was not common for them to take into consideration and make accommodations for students’ language difficulties when considering task performance; neither did they find it necessary to provide corrective feedback on language errors. Nevertheless, teachers who incorporated language improvement as one focus of assessment more often practiced these strategies. The findings of the investigation have implications not only for advancing our understanding of teachers’ assessment needs in the Taiwanese EMI context but also for developing and implementing useful training programs in support of EMI teachers’ professional development with regard to assessment. Last but not least, the contextual features of EMI assessment identified through the study may serve as a point of departure for the definition of good assessment practices for EMI.

Wednesday, July 4
10:30am - 11:00am
Venue: Case Room 4 (260-009)

**Measuring the Construct of Spoken Collocational Competence: Theoretical Underpinnings, Operational Measurement, and Implications**

Jing Xu, Cambridge Assessment English, UK

Collocation, known as words that commonly co-occur, is a major category of formulaic language. There is now general consensus among language researchers that collocation is essential to effective language use in real-world communication situations (Ellis, 2008; Nesselhauf, 2005; N. Schmitt, 2010; Wray, 2002). Although a number of contemporary speech-processing theories assume the importance of formulaic language to spontaneous speaking (Bygate, 1987; de Bot, 1992; Kormos, 2006; Levelt, 1999), none of them give an adequate explanation of the role that collocation plays in speech communication. In the practices of L2 speaking assessment, a test taker’s collocational performance is usually not separately scored mainly because human raters can only focus on a limited range of speech characteristics (Luoma, 2004).

This paper argues for the centrality of collocation evaluation to communication-oriented L2 oral assessment. Based on a logical analysis of the conceptual connections among collocation, speech-processing theories, and rubrics for oral language assessment, the author formulated a new construct called Spoken Collocational Competence (SCC). In light of Skehan’s (1998, 2009) trade-off hypothesis, he developed a series of measures for SCC, namely Operational Collocational Performance Measures (OCPMs), to cover three dimensions of learner collocation performance in spontaneous speaking: collocation accuracy, collocation complexity, and collocation fluency. He then investigated the empirical performance of these measures with 2344 lexical collocations extracted from sixty adult ESL learners’ real L2 oral assessment data collected in two distinctive contexts of language use: conversing with an interlocutor on daily-life topics (or the SPEAK exam) and giving an academic lecture (or the TEACH exam). Multiple regression and logistic regression were performed on criterion measures of these learners’ oral proficiency (i.e., human holistic scores and oral proficiency certification decisions) as a function of the OCPMs. The study found that the participants generally achieved higher collocation accuracy and complexity in the TEACH exam than in the SPEAK exam. In addition, the OCPMs as a whole predicted the participants’ oral proficiency certification status (certified or uncertified) with high accuracy (Nagelkerke $R^2 = .968$). However, the predictive power...
of OCPMs for human holistic scores seemed to be higher in the SPEAK exam \( (\text{adjusted } R^2 = .678) \) than in the TEACH exam \( (\text{adjusted } R^2 = .573) \). These findings suggest that L2 learners’ collocational performance in free speech deserve examiners’ closer attention and that SCC may contribute to the construct of oral proficiency somewhat differently across speaking contexts. Implications for L2 speaking theory, automated speech evaluation, and teaching and learning of oral communication skills are discussed.

### Assessing Interactional Competence in Contexts of Movement and Change

Jayanti Banerjee, Paragon Testing Enterprises, Canada
Noriko Iwashita, University of Queensland, Australia
India Plough, Michigan State University, USA

An ever-increasing body of research in second language (L2) assessment and second language acquisition has led to significant theoretical and methodological progress in the assessment of speaking proficiency. In particular, evidence from studies of speaking tests employing a face-to-face format support the claim that interactional competence (IC) is an essential component in the co-construction of dialogue. Simultaneously, technological advances have resulted in computer-mediated tests (TOEFL iBT), a fully-automated speaking test (Pearson PTE) as well as the exploration of video-conferencing for the delivery of ‘face-to-face’ speaking tests. It is not clear how this expansion in speaking test delivery modes accounts for and captures interactional competence as a feature of the speaking construct. In contexts of movement and change, such as immigration or entry to post-secondary study, a test score and its meaning can have substantial consequences for the test taker and other stakeholders. In these contexts, it is crucial to discuss the relevance of interactional competence to the speaking construct and how it might be operationalized. We argue that the field must consider the incorporation of IC into the speaking construct.

We begin with a brief review of the range of speaking tests currently available, discussing the tension between new approaches to delivering and scoring speaking tests and the operationalization of key aspects of the speaking construct such as the ability to develop, support, and maintain interaction. To advance the proposal for the inclusion of IC in the speaking construct, we present a schematic, “Creating Interaction.” The model brings together research in L2 speaking assessment as well as the work in first and second language gesture studies in order to portray the relationships between verbal features, nonverbal behavior, and communicative functions. Analyses of candidates’ behavior during the speaking portion of CaMLA’s Examination for the Certification of Proficiency in English will be used to explain and exemplify the model. For example, interconnections have emerged in a comparison of the verbal interaction features and the nonverbal behavior of candidates to the ratings awarded. Higher rated speakers tended to exhibit greater use of verbal features associated with turn management and turn development as well as increased use of nonverbal behavior associated with engagement.

In the final part of the paper we will return to the different delivery mode options for speaking tests and discuss how a computer-delivered test might gather evidence of (and evaluate) interactional competence as well as how a computer-mediated test (such as a test delivered via video conferencing) might interpret evidence of interactional competence. In this part we will focus on the evidential links that will need to be made between the speaking performance elicited and the claims about what the scores mean.

### Validating the Retelling Task of Test for English Majors Band 4 Oral in China: Evidence from a Corpus-based Exploration

Xiaoxiao Kong, University of Melbourne, Australia

The Test for English Majors Band 4 Oral (TEM4-Oral) is a nationwide English speaking test for undergraduate English major students in China. It is administered at the end of the second year of the undergraduate degree, which is the
transition point for students to progress from the learning of foundational English abilities (e.g. speaking, writing) to more advanced language knowledge (e.g. English linguistics and literature) (National Foreign Language Teaching Advisory Board, 2000). Despite its wide use and crucial place in the education system, little validation work has been carried out. In the first task of TEM4-Oral, ‘retelling a story’, the provision of a source story which candidates are instructed to retell raises questions about the use of source material and how it is linked to language proficiency (e.g. Plakans & Gebril, 2012).

This research is a validation study of the TEM4-Oral retelling task through the analysis of features of test-taker discourse across four test administrations. It aims to address the relationship between test-taker performance and scoring of the TEM4-Oral retelling task as well as its consistency of design and scoring across task versions. In order to investigate how the emphasis on ‘content’ in the marking criteria was reflected in the language features of candidates’ retellings across scoring levels, corpus analysis and qualitative explorations were carried out on discourse features at three lengths – single words, multi-word sequences, and longer stretches of language denoting segments of content. Quantitative corpus analysis suggested that higher- and lower-ranked retellings differed significantly in the frequencies of certain lexical items used and the number of multi-word sequences included. A further qualitative coding of the retold content produced by the candidates showed that compared to lower-ranked retellings, not only did higher-ranked retellings included significantly more content, those segments of content also resembled more closely the content of the source story. These findings partly support the score interpretation from test-taker performance in the TEM4-Oral retelling task. Across task versions, inconsistencies were found regarding the language features of the retellings as well as the scoring of them. This indicates that the generalisability of the task is threatened.

In terms of task design, discourse features of candidates’ retellings across administrations have pointed out inconsistencies in the difficulty levels of the source stories. Language features associated with the higher rank suggested that the scoring of the TEM4-Oral retelling task favored memorisation and replication, at the same time hindering the assessment of comprehension and inducing construct-irrelevant variance. In order for TEM4-Oral to serve as a better assessment of spoken English proficiency, the source text, instructions and marking criteria of the TEM4-Oral retelling task need to be reconsidered so as to minimise the identified issues, improve the defensibility of the task and provide positive washback for the oral English learning of undergraduate English major students in China.

Wednesday, July 4
11:05am -11:35am
Venue: Case Room 3 (260-055)

Classroom Assessment Literacy for EFL Teachers: Conceptualisation, Operationalisation, and Validation
Peter Gu, Victoria University of Wellington, New Zealand

Research on language assessment literacy has examined the knowledge, skills, and principles needed for various stakeholders (Taylor, 2013). Assessment literacy for language teachers has been explored from either a knowledge-base perspective for teacher education or the assessment training needs of language teachers (Fulcher, 2012). Most of these conceptualisations of assessment literacy cover a wide range of capacities relevant to language testing and assessment. The language teacher’s ability in making use of assessment for teaching and learning purposes, an aspect of assessment literacy crucial for the usefulness of classroom assessment, needs more research attention.

This presentation outlines the process in creating and validating a Classroom Assessment Confidence Index (CACI) for EFL teachers. I will first present a framework for conceptualising ‘classroom assessment competence’, followed by descriptions of the design and validation of the CACI.

The conceptual framework was developed based on a comprehensive review of theories and research on assessment literacy and classroom assessment, from classic conceptualisations of assessment literacy such as the Standards for teacher competence in educational assessment of students by the American Federation of Teachers, the National Council on Measurement in Education and the National Education Association (1990) and Stiggins (1995), to more recent reformulations such as Heritage (2007), Brookhart (2011), Willis, Adie, and Klenowski, (2013), and DeLuca, LaPointe-McEwan, and Luanga (2015). The framework was also informed by the latest applied linguistics research on classroom assessment practices (e.g., Davison & Leung, 2009; Rea-Dickins, 2001; Rea-Dickins & Gardner, 2000; Taylor, 2009).

A 40-item, Likert-scale CACI was created based upon the conceptual framework and upon an analysis of existing assessment literacy measures (Gotch & French, 2014), aiming to elicit EFL teachers’ self-assessment of their own
confidence in classroom assessment practices. Two hundred and eighty three (283) secondary school teachers teaching English as a foreign language in China participated in the construct validation survey. Exploratory factor analysis revealed 6 factors: 1) Believing in classroom assessment; 2) Having clear standards before assessment; 3) Ability to choose and design assessment tools; 4) Ability to evaluate assessment results and diagnose problems; 5) Ability to adjust teaching and learning; 6) Ability in ethical assessment. Twenty six (26) items remained after removing items that did not load highly onto these six factors. The 6-factor model was then subjected to a confirmatory factor analysis. Model-fit statistics indicated a reasonably good fit between the 6-factor CACI model and the observed data. Satisfactory values were also obtained for the convergent and discriminant validity of the CACI.

Suggestions and discussions for using and interpreting the CACI for teachers and teacher education programmes will next be presented.

Wednesday, July 4
11:05am - 11:35am
Venue: Case Room 4 (260-009)

Examining the Effect of Delivery Mode on Scores in the Speaking Domain of ACCESS for ELLs 2.0
Kyoungwon Bishop, WIDA at University of Wisconsin Madison, USA
Shinhye Lee, WIDA at University of Wisconsin Madison, USA
Howard Gary Cook, WIDA at University of Wisconsin Madison, USA

ACCESS for ELLs 2.0 (ACCESS 2.0) is a large-scale English language proficiency assessment administered to K-12 English language learners (ELLs) in 39 states, districts and territories in the United States. Since 2015, the speaking domain test of ACCESS has been administered in two different delivery modes: paper and online forms. In the paper mode, ELLs speak their responses to local educators upon listening to pre-recorded speaking prompts, and the same educators score their speech. By guidance teachers are not supposed to score their students’ speaking tests. In the online mode, ELLs record their responses on a computer, which are rated by centrally trained scorers. Since the inception of both modes, substantial score differences have been found with effect size difference, favoring the paper mode, ranging from .4 to 1.0 across grades.

Consistent gaps in scores have led to questions such as: (1) Are there differences in the test administration settings that cause the score discrepancies? (2) Are there differences in the scoring behaviors between the local administrators and the trained raters? If so, would such behaviors vary/hold constant when the two groups of scorers cross-examine responses produced from either delivery mode? (3) Are there qualitative differences in the spoken responses that contribute to the rating discrepancy?

The current investigation takes a mixed-methods approach, collecting from a wide array of data sources. ACCESS 2.0’s Speaking tests consist of five cluster tests in grades 1 -12. From each cluster test, 20 students who take retired paper forms of the Speaking test will be selected from schools that have exhibited large score differences between paper and online tests. Participants’ ACCESS 2.0 test taking experiences will be captured and their responses will be rated by both local scorers and central scorers. For online test takers, 20 online test responses will be collected and rated by both central scorers and local scorers. The total sample size will be 200 (paper participants n=100; online responses n=100) and their doubled ratings will total of 400.

For quantitative analysis, ratings will be modeled using Hierarchical Rater Model (HRM: Patz, Junker, Johnson, & Mariano, 2016) to estimate individual rater severity and consistency effects. For the qualitative analysis, a video recording of the proctoring and rating processes of the paper administration will be examined. Additionally, an interview with local administrators will be conducted to illuminate rating behaviors and examine potential scoring bias or consistencies across local scorers. Spoken responses of both modes will be further qualitatively analyzed through three discourse quality measures of fluency, accuracy, and complexity. The intent here is to discover if responses exhibit any consistent linguistic patterns between local and central scorers’ gaps.

Findings from quantitative and qualitative analyses will inform practitioners in the nature and influence (if any) of how scoring procedures and rater effects contribute to mode difference. Results can guide the design of the Speaking assessment and prevent disadvantaging students’ performance based on mode. Outcomes on potential sources of score bias will enhance dual-mode assessments in light of score equivalence and comparability issues.
Re-conceptualizing the Construct of Computer-based Speaking Assessments: Use of Communicative Strategies in Paired Discussion
Lin Zhang, Shanghai Jiao Tong University, China
Yan Jin, Shanghai Jiao Tong University, China

In recent decades, the application of computer technology has become an increasingly important feature of language assessments. For a long time, computer mediation has been regarded as a potential source of construct-irrelevant variance in making inferences about test-takers’ language abilities. Numerous studies have been undertaken to investigate the equivalence between computer-mediated and traditional forms of assessment. With the rapid improvement in language learners’ computer literacy, language testers need to reconsider the place of computer mediation in defining the construct of computer-based language assessments. In this presentation, we report a recent study of the impact of computer mediation on performances in a large-scale, high-stakes speaking test, and discuss the role of computer mediation in conceptualizing the test construct.

Our research focused on the use of communicative strategies in the paired discussion task of the computer-based speaking test, which is designed to assess some aspects of interactional competence (He & Young, 1998; Young, 2003, 2010, 2011). The research question is to what extent test-takers’ strategy performances in paired discussion would be affected by the form of assessment: computer-based or face-to-face. A total of 66 participants at various proficiency levels were invited to sit a face-to-face interview test following the computer-based speaking assessment. Particular attention was paid to establishing equivalence between the two assessments in terms of testing conditions and discussion topics. A coding scheme was developed based on literature review and a pilot study among 12 volunteers. Conversational analysis was conducted to identify two major categories and five sub-categories of communicative strategies employed in computer-mediated and face-to-face discussions. Participants’ performances were also double rated for communication effectiveness. The study revealed a high level of similarities in the quantity and variety of communicative strategies in computer-based and face-to-face interactions. The data also showed that participants were generally capable of turn-taking effectively in the two discussion tasks. Moreover, participants scoring high on communication effectiveness used more interaction strategies than low scorers in both discussion tasks. Despite the marked similarities in overall strategy use, some variations were also observed in the study. Cooperative strategies, particularly the strategy of asking a question, were used more frequently in computer-based discussions than in face-to-face discussions. Analyses of the turn-taking strategy indicated that while overlaps, interruptions and backchannel responses occurred more often in face-to-face dialogues, simultaneous talk and inter-turn pauses were more frequent in computer-mediated conversations.

The research has important implications for construct conceptualization of computer-based speaking assessments. In an era when language learners have developed sufficient computer literacy, computer mediation, which has become an important contextual feature of language activities, should no longer be considered as an interference factor, but as an integral part of the construct being measured in computer-based language assessments. Language testers need to re-conceptualize the construct of computer-mediated assessments by taking into account a local context-bound view of the test construct (Chalhoub-Deville, 2003). Further analyses of contextual features that may influence a test’s construct definition and content specifications need to be conducted under the new research paradigm.

Development and Examination of a Scenario-based, Semi-direct Pragmatic Speaking Test for an Academic Target Language Use Domain
Fred S. Tsutagawa, Teachers College, Columbia University, USA

Despite the fact that pragmatics testing research has long understood the importance of context, especially in regard to setting up a situation in terms of topic, setting, participant roles, power relationships, degrees of social distance between interlocutors, etc. (Blum-Kulka, House, & Kasper, 1989; Chapelle, 1998; Halliday & Hasan, 1989; Hudson,
Detmer, & Brown, 1992, 1995; Timpe-Laughlin et al., 2015), this kind of contextual information has traditionally been conveyed, and arguably controlled, to limit individual test-taker knowledge and attributes by using lengthy written task descriptions that rely on strong reading comprehension skills on the part of the test taker in order to do well on the tasks (Grabowski, 2009). The current pilot study, therefore, expands upon and incorporates aspects of context and background knowledge into the design of a pragmatics speaking assessment by using realistic semi-direct audio and video prompts in a “life-as-a-graduate-student” scenario to determine the extent to which adding real-life contexts such as richer, more detailed interpersonal background information onto the people in the scenario and actually seeing the physical environment where the scenario is taking place, affects test-taker spoken responses. The subjects will interact with six scenario-based video tasks in which they will be asked to make requests of differing degrees of imposition, show appreciation, make complaints, and discuss difficult interpersonal problems, etc., across multiple turns, with all spoken responses audio recorded. For the present pilot study, approximately forty-five total participant samples will be randomly collected, selecting participants from three beginning (n = 15), intermediate (n =15), and advanced (n = 15) proficiency level groups from a Community Language Program (CLP) at a major North American university. In addition, data from approximately 5 doctoral students in an Applied Linguistics graduate program will be collected as a nonequivalent control group. Two native speaker raters will rate the responses for 1) grammatical accuracy, 2) semantic meaningfulness, 3) functional appropriateness, 4) sociolinguistic appropriateness, 5) sociocultural/intercultural appropriateness, and 6) psychological appropriateness (Purpura, 2004, 2017). Descriptive statistics will be gathered for each of the six pragmatic tasks that are rated by the two native speaker raters, and t-test and/or analysis of variance (ANOVA) analyses will be conducted on the three proficiency subgroups to search for possible differences between them. Next, many-facet Rasch measurement (MFRM) will be used to examine the facets of examinee, task scenario type, raters, and rating categories. The results from SPSS and Facets will then be further compared to qualitative post-test survey responses to see if test-taker perceptions matched or digressed from the actual scored pragmatics speaking test results. Sources of rater variance and bias will also be carefully investigated for interactions between the raters and the examinees, task scenario types, and rating categories. As this kind of scenario-based, semi-direct speaking test of pragmatics has not been attempted before, it is hoped the results will provide valuable proof of concept for adoption as a possible large-scale testing solution for speaking tests. Finally, recommendations will be made for designing future pragmatic test tasks of this kind.

Wednesday, July 4
11:40am - 12:10pm
Venue: Case Room 4 (260-009)

The Discourse of the IELTS Speaking Test: Interactional Design and Practice
Paul Seedhouse, Newcastle University, UK

The IELTS Speaking Test (one component of IELTS) is used worldwide to assess whether a candidate has the ability to communicate effectively on programmes in English-speaking Universities. This paper reports on the findings of 3 British Council/ IELTS-funded projects into aspects of the interactional organisation of the IELTS Speaking Test. A brief description is provided of the 3 corpora developed, made up of 257 audio recordings and transcriptions of 11-14 minute speaking tests from around the world. I briefly explain what happens in the 3 parts of the Speaking Test and how performance is assessed. Adopting a Conversation Analysis perspective, findings are presented on the organisation of turn-taking, sequence, repair, and topic development in relation to the institutional goal. Interactional design is compared with interactional practice in terms of what is observable in the data. I then provide examples of application, namely how analysis of data can inform test design and examiner training, looking at topic disjunction and recipient design in particular. I also report on two studies of how features of candidate discourse relate to scores allocated to candidates. I consider the universal problem: to what extent can oral performance in one variety of discourse predict future performance in another variety of discourse? Finally, I discuss the relationship between design and practice: how do the interactional design of the test and the interactional evidence of what actually happens in practice relate to each other (Seedhouse & Nakatsuhara, in press)?

References
Testing Readiness for University Entry: An Investigation of the Relationship between Writing Proficiency Level and the Strategic Behaviours Involved in Integrated Listening- and Reading-to-write Performances

Sally O’Hagan, University of Melbourne, Australia
Ute Knoch, University of Melbourne, Australia
Michelle Czajkowski, University of Melbourne, Australia
Susy Macqueen, Australian National University, Australia
Craig Wattam, University of Melbourne, Australia

The use of skills integration is becoming more popular in the assessment of second and other languages, in particular in assessments designed to assess university preparedness. The TOEFL IBT, for example, added the integrated writing task to “improve the measurement of test takers’ writing abilities, create positive washback on teaching and learning as well as require test takers to write in ways that are more authentic to academic study” (Cumming et al. 2006, p.1). Research on integrated tasks has focussed on the discourse produced and writers’ use of source material. Although a limited number of studies have examined the processes and strategies test takers employ to complete integrated writing tasks (e.g. Spivey, 1990; Plakans, 2008), existing research has mainly focussed on reading-to-write tasks and been restricted to very small groups of learners and has generally failed to link the findings to models of information or text processing. A deeper understanding of these processes and strategies will provide insights into how test takers use source texts to inform their writing and whether they complete tasks in a way that reflects processes employed in target language use situations, e.g. the academic domain.

This study investigates strategic behaviours displayed by test takers completing TOEFL iBT integrated listening-and reading-to-write tasks. Three groups of 15 test takers at each of three proficiency levels took part: two in university preparatory English programs and one in first year undergraduate study. Participants provided verbal protocols while completing the integrated writing task and were subsequently interviewed. Through qualitative analysis of the data, strategic behaviours of test takers were identified and matched to existing models of information or text processing (e.g. Purpura, 2014; Van Dijk & Kintsch, 1983; Plakans, 2008). The findings offer insight into the interaction between integrated task processes and proficiency level and, in this way, build on current understanding of the skill integration processes used by second language learners. The study also showed that existing models are not able to fully capture test taker cognition on these tasks due to their more general nature and the specific nature of the integrated tasks. Implications for both test taker strategy research and model building are discussed.

International Undergraduate Reading Demands: The Case for Disciplinary Literacy Perspectives in Language Test Development

Elizabeth Jean Larson, OISE, University of Toronto, Canada
Megan Vincett, OISE, University of Toronto, Canada
Hyunah Kim, OISE, University of Toronto, Canada
Bruce Russell, OISE, University of Toronto, Canada
Chris Barron, OISE, University of Toronto, Canada

The number of international students studying at universities around the world is rapidly increasing, with seven million students expected by 2022 (ICEF, 2015). English-dominant countries (e.g., USA, UK, Australia, Canada) are among the most popular destinations (UNESCO, 2014), making proficiency in English essential for students to succeed. However, the definition of proficiency and the specificity of language skills have been topics of debate. The field of English for Academic Purposes has predominately focused on a common set of language needs in higher education (Hamp-Lyons, 2011) captured in commercial English tests (e.g., TOEFL, IELTS) required for admission. Yet, there is a lack of consensus regarding whether these tests truly reflect the language skills necessary for university students.
Digging for Gold or Sticking to the Criteria - Teachers Training to be Raters of Texts
Lennart Jøran Jølle, Norwegian University of Science and Technology, Norway
Gustaf Bernhard Uno Skar, Norwegian University of Science and Technology, Norway

This paper reports findings from a project that took place within a writing test program in Norway between 2010 and 2016. The project was called “The National Panel of Raters” (NPR) and the objective was to reliability rate the national sample-based writing test (NSBWT) (Author 2017a).

A recent research project (Author 2017b) found that as a group the NPR produced highly reliable ratings of students writing. However, individual differences between the raters remained, with - for example - statistically significant differences in severity and different use of assessment scale categories. This finding is in line with studies of rater reliability in similar settings (c.f. Author 2017c).

A more puzzling result, however, was a finding that indicated that NPR members expressed different approaches to the NPR job and the materials (e.g. assessment criteria, benchmark texts) and especially the NPR “rules”—the codified ways of applying criteria and using benchmarks. For example, there were teachers who refused to use benchmark texts, because they disagreed with the official marks.

This paper reports results from an explorative follow up study that sought to better understand the different approaches toward the NPR job and rules. The purpose was to identify ways of reasoning in rating situations that would illuminate how individual teachers negotiated the rules in concrete assessment situations. Increased knowledge can contribute to our understanding of the outcomes of this and similar programs.

Sixty-three NPR members were surveyed with 23 items in a questionnaire during a workshop and rating session in the fall of 2016. The items in the questionnaire asked participants to “agree” or “disagree” with certain rater behavior. The items were constructed to be dilemma-like in the sense that deviating from the NPR rules would follow another - but socially acceptable - rationale. A high score on the scale indicated large deviance from the NPR rules.
Four NPR members participated in a follow-up interview in which they explained their answers to the questionnaire. Two high scoring and two low scoring NPR members participated in interviews in which they motivated why they had agreed or disagreed with certain items.

The high scoring members seemed more willing to picture an actual student writer. It seemed as if they saw it as their job to motivate the student to future learning. Typically they said things like “it is about digging for gold” and “the most important thing we do is to find the qualities in the texts and then build on that, rather than searching for shortcomings or errors.” The low scoring raters constantly foregrounded the guidelines and the way they had been schooled as raters. They would typically utter things like “I believe she (i.e. the fictional rater) has to stick to the criteria”.

The results indicate two distinctly different stances toward rating work, with one stance threatening the validity of the scoring process. The results and its implications will be thoroughly presented and discussed in the paper.

Wednesday, July 4
4:00pm – 4:30pm
Venue: Seminar Room B (260-040B)

A Corpus-driven, Curriculum-based Approach to Assessing Chinese Proficiency Development in US Universities: Textbooks and Core Content Domains

Xun Yan, University of Illinois at Urbana-Champaign, USA
Yuyun Lei, University of Illinois at Urbana-Champaign, USA
Chilin Shih, University of Illinois at Urbana-Champaign, USA

This study explores a corpus-driven, curriculum-based approach to assessing proficiency development of Mandarin Chinese, a less commonly taught language (LCTL) in US universities (Walker, 1989). One common challenge faced by LCTL programs is the development of reliable and valid assessments to measure students’ proficiency for placement, achievement and diagnostic purposes (Nier, Donovan, & Malone, 2009; Winke, 2006). Because there are a small number of test-takers and limited testing resources, LCTL instructors need to look for reliable content materials to create test items with appropriate difficulty levels and content coverage for the target learners. To address this challenge, this study employs a corpus-driven approach to examine language use in commercial Chinese textbooks, the primary source of language exposure and use for LCTL learners, as an effort to identify core content domains for Chinese assessment in US universities.

We compiled and annotated a corpus of 22 Chinese textbooks used widely in US universities. These textbooks were selected based on a recent large-scale survey conducted by the Chinese Language Teachers Association, USA (Li, Wen, & Xie, 2014), covering three general proficiency/course levels: beginning, intermediate and advanced. The corpus consisted of 293 unit texts, amounting to a total of 807,433 characters as well as a list of 18,559 key vocabulary words and 2,727 key grammar items. Using corpus-analytic techniques, we found similar distribution patterns in the complexity and diversity of lexico-grammar in the textbooks across proficiency/course levels. Based on these distribution patterns, we identified a core set of linguistic forms and functions targeted by Chinese curriculums in US universities.

To test the appropriateness of the core content domains for local institutional contexts, we developed a curriculum-based elicited imitation (EI) test and piloted it on 15 learners across three proficiency/course levels in the Chinese language program at a large US university. EI was chosen because it is a well-attested quick and effective method to measure language proficiency (Vinthner, 2002; Yan, Maeda, Lv & Ginther, 2016), and there are Chinese EI tests developed in other approaches that can be used as a baseline comparison. We created three sets of stimulus sentences (k=30), each targeting the key linguistic forms and functions at a proficiency/course level, and then compared its psychometric qualities with those of a non-curriculum-based EI test developed by Wu and Ortega (2013) (parallel translated versions of the EI tests in Ortega, Iwashita, Norris, & Rabie, 2002).

Results of item analysis suggest the potential of the curriculum-based test for placement, achievement, and diagnostic purposes. The curriculum-based test demonstrated better psychometric qualities. Scores and responses on the test could better predict individual test-takers’ proficiency/course level and indicate their acquisition of key forms and functions in the curriculum. Taken together, findings of this study support the usefulness of the corpus-driven, curriculum-based approach to assessing proficiency development of Chinese as well as other LCTLs in local institutional contexts. Suggestions on how to use this approach to guide content selection and item development will also be discussed.
Assessing Content for Enhancing Construct Representation in Source-based Writing Performance Assessment

Yasuyo Sawaki, Waseda University, Japan

The ability to understand source-text content accurately and represent it appropriately in writing is a critical component of academic literacy. Except for a small number of previous studies (e.g., Li, 2014; Sawaki, Quinlan, & Lee, 2013; Yang & Plakans, 2012; Yu, 2007), however, L2 assessment research has focused relatively little on source-text content representation compared to other aspects of integrated writing task performance. This is possibly due to the frequent conceptualization of scores on such task types as measures of writing ability. Thus, an important direction for future research is to explicate the role of source-text comprehension and its representation in task performance and devise criteria for assessing content that promote the balanced development of academic reading and writing skills that underlie source-based writing performance.

This study examines the functioning of a rating scale designed to assess the content of learner-produced summaries in university EFL academic writing courses in Japan. The integration rating scale is designed to assess the degree to which a learner-produced summary succinctly and yet accurately represents main points of the source text and their interrelationships based on Kintsch and van Dijk’s (1978) discourse processing model. The functioning of this rating scale is examined from three perspectives: (1) score dependability, (2) relationships with other measures of L2 reading, writing, and summarization performance, and (3) relationships to the use of Kintch and van Dijk’s macrorules for text comprehension and summarization (e.g., Hare & Borchartd, 1984; Cordero-Ponce, 2000) represented in learner-produced summaries. Two cohorts of English majors enrolled in EAP courses at a private university in Japan in 2016 and 2017 wrote 60-80 word summaries of two expository texts randomly assigned to each individual out of four. Each summary was scored by two independent raters. Phase 1 of the study focused on the analysis of score dependability on this rating scale. A multivariate generalizability analysis of summaries written by the 2016 cohort (N=162) for the fully-crossed person-by-rater (ρ* x r*) design suggested sufficiently high dependability of the ratings for low to medium-stakes use (Φ=.71~.78 across texts) and generally low universe-score correlations with other analytic measures of summarization performance, suggesting the relative distinctness of the construct this measure taps into. Data obtained from the 2017 cohort (N=135) are being analyzed in Phase 2 for a partial replication of the Phase 1 generalizability theory analysis and for exploring the degree to which integration is predicted by measures of EAP ability (TOEFL ITP® total and section scores and Criterion® essay scores) and the frequency of macrorule use in learner-produced summaries identified by human coders (20% double coding) by conducting hierarchical regression analyses. In this session, the presenter will describe the theoretical background of the integration rating scale, summarize Phase 1 and Phase 2 study results, and discuss how the incorporation of a content measure like this rating scale can enhance the construct representation in integrated writing assessment as well as the design of performance feedback provided to learners.

Measuring Cohesion: Fair Partial-credit Scoring in Sentence-sequencing Reading Tasks

Richard David Spiby, British Council, UK
Karen Dunn, British Council, UK

This paper addresses issues arising in the fair allocation of scores for sentence-sequencing items in reading tests, in which test-takers are required to re-order a set of sentences to display knowledge of intersentential cohesion. Such tests have recently undergone a resurgence of interest (O’Sullivan, 2015; Pearson, 2017). A common means of scoring such items is to credit only a perfectly ordered set of sentences. However, this approach loses a large amount of information about test-takers who might misplace a couple of sentences. Different methods of allocating partial credit have been suggested, including ‘exact matching’, crediting pairs and triplets of sentences, correction steps and distance from correct position. However, no consensus has been reached in the literature as to the optimal partial-credit scoring solution for such tasks.
The study reported here investigates methods of scoring sentence-sequencing items in terms of their reflection of candidates’ understanding of intersentential cohesion, and evaluation of these scoring methods with live response data. The data analysed were taken from the reading component of a large international test. A sentence-sequencing item comprising six sentences appears alongside three other subtests. A subset of data from two test forms were used in the analysis, with approximately 1,400 test-takers. All test-takers also completed a separate test of grammar and vocabulary.

The response data were re-scored under four distinct partial-credit scoring methods. However, none of the scoring methods were found to maximize the potential of the interim score points along the scale. Conflation of interim categories yielded a much clearer picture. However, at a candidate level, the test-takers scoring in interim categories varied across methods. This was most distinct in the comparison between adjacent scoring methods (i.e. pairwise) and location scores (i.e. exact matching), for which there was movement in both directions of approximately 10% of test-takers between category 0 and interim categories. This movement between categories highlights a fairness issue in determining which scoring approach to apply to tasks of this nature. The scoring method that best reflects the construct being tested will justly credit the candidates that partially fulfil task requirements.

In order to address this issue, the scores derived from the different approaches were examined with respect to other items within the test, and also grammar and vocabulary items, which are believed to be a good predictor of reading ability (Urquhart and Weir, 1998). Scores employing the pairwise method displayed a stronger relationship with the grammar and vocabulary test results. Relationships between the exact-matched and pairwise-scored sequencing task were found to vary in the strength of correlation with other items in the reading test across difficulty levels. Furthermore, single-factor CFA of the polychoric item response matrix revealed loadings more evenly dispersed for the dataset containing exact matching scores.

This paper discusses the issues that this analysis highlights with respect to scoring sentence-sequencing tasks, and the possibility that whilst the scoring of adjacent sentences may be felt to best reflect the construct informing the test design, it is less clear whether such an approach has optimal measurement properties.
analysed deviations statistically, and discussed underlying reasons with the teachers. This facilitated refining the checklists, and it helped teachers voice their concerns regarding the transition from intuitive grading on a 16-point scale to rating student scripts against proficiency-based descriptors in a checklist.

The insights we gained have implications beyond our local context. Many language course providers face the challenge to move from internal teacher-defined criteria to comparable standardised criteria that are aligned to an internationally recognised framework. In such contexts, our approach can illustrate how such a transition can be managed, and how the expertise of teachers, course organisers and researchers can be drawn upon in a synergetic endeavour. Moreover, our approach to developing a proficiency-based checklist adaptable to fit individual teachers’ needs and classroom practices can serve as example for contexts where achievement testing is to be aligned to proficiency frameworks.

Wednesday, July 4
4:35pm – 5:05pm
Venue: Seminar Room B (260-040B)

A Corpus-driven Receptive Test of Collocational Knowledge
Ivy Win-wen Chen, 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 drawbacks, most notably issues with item selection, such as the inclusion of only one collocation type (usually restricted verb-noun). The current research aimed to to create a corpus-driven and reliable receptive test of the knowledge of the word parts of high-frequency collocations in English by considering how the different properties of collocations (e.g. frequency) affect 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. Two large corpora (COCA and BNC) were used to produce a list of the most frequent 2,000 collocations in English. Test items were selected from five common types of word combinations: compound words and four collocation types (verb-noun, adjective-noun, noun-preposition, and other phrase). Participants were non-English dominant undergraduates studying at an Australian university (N=132), with the majority from a Chinese background (dominant in Mandarin, N=79; Cantonese, N=6). 47 English dominant students served as a comparison group.

This paper will focus on two main aspects of the study. The first part concentrates on 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: length in number of syllables, collocation type (grammatical versus lexical), frequency (by genre), degree of coherence, and semantic transparency. While a few of these could significantly predict item difficulty, it was collocation type that did this best. As expected, both lexical collocations (verb-noun, adjective-noun) were significantly easier than the two grammatical ones (noun-preposition, other phrase), mostly due to differences in semantic transparency.

The construct is the focus of the second part, including the checking of the internal structure of the test to see whether the assumption of a single construct was supported when a larger range of collocations were incorporated. The data suggest that the construct may not be limited to collocational knowledge but to a larger construct of vocabulary knowledge. This is because internal consistency was high with the compound word subtest (α=.98) and without (α=.97), and all five subtests (i.e. including compound words) were significantly correlated with each other (apart from adjective-noun and other phrase). The internal consistency of all subtests (15 to 18 collocations each) was reasonably high (α between .72 to .91). The test also significantly correlated with criterion measures of general English proficiency (c-test) and knowledge of routine formulae. In sum, this paper provides strong support for the inclusion of multiple collocation types ranging in semantic transparency and frequency. This careful item selection process is an important and often neglected aspect of the development and validation of collocation tests.
An Investigation of How Test Takers Integrate Content from Reading and Listening Stimulus Materials into Performances on a TOEFL iBT Speaking Task Designed to Measure Readiness for Academic Entry

Kellie Frost, University of Melbourne, Australia
Joshua Clothier, University of Melbourne, Australia
Gillian Wigglesworth, University of Melbourne, Australia

Integrated speaking tasks, which require test takers to incorporate content from reading and/or listening texts into their oral performances, are thought to provide a more authentic and comprehensive measure of the construct of speaking ability needed for students to manage the transition into higher education contexts. However, the ways in which test takers integrate information from stimulus materials into speaking performances, particularly the cognitive processes and strategic behaviours involved in understanding, selecting, and reproducing content from these materials, is not well understood. Moreover, while language learning strategies have been widely researched in relation to reading tasks (for example, Grabe, 2009; Phakiti, 2003) and listening tasks (for example, Richards, 1983; Vandergrift, 1997), research into strategy use and its impact on integrated speaking task performance remains extremely limited.

This paper reports on a study aimed at extending beyond the taxonomies of strategies that have emerged from previous research. It offers a way to conceptualise the complex and varied relationship between language skills, strategy use, task characteristics and the quality of spoken discourse produced by test takers, particularly in terms of the accuracy and organisation of content drawn from stimulus materials. To this end, we investigated how content from stimulus materials in a TOEFL iBT integrated reading and listening-to-speaking task was processed and integrated into oral performances by test takers across three levels of proficiency: low, middle and high (based on previous TOEFL overall scores). The dataset consisted of transcripts of 45 speaking task performances and 45 immediate-retrospective-stimulated verbal reports from the same test takers. Test takers’ performance discourse was examined using a range of content analytic measures, developed by Frost, Elder and Wigglesworth (2012), to determine the accuracy with which test takers across proficiency levels reproduced idea units from the source texts, the extent to which they summarised and organised source text information, and the coverage of main ideas they achieved. The content analytic approach was combined with a verbal reporting methodology to examine the complex and interactive relationships between the skills and strategies involved in generating oral performances, particularly those engaged by test takers to reproduce and/or transform content from the stimulus materials, and the content-related dimensions of task performances. Our study provides important theoretical insights into the nature of the reading- and listening-to-speak construct and, based on findings of between-group differences, a model of these relationships is proposed.

Designing a Diagnostic Assessment of EAP Reading: A Needs-driven Approach

Xiaohua Liu, University of Auckland, New Zealand

Recent years have witnessed a surge of interest in exploring diagnostic language assessments from various perspectives. However, few of those projects have incorporated assessment users’ needs; instead, they have generally relied on a retrospective approach by retrofitting assessments developed for other purposes (e.g., Jang, 2009; Lee & Sawaki, 2009; Li, 2011; Kim, 2014; Chen & Chen 2015). It is argued that neglecting users’ needs in assessment development may undermine the meaningfulness, generalisability, relevance and sufficiency of the diagnostic information generated due to potential construct irrelevance and construct underrepresentation (Bachman & Palmer, 2010; Messick, 1987). In view of this, the present study applied a needs-driven approach to re-designing the reading section of the Diagnostic English Language Needs Assessment (DELNA), which currently yields only an overall band score.
In order to achieve this, a theoretical framework for designing diagnostic assessments was developed based on the Evidence-Centred Design (Mislevy, Almond, & Lucas, 2003) and the argument-based approach to assessment (Kane, 2004). This framework highlights that, for a diagnostic assessment to be defensible in terms of extrapolation and utilisation, the construct model of the assessment should reflect assessment users’ needs. Therefore, a needs analysis in a sequential design (Creswell & Clark, 2011) was conducted at the university where the DELNA is operated. Twenty-nine semi-structured interviews with undergraduates and language teachers identified a wide range of reading and reading-related skills that were needed in academic situations. A subsequent questionnaire survey with more than 200 undergraduates based on the interview findings further revealed that these skills consisted of five subdomains of academic reading of differing importance. Taking into account both validity considerations (i.e., utilisation, extrapolation, explanation, generalisation, and evaluation) and practical constraints (i.e., assessment administration, scoring, and feedback), the application of the theoretical framework led to an 11-skill construct model and a 3-module assessment design based on the needs analysis.

Although conducted specifically within the DELNA context, this study represents one of the few assessment projects which have attempted to develop a diagnostic reading assessment based on assessment users’ needs. Thus it has implications for developing diagnostic assessments in other contexts, especially those in which assessments are used at the outset of academic programmes to produce linguistic profiles for guiding further language enhancement. The presentation will elaborate on these specific implications. It will also discuss the issues involved in converting findings of needs analysis into a conceptual framework for assessment, such as the challenge of balancing validity considerations and practical constraints in the choice of skills/constructs as well as operational features.

Wednesday, July 4
5:10pm – 5:40pm
Venue: Case Room 3 (260-055)

Linguistic Dimensions of TOEFL iBT Essays Compared with Successful Student Disciplinary Writing
Eric Frigina, Georgia State University, USA
Sara Cushing, Georgia State University, USA
Scott Grapin New York University, USA
Margaret Malone, Georgetown University, USA

Establishing that performance on an assessment is comparable to performance in a target language use (TLU) domain (Bachman & Palmer, 2010) is critical to supporting the claim that an assessment is useful for decision-making. For a high-stakes test like TOEFL iBT, for example, it is important to establish that performance on the test will predict performance in the university setting. Weigle and Friginal (2015) conducted an investigation to begin to address this validity concern by examining the linguistic features of essays produced in response to the TOEFL Independent task and a corpus of successful college writing across several disciplines. They found that test essays differed in significant ways from disciplinary writing, particularly in the natural and health sciences. Their study, however, did not include the TOEFL Integrated task, the other task type in the TOEFL iBT Writing Section. Whereas the Independent task asks students to agree or disagree with a statement and to provide specific reasons and examples to support their answer, the Integrated task asks students to write an essay that summarizes a lecture and compares the information on the lecture to the information on a reading passage. Previous studies that have compared the linguistic features of the Independent and Integrated tasks have found that they do, in fact, elicit different features (Biber & Gray, 2013; Guo, Crossley & McNamara, 2013). These studies also suggest that the Integrated task, in comparison with the Independent task, may elicit writing that is more similar to disciplinary writing. The purpose of this study is to investigate whether this hypothesis is supported.

We compared the linguistic features of TOEFL essays written by 103 international, non-native undergraduate students from 8 universities in the U.S. with the same corpus of successful student writing used in Weigle and Friginal (2015). Using Biber’s (1988) multidimensional analysis framework, the TOEFL essays (one Independent and one Integrated) were tagged and analyzed along four dimensions of successful student writing, and then compared to the corpus using multiple analysis of variance (MANOVA) following the same approach as Weigle and Friginal (2015).

We found that there were differences between the linguistic features of the TOEFL essays and the corpus of disciplinary writing. In general, TOEFL essays had higher dimension scores than disciplinary essays, and were closer, in terms of linguistic features, to humanities and social sciences essays than those in the natural and health sciences. We also found differences in the linguistic features of the Independent and Integrated essays on 2 of the 4 dimensions: scores on the Independent essay were higher than scores on the Integrated essay. On these 2 dimensions, the
linguistic features of the Integrated essays were closer to those of the disciplinary essays, providing evidence to support the hypothesis that the Integrated task, in comparison with the Independent task, may elicit writing that is more similar to disciplinary writing. Our results differ in some respects from those of Weigle and Friginal (2015); we discuss these differences and implications for the TOEFL validity argument.

Wednesday, July 4
5:10pm – 5:40pm
Venue: Seminar Room B (260-040B)

Developing a Validity Argument for a Diagnostic Computer-adaptive Vocabulary Knowledge Profiler
Benjamin Kremmel, University of Innsbruck, Austria

Vocabulary knowledge is key to the successful use of any language skill (Nation & Webb, 2011) and vocabulary assessments can play a facilitating role in this learning process, which is why there is now an abundance of assessment tools to measure lexical knowledge. However, vocabulary assessments do not seem to have moved with the times and taken into account the changing landscapes in language learning. Few vocabulary tests have undergone sophisticated validation, and even fewer have attempted a transition to newer technologies. Webb and Sasao (2013) claim that there has not been much innovation in the development of new vocabulary tests. Instead, conventionalized traditions are being reiterated without questioning them. This talk will present a PhD project that has set out to address this gap of an innovative measure of vocabulary knowledge by developing a new diagnostic computer-adaptive measure of form-meaning link knowledge: The Vocabulary Knowledge Profiler.

The talk will introduce a test development project that started from scratch by questioning underlying assumptions and trying to make design decisions based not only on theoretical considerations but empirical evidence. The paper will outline briefly the design of the instrument and will illustrate backing for the meaningfulness of score interpretations from four studies.

First, a study on 99 EFL learners, which investigated the interpretability of four different test formats for discrete vocabulary tests, found that all formats show considerable measurement error but the MC format may be the most useful because of its systematicity in overestimating scores. Data from the same study also showed that the lemma is a more interpretable counting unit in tests than traditional word families. Second, a study using corpus analyses will argue for a new approach to frequency banding that takes into account the relative importance of frequency bands in terms of the coverage they provide for more meaningful interpretation. Third, another study related the scores from the Vocabulary Knowledge Profiler to those of a reading comprehension measure and found that the test manages to distinguish between readers at different CEFR levels to some extent, but that further validation research is required prior to the test launch. Fourth, the claim of consistency of assessment records is backed by a study that compared two different computer-adaptive algorithms for implementation in the test design. 157 EFL learners took two test versions twice in a row to explore which yielded more reliable scores. 34 learners also took an extended test to see which version would generate more representative score profiles.

All of this evidence will be presented as a validation argument based on Bachman & Palmer’s (2010) Assessment Use Argument for this new measure. Pulling together this initial validation evidence from different studies, a tentative argument will be built that the Vocabulary Knowledge Profiler and the diagnostic decisions that are made based on its results and use are beneficial to EFL learners and EFL teachers for classroom learning and teaching.
Thursday, 5 July 2018
1:50pm - 2:20pm
Venue: Case Room 2 (260-057)

Cloze Test Item Validation Study of L2 and L1 Performance Using Mechanical Turk
Jonathan William Trace, Keio University, Japan

The use of cloze procedures has been a staple of second language assessment for over forty years, and in that time concerns have continuously been raised about the constructs being measured by them, and in particular the degree to which cloze items can be said to draw on comprehension that is limited to bottom-up or local levels of information only. Though much of the research points to cloze tests as measuring beyond L2 reading ability alone, many of those same studies argue that cloze items are limited in their scope to drawing on local syntactic and semantic information.

Trace, Brown, Janssen, and Kozhevnikova (2016) have recently argued against this proposition, finding evidence for cloze items being able to also measure at the intersentential or more global levels of passage comprehension, following similar arguments by Jonz (1990) and Sasaki (2000). No connection, however, has been drawn between these findings and what it means for our understanding of the larger construct being measured by cloze items. Given that these tests are still widely used by classroom teachers and in high-stakes placement contexts such as Japan, a lack of a full understanding of what interpretations can be drawn from these assessments remains a critical issue for test use and development, particularly in local contexts by teachers and test designers where experience in assessment validation practices may be limited.

To that end, the current study gathered extensive information about fifteen 30-item pseudo-randomly designed cloze tests administered to both L1 and L2 test-takers in order to pinpoint how cloze items function. Native user of English data were collected using Amazon’s Mechanical Turk crowd-sourcing data platform to gather test data from L1 proficient users of English in the United States (n = 675), which were then compared to performances on those same tests by Japan and Russia EFL university students (n = 2274). Using a 10-item anchor passage that was also administered to all participants, many-faceted Rasch measurement was run to place all fifteen tests onto a single interval scale to identify item difficulty for 450 individual cloze items. Item difficulty was then further compared to L1 data collected via Mechanical Turk (n = 2475) from an experimental study that pinpointed the minimal level of context required to answer each of the items. Through an examination of Rasch item difficulty, item discrimination, and the level of context targeted by each item, several conclusions were drawn about cloze item function for L2 learners.

Importantly, there was clear evidence that cloze items can and do draw on both local information as well as intersentential or global information in order to be correctly answered by examinees. Furthermore, the results point to cloze items functioning very similarly for both L1 and L2 examinees, which taken with other evidence suggests that they appear to be construct valid as specifically measures of reading ability. These results, therefore, have clear implications for the continued use of cloze tests for L2 assessment going forward.

Thursday, 5 July 2018
1:50pm - 2:20pm
Venue: Case Room 3 (260-055)

Pressure for High-stakes Tests to Contribute towards Positive Washback: The Case of a Locally Developed Test
Carol Dabarera, Nanyang Technical University, Singapore

Although assessments have been categorized frequently according to the broad purposes they serve, they could also be viewed as vehicles for providing information to various stakeholders for varied purposes (Murchan & Shiel, 2017). This information is crucial to enable evidence-based decisions and practices to improve teaching and learning. There has been an increased awareness on the necessity to understand and manage the impact of washback brought about by large-scale high-stakes language assessments (Alderson, Brunfaut & Harding, 2017). Accordingly, there were attempts over the years to examine various large-scale high-stakes language assessments for their potential to contribute towards positive washback. However, it is noted that there is a misalignment between the aims of these tests and the feedback that the researchers sought to provide.
The test examined in this study is a locally developed high-stakes language test. In addition to providing a benchmark for institutions and schools to evaluate potential and current employees and students, it aims to see to the needs of the locals for self-assessment, and to promote the learning of English and lifelong learning. This unique combination of summative and formative aims places this test in a good position to explore the feasibility of large-scale assessments to fulfil additional purposes (Chudowsky & Pellegrino, 2003; Doe, 2015; Klinger, 2016) such as making known the nature of the test-takers’ achievements (National Research Council, 2001) which will contribute to the key function of assessment to facilitate and benefit the growth of the people and society (Murchan & Shiel, 2017).

This study utilized a sequential mixed method design to explore the skills assessed by a locally-developed language proficiency test, the GEPT Advanced Reading. The potential of the test to provide feedback is evaluated according to DiBello, Roussos, and Stout’s (2007) framework of diagnostic assessment implementation. Since there has been much criticism on the use of the same psychometric model in the refinement and validation stages of the framework, this study attempted a variation by using different statistical procedures for these two stages. This study identified five skills evaluated by the test. In addition, unlike the findings of previous studies, this study found that it is feasible to utilize the entire range of subskill mastery classification. Other findings include evidence of psychological multidimensionality and support for non-compensatory relationship between the identified skills. The latter is in contrast to studies that assumed non-compensatory relationships without much justification. The findings of the study are situated within the general theoretical framework for diagnostic language assessment (Lee, 2015) to explain its contribution to enhancing learning outcomes.

The test examined currently provides a holistic score for the test and is hopeful that the findings of this study will contribute towards an enhanced score-reporting practice that will see to the formative aims of self-assessment and lifelong learning that the test aims to achieve, and contribute towards positive washback. On a wider scale, it is hopeful that the study furthers the notion that at the heart of assessment is, as succinctly articulated by Hayward (2015), Assessment is learning.

Thursday, 5 July 2018
1:50pm - 2:20pm
Venue: Case Room 4 (260-009)

An Investigation of Strategy Use on Expeditious and Careful Reading Tasks
Presenter: Limei Zhang, Nanyang Technical University, Singapore

Messick (1989) pointed out that test takers “differ consistently in their strategies and styles of task performance” (p. 54). Bachman (1990) argued that a critical limitation to the quantitative approaches to construct validation is that only test scores are investigated and no means are provided to examine test-taking processes. On the other hand, researchers have long been interested in the differences between expeditious and careful reading (e.g., Weir, Hawkey, Green, & Devi, 2012; Urquhart & Weir, 1998; Aryadoust & Zhang, 2016). According to Weir, Hawkey, Green, Ünaldi, and Devi (2012), expeditious and careful reading are two types of reading operationalized through a series of skills at both local and global levels. Urquhart and Weir claim that relative to careful reading, expeditious reading skills have been overlooked in reading instruction and research, though it is believed they are extremely important in academic contexts and daily life (Rayner & Pallatsek, 1989; Weir, 1993).

This study was designed to examine the test-taking process through the investigation of test takers’ strategy use on expeditious and careful reading tasks. It addresses the following research questions: 1) What strategies do test takers use in expeditious and careful reading tasks? 2) Are there any differences in test takers’ strategy use when they respond to expeditious and careful reading tasks? 3) What are the implications of the results for test validation and reading instruction? Twenty-six EFL college students were invited to sit for a reading comprehension test of expeditious and careful reading tasks while verbalizing their understanding of the test in the language they were most comfortable with.

With Nvivo software program, the think-aloud data were analyzed following the framework informed by Cohen and Upton’s (2006), Gernsbacher (1990), Pressley and Afflerbach (1995), Zhang, Goh and Kunnan (2014). Results indicate that different strategies are used in expeditious and careful reading tasks, indicating that the cognitive processes of the two types of reading are different. Implications for test validation, test preparation and classroom instruction of reading comprehension are provided. This study is expected to make contributions to the knowledge and research of reading assessment and classroom instruction of reading comprehension.
How Young Learners’ Interest Can Be Used to Support a Learner-oriented Assessment Approach

Megan Vincett, University of Toronto, Canada
Clarissa Lau, University of Toronto, Canada

Within a learner-oriented assessment approach, tasks are developed as integrated tools of assessment and learning, with the intention of generating authentic reading experiences to promote learning through assessment (Carless, 2007; Saville, & Jones, 2016). As the issue of student interest is one of the foremost challenges faced by young learners (O’Flahavan et al., 1992), so to must interest be a key consideration within authentic assessment. Interest is found to influence recognition, discourse processing, cognitive attention and expository text recall (Hidi, 1990) as well as the ability to engage in deeper text processing (Schiefele & Krapp, 1996). As such, research exploring the extent to which interest influences young learners’ assessment performance is overdue. One key means of promoting interest among readers is through the provision of choice, known to be associated with improvements in self-regulation (Randi & Corno, 2000), affect (Ryan & Deci, 2000), motivation, effort, and learning (Flowerday & Schraw, 2000).

The current study explored student interest through provision of choice during mediated diagnostic assessment intervention sessions, and its role on students’ responses to the intervention. 14 Grade 5 and 6 readers received mediated assessment interventions. Each student received diagnostic reading profiles with information detailing skill mastery, perceived reading ability, and goal orientation. In eight one-to-one intervention sessions, students were guided to work on specific reading goals using a variety of texts and think-alouds. Nine passages were selected based on the analysis of text density using TextEvaluator (Sheehan, Flor, & Napolitano, 2013). Across sessions, students selected four passages and used think-alouds to explain their reasoning behind choice of goals and text selections. An inductive thematic analysis of transcribed verbal data was conducted to better understand the role that student interest played during assessment tasks.

Student data revealed a seven-category continuum of interest: interest with expansion, preference of particular literary device, prediction of text content related to interest, interest without expansion, perceived difficulty based on surface structure, lack of clear reasoning, and lack of any reasoning. The continuum was used to further examine how student interest mediated changes in learning by employing inferential statistics to identify impact of interest across sessions. The study results indicated that provision of text choice did facilitate student interest, improving performance, although this was dependent upon where students were located on the interest continuum. That is, students lower on the continuum often made ill-informed choices guided by interest that didn’t facilitate their learning, in fact at times leading to increased reading difficulties. For example, if choices were made based on inaccurate assumptions about text content using visuals, titles, or words, interest was found to decrease, resulting in decreased text comprehension.

The study results offer important implications for designing learner-oriented assessment tasks that incorporate interest within the context of dynamic assessment. Findings suggest that choice alone does not facilitate student learning, but that students would benefit from being taught specific strategies as well as being provided indirect guidance (e.g. tailored text selections and goal options) to effectively cultivate interest in authentic assessment tasks.

Classroom Assessment for Engaging Chinese First-year University Students: A Didactic-experiential Continuum

Jiming Zhou, Fudan University, China

Improving students’ current academic achievement and fostering their capacity for life-long learning are core issues on the worldwide university agenda. Classroom assessment is arguably one of the most powerful strategies for achieving these aspirations. While large-scale surveys in different higher education settings offer snapshots of the challenges confronting students in their first university year, the complex and dynamic nature of students’
acculturating process, and the potential role of classroom assessment during that process, warrant further exploration with qualitative data. This study explores how to use classroom assessment to engage Chinese students in an English for Academic Purposes course in their first university year. It addresses the following two research questions: 1) What are students’ perceptions of and approaches to learning at different time points in their first university year? 2) How do teachers’ assessment practices affect students’ learning perceptions and engagement? This study’s datasets include focus group interviews with twenty students at two different time points over one academic year, self-reflective essays written by seven students, and 36 sessions of audio-recorded English classes taught by two English tutors. Content analysis procedures (Merriam, 2009) and constant comparative analysis (Strauss & Corbin, 1990) were applied to transcripts of the audio-recorded classroom assessment episodes, interview transcripts, and students’ reflective essays. The findings show that Chinese students, at the outset of their university learning, demonstrated an obsession with knowledge accumulation, and relatively low-level engagement in activities designed to develop their higher-order learning skills (such as synthesizing and evaluative skills). Teachers’ divergent approaches to conducting assessment were identified, based on which a didactic-experiential continuum of classroom assessment implementation was proposed. A didactic approach involves a teacher talking about a practice from an external perspective, whereas an experiential approach involves a teacher generating participatory opportunities for students, and scaffolding them during the participatory process. It was through iterative participation that students in this study gradually realized and accepted that their teachers ‘played cards in a different way’; they were given opportunities to reflect on whether and how they would like to pursue externally-set learning goals, and subsequently became more active, engaged, and competent in their university learning. In brief, the driving power of classroom assessment in engaging students lies in the experiential access it grants them. This study illuminates the interplay between classroom assessment and first-year university students’ engagement. A comparison of the two teachers’ classroom assessment enactment carries pedagogical implications. First, clearly-defined learning goals are essential for students, especially in the first year, when they are adapting themselves to university study. Second, a variety of assessment tasks need to be implemented, to provide students with sufficient opportunities to display their abilities. Considering the increasing number of international student intakes (including Chinese students) in English-medium universities, the study has contextual relevance to Western education practitioners, in terms of using classroom assessment to enhance students’ learning engagement.

Thursday, 5 July 2018
3:50pm - 4:20pm
Venue: Case Room 3 (260-055)

Can the Strength of a Speaker’s Accent Be Reliably Measured?
Gary John Ockey, Iowa State University, USA

To better ensure that a rich construct of listening is assessed, concerns about the need for assessing multidialectal listening skills are becoming increasingly prevalent. However, the inclusion of speakers who have diverse speech varieties on listening assessments may threaten test fairness because test takers’ listening could be affected differently depending on their experience with certain speech varieties; familiarity with a given speech variety has been shown to be an advantage (e.g., Gass & Varonis, 1984; Ockey & French, 2016), and it would not be feasible to include every speech variety in a given community on an assessment. One possible approach to this challenge is to select speakers with somewhat different speech varieties but not so different that a familiarity effect would result. To realize this approach, it would be necessary to have a reliable instrument that could be used to measure a speaker’s accent.

This study aimed to investigate the reliability and possible sources of score variance in a strength of accent measurement instrument designed to aid in selecting speakers for second language (L2) listening assessments. Based on previous research, accent was defined, and a strength of accent scale was developed. After numerous tryouts of the scale and resulting revisions, a study was conducted to investigate its reliability and possible sources of score variance. Twenty English speakers, two of whom were believed to have accents representative of the local speech variety (Standard American English) and 18 of whom were believed to have accents that differed from the local speech variety, were recorded giving a short lecture. Two 20-second audio clips for each of the speakers were then taken from the short lectures. These 40 audio clips were played to 69 L1 and 31 L2 listeners, who were users of the local speech variety. The L2 listeners were all advanced speakers of English, who were studying in subject matter courses in three different English-medium universities. The L1 listeners studied at the same three universities as the L2 listeners. The listeners used the strength of accent scale to judge each of the 40 audio clips. A Many-Facet Rasch
Measurement analysis with two facets: 1) speakers with different accent strengths and 2) listeners with different proficiencies, was used to analyze the data. Results indicated that the listeners’ use of the Strength of Accent scale resulted in ratings which clearly distinguished speakers of the local variety from those of other varieties, who in turn were reliably distinguished from each other. The results also indicated that L1 and L2 users of the local speech variety differed in their judgements of the speakers’ accent strengths, with L2 users being slightly harsher and less consistent. Implications of the study are that the development of a reliable strength of accent scale that can be used to aide in selecting appropriate speakers for listening assessments is achievable. Moreover, the study indicates that less proficient English listeners may be more impacted by different speech varieties than their more proficient counterparts.

Thursday, 5 July 2018
3:50pm - 4:20pm
Venue: Case Room 4 (260-009)

Creating a Validity Argument for an Existing Testing Program for University Admissions and Professional Certification
Angel Arias, University of Ottawa, Canada
Beverly Baker, University of Ottawa, Canada
Carol Chapelle, Iowa State University, USA

Testing bodies need to manage a diverse set of test development and validation activities in addition to communicating with various stakeholders about the tests and their intended score uses. An argument-based validation framework is intended to organize these activities in a systematic way, but this approach can seem out of reach to many existing testing programs, and has not been widely adopted (Newton & Shaw, 2014). One reason may be that creating a validity argument is presented as a resource-heavy two-step process consisting of 1) development of an interpretation/use argument and then 2) seeking evidence to support each of the claims in the argument (e.g., Chapelle, Enright, & Jamieson, 2008; Kane, 1992, 2006, 2013). Researchers wanting to develop validity arguments in real testing programs face obstacles such as a lack of resources and institutional understanding of the need to conduct ongoing validation research.

This paper provides an example of how one testing organisation transformed its existing analysis and research practices into a program of research and communication informed by a validity argument framework. More specifically, we focus on the reading and listening sections of one of our high stakes tests used for mobility purposes; i.e., for university admissions as well as for licensing requirements for professionals educated abroad. Past activities for test validation include updating of construct definitions for the listening and reading sections of the test with support from existing literature; collecting evidence for the internal structure of these constructs in view of revised construct definitions; developing protocols for standardized administration and grading procedures; examining test taking processes through the use of think-aloud protocols, and consulting with users of test scores. In this paper, we describe how these ongoing test validation activities were re-organized and re-interpreted in view of the backing they can provide for claims in the validity argument for test interpretation and use.

The paper will show the resulting interpretation/use argument with the existing backing, in addition to our future research agenda. We also include a discussion of how to negotiate obstacles related to resource constraints and assessment literacy of colleagues in our quest to adhere to standards for the field and to communicate about the validity of the intended uses of this test. This paper will be of interest for anyone who is interested in the potential of re-organizing their test validation activities with reference to an argument-based validity framework.

References
The Effects of Administration and Response Modes on Grades 1-2 Students’ Academic Writing Performance

Ahyoung Alicia Kim, WIDA at University of Wisconsin-Madison, USA
Shinhye Lee, Michigan State University, USA
Mark Chapman, WIDA at University of Wisconsin-Madison, USA
Carsten Wilmes, WIDA at University of Wisconsin-Madison, USA

With broader access to ever improving technology, large-scale standardized assessments have increasingly transitioned from traditional paper-and-pencil to online platforms. Administering online assessments can have numerous benefits, such as allowing large numbers of students to take the test at the same time, thereby easing the administrative burden. Online assessments may also be designed to leverage the technology and deliver test content to students in a more engaging way, compared with paper-and-pencil tests. However, online tests may disadvantage test-takers who are unfamiliar with technology. This is especially concerning for second language (L2) writing assessments, in which students need to keyboard extended responses. Although a number of studies have been conducted on the topic of paper vs. online writing exams, they generally focused on adult learners (e.g., Barakaoui, 2014; Chen, White, McCloskey, Soroui, & Chun, 2011; Laborda, Royo, & Bakieva, 2016). Comparatively few studies exist regarding Grades K–12 English language learners’ (ELLs) paper vs. online writing performance (e.g., Renn, DeMarco, & MacGregor, 2015). Moreover, study findings have been inconsistent regarding the benefits of paper vs. online writing test modes. These studies mostly attended to test scores, while leaving much room for interpretation as to the actual quality of the written texts (McDonald, 2002).

The purpose of this study was to investigate how young students differ in their writing performance on paper vs. online test modes. This was done by drawing on both quantitative and qualitative research methods. Participants were 139 Grades 1–2 ELLs in the U.S. They completed three writing tasks, representing three different test modes: (1) Paper task (traditional paper-and-pencil format), (2) Online-Handwriting task (students read the prompt online and handwrite responses), (3) Online-Keyboarding task (students read the prompt online and keyboard responses).

Up to 20 students were tested during each 90-minute session, in which researchers observed two or three students to examine any differences in their writing behaviors across tasks. Immediately afterwards, researchers asked the same students about their preferred test modes. Moreover, four educators who worked with the participating students were interviewed to understand their perception of students’ computer familiarity (typing ability) and educators’ preferred test modes. Students’ performance was analyzed both quantitatively (using descriptive statistics and ANOVA) and qualitatively (fluency, accuracy, and complexity). Observation and interview data were analyzed qualitatively.

Findings show that children received the highest scores under the Paper condition. Specifically, Grade 1 students performed significantly better on the paper task than on the two online tasks, whereas Grade 2 students performed significantly better on the two handwriting tasks than on the keyboarding task. Results from qualitative analyses yielded similar patterns. Regarding fluency, students produced the most number of words on the Paper task and the least number of words on the Online-Keyboarding task. Interview findings provided additional insights. Students typically preferred the keyboarding task over the handwriting task due to the efficiency of keyboarding, which contradicts the fluency results. Meanwhile, most educators preferred the handwriting task over the keyboarding task.

Study results provide practical implications for L2 writing test administration in the K-12 context.
Analysis of Written and Oral Comments within Peer Feedback Processes from the Perspective of Recipients
Qiyun Zhu, University of Hong Kong, China

Recent conceptualization of feedback as a dialogic process has highlighted the receiver’s role in the effectiveness of feedback. Yet, the literature on student engagement with feedback often narrowly focuses on feedback provided by educators. Further, extant studies on written and oral feedback among peers concentrate on the delivery of comments, overlooking the perspective of feedback recipients. Against this backdrop, the study aims to understand the characteristics of different modes of feedback within peer feedback processes and the receiver’s role in these modes.

Contextualized in a writing curriculum for English language learners in a Chinese university, this two-semester-long study collected two sets of data from students. The first set included annotated drafts with peer comments (77 copies), completed peer review forms (44 copies), and transcripts of audio-taped peer discussion (6 groups). The second set were drawn from stimulated recall interviews with 11 focal students and their reflective journals. A coding scheme was adapted to analyze the first dataset, meanwhile content analysis was conducted on the second dataset. The analysis revealed distinctive characteristics of written and oral peer comments in terms of the type and depth of feedback, and different functions of these feedback modes for the reviewer and the receiver. Moreover, the analysis of oral interaction and student self-reports unfolded a complex process of knowledge co-construction. A continuum of receiver proactivity was also identified.

Based on these findings, a model of written and oral comments within peer feedback process mediated through receiver proactivity is proposed. Identifying the linkage between recipient proactivity and oral feedback, the study contributes to the knowledge about favorable peer feedback design that encourages students’ proactive recipience of peer feedback. Pedagogical implications are drawn accordingly.

Developmental Patterns of Young EFL Students’ Oral Narratives
Ching-Ni Hsieh, Educational Testing Service, USA
Yun Wang, Educational Testing Service, USA
Florence Tolentino, Educational Testing Service, USA

In language classrooms, storytelling is frequently used as an activity to help develop and evaluate young language students’ oral language proficiency (McKay, 2006). This study explored the developmental patterns of young EFL students’ picture-based oral narratives and evaluated the performance of students’ narratives within the context of a large-scale standardized language assessment designed for young EFL students—TOEFL Junior speaking test. A sample of 382 EFL students across four proficiency levels, between below A2 and B2 on the CEFR, was selected for the study. The students were ages between 9 and 16 (M = 12.8). These students’ responses to the picture narration task of the TOEFL Junior speaking test were used for a discourse-based analysis. The students’ narratives were rated for story completeness on a holistic, six-point scale and coded for narrative organization skills adapted from the Narrative Scoring Scheme (e.g., character development, use of mental state words, referencing) (Heilman et al., 2010). Aspects of vocabulary (e.g., number of word type, type-token-ratio) and grammatical accuracy (number of grammatically accurate speech units) and complexity (e.g., number of clauses per speech unit) were analyzed using human coders and natural language processing tools. Descriptive, correlation, ANOVAs, and regression analyses were conducted to analyze the data. Results of mean-based comparisons showed significant gains from low to high proficiency levels for story completeness, narrative organization skills, and measures of grammatical accuracy and complexity. Compared to students at the lower proficiency groups, students at the higher proficiency groups produced oral narratives with greater numbers of complete events and details, and by using more accurate and complex grammar. Lexical measures, however, did not show consistent differences across proficiency levels. The regression analysis revealed that the students’ scores on the picture narration task awarded by operational raters of the TOEFL Junior speaking test
was best predicted by global grammatical accuracy and story completeness. Results of the study have implications for the use of picture-based narration tasks to measure young students’ speaking proficiency.

Thursday, 5 July 2018
4:25pm - 4:55pm
Venue: Case Room 4 (260-009)

Evaluating Validity Claims for Existing Diagnostic Assessments: Consequences Revisited
Yong-Won Lee, Seoul National University, South Korea

Diagnostic language assessment (DLA) has recently garnered a great amount of attention from language testing researchers and practitioners. DLA is purported to identify learners’ strengths and weaknesses in a targeted domain of linguistic and communicative competence (Alderson, 2014, Lee, 2015). DLA can be regarded as effect-driven assessment (Fulcher & Davidson, 2007), for it has an explicit goal of positively impacting future learning by providing useful feedback and remedial advice based on diagnostic results. This implies that DLA should be evaluated with intended instructional impact clearly in mind, particularly the specific effects of the feedback and advice on test-takers’ further learning.

In conjunction with this, it is also worthwhile to note that there have long been heated debates in the fields of psychological and educational measurement on whether it is necessary to include consequences of testing in validity frameworks for tests. Some reductionists, such as Borsboom (2006) and Lissitz and Samuelson (2007), claim that the notion of validity should be confined only to the accuracy of score-based inferences, whereas expansionists, like Kane (2009) and Messick (1989), argue for inclusion of the consequences of test use and score-based actions in validity frameworks. Interestingly, DLA and other learning-oriented assessments provide nice testing grounds for such conflicting views on consequences in validity theorizing.

The main purposes of the present investigation are to: (a) examine the major arguments for, and against, the inclusion of consequences in the validation process for language assessment, particularly in the context of DLA; and (b) analyze some of the existing language assessments with diagnostic claims (e.g., DIALANG, DELNA, DELTA, and others) in terms of how the assessment-induced educational effects are conceptualized, measured, and evaluated in the form of validity evidence. A preliminary analysis has revealed that (a) most of these tests do have some feedback, advisory, and recommendation components that are intended to guide subsequent learning activities; (b) the granularity and actionability of the feedback information vary greatly across tests; and (c) there seems to be a strong need for systemically examining the positive (or negative) effects of such diagnostic/advisory information on learning from the consequential validity perspective.

In the presentation, the implications of the major study findings will be discussed in terms of not only planning appropriate validation strategies for DLA but also theorizing validity in general. The researcher also strongly argues for the need to expand the scope of validation in DLA to include the effects of not only quantitative (score profiles) but also qualitative information (feedback/advice) generated from the results of test-based diagnosis.

Thursday, 5 July 2018
5:00pm - 5:30pm
Venue: Lecture Theatre OGGB 4 (260-073)

Young Korean EFL Learners’ Reading and Test-taking Strategies in a Paper and Computer-based Reading Comprehension Test
Soohye Yeom, New York University, USA
Henry Jun, Seoul National University, South Korea

This study investigated whether young Korean students learning English as a foreign language (EFL) engage in similar processes when responding to reading comprehension questions delivered in two different test-delivery formats: paper and computer. Previous studies comparing reading comprehension tests in the two formats have mostly investigated first language (L1) reading comprehension and examined whether the test scores obtained from the two formats were equivalent. The present study examined differences in the processes that young EFL students employ while interacting with reading comprehension questions delivered in the two formats. More specifically, the study
investigated the relationship between reading test-delivery format (paper and computer) and test-takers’ reading comprehension strategies, as well as the degree to which the test-takers’ reading proficiency moderates this relationship.

Two forms of a reading comprehension test were developed and both were delivered via paper and computer. Eighty-four Korean middle school EFL learners took both forms of the test. Half of them were randomly assigned to take Form 1 on the computer and Form 2 on paper; whereas the other half took Form 1 on paper and Form 2 on the computer. Each form consisted of three passages and follow-up multiple choice questions. After finishing each passage, participants filled out a reading and test-taking strategy checklist, developed based on Cohen and Upton’s (2006) framework. The first 20 participants also participated in a stimulated recall interview after completing each checklist. In addition, all participants completed a background questionnaire and a posttest survey about their perceptions of the reading tests in the two formats.

Results revealed that there was no statistically significant difference between the participants’ scores on the reading tests delivered on paper and computer, and there were very few statistically significant differences in strategy use between the two formats. On the other hand, there were clear differences in the strategies used by learners with different levels of English reading proficiency. The high-proficiency group employed significantly more strategies to enhance the quality of their reading comprehension, whereas the middle-proficiency group utilized more strategies to compensate for their insufficient understanding of the passage and the questions, and the low-proficiency group reported more guessing moves.

Despite the similarities in test performance and strategy use between the two test delivery formats, the majority of participants preferred taking the test on paper and reported discomfort with the computer format. This might be due to the participants’ lack of familiarity with reading English on a computer screen, given that the background survey revealed that the participants rarely read English on a computer.

The findings of the present study support the equivalence of the computer and paper-based reading comprehension tests, in relation to the reading processes and performance of young Korean EFL students. However, the students’ lack of familiarity with reading on a computer had a negative effect on their perception toward the computer format, despite their similar performances on both formats. The implications of this study will be discussed in relation to current global trends towards switching from paper-based tests to computer-based tests.

Thursday, 5 July 2018
5:00pm - 5:30pm
Venue: Case Room 3 (260-055)

A Comparison of ESL Students’ Performance in Face-to-face and Virtual-world Group Oral Tests
Jayoung Song, Rice University, USA

For the past decades, computer technology has been increasingly used in all aspects of language testing, including test design and development, test administration, scoring, and analysis in order to accurately measure one’s second language speaking ability (Chapelle & Douglas, 2006). Although assessing oral language proficiency via Internet with computers or other forms of multimedia technology has been a popular trend in the field of language testing, it has some limitations. Some studies have found that semi-direct interviews such as tape-mediated interview (e.g., SOPI) elicit different language functions and skills (Shohamy, 1994; van Lier, 1989). More importantly, talking to the computer or tape-recorder does not measure interactional competence, that is, how a person manages a conversation in real contexts. In light of these issues, it seems timely to develop an interactive computer-based speaking assessment, in which test takers can discuss topics with each other using a computer (Ockey, 2009). One of the possible platforms could be a virtual world. The purpose of the present study is to explore if a virtual world could be a suitable platform for second language assessment. It specifically compares ESL students’ performance in a virtual-world group oral test and a face-to-face group oral test. A total of sixty-four ESL students enrolled in a large language institution in the southwestern part of the United States participated in the study. Thirty students served as a control group and took two sets of face-to-face and virtual world tests. Thirty-four students served as an experimental group and took two sets of face-to-face and virtual world tests after receiving training in the virtual world. Data was drawn from students’ group oral test scores, a survey asking their perceptions towards the two testing modes, and interviews. The findings showed that students produced similar scores when tested again in the virtual world, which assured test-retest reliability of the virtual-world testing mode. The results also revealed that students’ group oral test scores in the virtual world were comparable to their face-to-face group oral scores, which provided concurrent validity of virtual-world testing mode. It was noteworthy that students produced comparable scores in the virtual world when
they were tested again in the virtual world. Students’ perceptions towards virtual-world testing mode were promising in terms of self-efficacy, anxiety, and easiness of the testing mode. However, students reported that it was difficult to take turns in the virtual world due to the lack of visual cues in the virtual world. Qualitative analyses of interviews showed that virtual worlds have some benefits including interesting and relaxing testing conditions as well as a feeling of co-presence in the virtual world. The present study has several implications in the areas of language testing research and practice. Theoretically, the present study adds valuable knowledge to second language testing theory by providing validity evidence of a virtual-world group oral test. Practically, the study provides practical implications for test developers or language practitioners in terms of test design, the importance of training in the virtual world, and the importance of rater-training.

Thursday, 5 July 2018
5:00pm - 5:30pm
Venue: Case Room 4 (260-009)

Standards of Language Assessment for Placement Purposes: A Survey of University-level English Placement Tests
Jason Fan, Fudan University, China
Yan Jin, Shanghai Jiao Tong University, China

Recent years have witnessed the development and implementation of a host of professional standards, in the forms of Code of Ethics (e.g., ILTA, 2000), Guidelines for Practice (e.g., EALTA, 2005; ILTA, 2007), and Standards (e.g. ETS, 2014), representing the collective efforts of the field to strive for professionalism. However, empirical research of language testing practices is only piecemeal and has been slow to catch up. Adopting the good testing practice in Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014, hereafter the Standards) and ILTA Guidelines for Practice (ILTA, 2007, hereafter the Guidelines), this study made a foray into the practices of university-level English placement tests in China, which are typically developed and used by universities on their own. This study investigated two research questions: 1) What standards do universities follow in the development, administration and use of English placement tests? 2) What are difficulties and challenges facing universities in English placement testing?

The participants of this study were 63 representatives in charge of the English placement test in their universities from across China. The research was divided into two phases. During the first phase, a questionnaire was constructed on the basis of good testing practice prescribed in the Standards and the Guidelines, as well as a review of literature on placement testing. The questionnaires were then administered to the 63 representatives. During the second phase, 15 representatives participated in one-on-one semi-structured interviews. Due to the small sample, mainly frequency statistics were calculated to portray a picture of the testing practices of the participating universities. The interview data were analyzed using the inductive analysis approach.

Results identified a number of mismatches between the current placement testing practices as reported by the universities and the good testing practice in the Standards and the Guidelines. On the whole, the participating universities did not indicate clearly which standards they followed in test development, administration and use, and much variability could be revealed between the practices of different universities. Regarding test content and format, receptive language skills were much more prominently represented than productive skills. Analyses of interview data indicated that placement test results were used to make a variety of decisions, some of which were high-stakes in nature. The typical challenges in placement testing included time pressure, lack of human resources, lack of expertise in language testing, and the demand from and communications with Academic Affairs Office (AAO) which is responsible for formulating and implementing policies regarding English teaching and learning in Chinese universities.

As one of the few attempts to understand the role of standards in current language testing practices, this study has implications for the development and implementation of standards in language testing in the interest of quality and fairness. Furthermore, it also highlights the social and political dimensions of language testing practices, and the demand from and communications with Academic Affairs Office (AAO) which is responsible for formulating and implementing policies regarding English teaching and learning in Chinese universities.
Developing a Social Justice Questionnaire for Language Assessment
Matthew Wallace, University of Macau, China

Fairness and justice are receiving renewed attention in the language testing literature of late, as both are considered essential factors for inclusion into a validation argument for test administration and use. Fairness research has focused on providing evidence of unbiased test interpretation and use through statistical analysis of test scores (Kunnan, 2017), while research into justice has examined the consequences of test use and how outcomes can be used to influence society and represent societal values, both explicitly or implicitly (McNamara & Ryan, 2011). Test developers and testing researchers use findings from these studies to provide evidential and consequential evidence in support of test validity (Messick, 1989). However, noticeably absent from the literature are the perceptions of test administration and use of the test-takers themselves. Fairness judgments can have a rather profound influence on the attitudes and motivations of individuals being assessed, so it is essential to gain a better understanding of how fair and/or just test takers view test administration episodes and testing entities. Doing so will provide key additional information for the validation argument. Drawing from the social justice literature, a questionnaire was adapted from Colquitt’s (2001) organizational justice survey to measure Japanese EFL test-taker perspectives of a standardized English test. The 20-item questionnaire consisted of three sections, each designed to measure one of the dimensions of social justice—distributive (4 items), procedural (7 items), and interactional (9 items). *Distributive* justice refers to fairness of outcomes (e.g., whether test scores reflected the efforts made by test takers), *procedural* justice concerns fairness of procedural elements (e.g., whether test scores were unbiased), and *interactional* justice relates to fairness of interpersonal interactions (e.g., whether the test administrator provided feedback in a timely manner) (Blodget et al., 1997). Items were presented as statements and the 100 senior high school participants indicated how much they agreed with each using a five-point Likert scale, ranging from “to a small extent” to “to a large extent.” Data from the survey was examined using Rasch analysis to confirm the three-factor structure of the instrument. Results revealed the degree to which participants considered the test administration to be fair and just. This unique perspective provided test developers, users, and administrators with useful information from which to evaluate the quality of the test administration. Overall, this study contributes a new, test-taker oriented perspective of fairness and justice of language tests, as well as a tool with which to measure it.

Lecturer English Proficiency Assessment: What Norms?
Slobodanka Dimova, University of Copenhagen, Denmark

The growing lecturer and student mobility in European and Asian higher education has resulted in establishment of English-medium instruction (EMI) programs at traditionally non-Anglophone universities. This raised concerns about the EMI quality, so many universities implemented policies for assessment of lecturers’ English proficiency (Ball & Lindsay, 2013; Klaassen & Bos, 2010). Given the linguistic and cultural complexity of the academic contexts in which these assessments are used (Björkman, 2010; 2011), making decisions about what English norm lecturer performances are assessed against has been a struggle. Selection of appropriate norm(s) can warrant the quality of the assessment instrument and the validity of the procedure. Therefore, this paper discusses the validity of the “educated native-speaker” norm in the scalar descriptors of a performance-based EAP test used for EMI lecturer oral proficiency assessment. More specifically, it presents two complementary studies that investigate 1) rater behavior, i.e., applications of the native-speaker reference when assigning scores and writing feedback and 2) EMI students’ preference for lecturers in relation to their perceptions of lecturers’ speech accentedness, proficiency, and comprehensibility.
Analytic test scores on a five-point scale (pronunciation, vocabulary, grammar, fluency, and interaction) and written formative feedback reports (n=400) from nine trained raters were used to analyze rater behavior. Then, EMI students’ (n=95) rated sixteen test performances on a nine-point scale in relation to the strength of test-takers’ (lecturers’) non-native accent, proficiency and comprehensibility levels, and lecturer preference, i.e. how much they would like to take a course with the test-taker.

Data from trained raters were examined using multi-facet Rasch analysis (Linacre 2010), while feedback reports were analyzed in Nvivo10. Descriptive frequencies and Pearson correlation were used for student data analysis. Results suggest that raters focused on pronunciation and assigned top proficiency level only to native-like performances based on inner-circle norms. The written reports extensively used native-speaker references in description of test-takers’ oral performances. Student ratings for test-takers’ accent, comprehensibility, and proficiency were aligned with the five-point scale of the test; test-takers who were rated by the trained raters at the higher end of the scale (levels 4 or 5) received high proficiency and comprehensibility scores and low accentedness scores by the students. However, while students’ test-taker ratings were similar for test-takers at levels 4 and 5 (comprehensibility means, L4=7.4, L5=7.5; proficiency means, L4=6.8, L5=7.3), level 5 was rated higher than level 4 regarding accent (means, L4=4.6, L5=6.5). Moreover, the correlations between lecturer preference and proficiency (r=.78) and comprehensibility (r=.67) were stronger than that between accent and lecturer preference (r=.56).

Drawing from these findings, the educated native speaker reference was removed from the scale descriptors, and a rescaling procedure was undertaken to expand the highest scalar level to include more variation.

Friday, July 6
2:00pm – 2:30pm
Venue: Case Room 3 (260-055)

The Effect of Multilingual Assessment on the Science Achievement of Pupils
Fauve De Backer, Centre for Diversity and Learning - Ghent University, Belgium

Migration waves have resulted in an increasing number of pupils that speak another language at home than the language of schooling. There is an increasing concern that multilingual pupils are being disadvantaged in traditional testing situations, since their linguistic ability in the language of schooling may impact their results on content-related tests. In mathematics and science tests, the goal is to assess what a pupil knows and is able to do in mathematics or science. While language abilities should not confound test results, every test that uses language cannot avoid being also a language proficiency test. To ensure fair and valid testing, test accommodations have been proposed. The Educational Testing Service (ETS) divides accommodations into two categories; indirect linguistic support results in adjustments to the test conditions, e.g. modifications of the protocol, such as giving extra time or testing pupils individually. Direct linguistic support entails adjustments to the language of the test and thus involves modifications to the test itself. To date, there have been few empirical investigations of the effects of test accommodations on actual achievement and the available research is inconclusive. In the present study, 752 pupils in fifth grade (age 10) of 35 primary schools in Flanders were randomly assigned to one of six research conditions: ‘a Dutch-only test’ (DU/A-), ‘a Dutch-only test with audio-support’ (DU/A+), ‘a Dutch-Polish test’ (PO/A-), ‘a Dutch-Polish test with audio-support’ (PO/A+), ‘a Dutch-Turkish test’ (TU/A-), ‘a Dutch-Turkish test with audio support’ (TU/A+). In this presentation, results from ANOVA analyses will be reported that indicate the effectiveness of oral accommodations in both L1 and in the language of schooling and the effectiveness of the written version of the bilingual test. Implications of these results for large-scale international assessments will be discussed.

Friday, July 6
2:00pm – 2:30pm
Venue: Case Room 4 (260-009)

The Effect of Nonverbal Communication on Listening Item Type Performance
Aaron Olaf Batty, Keio University, Japan

Comparative studies between audio- and video-mediated listening tests abound in the literature, but findings have been inconsistent. A small majority find that the presence of visual information facilitates test performance, while
most of the remainder finds no effect. A closer inspection of the literature, however, reveals gross dissimilarities in methodology, test design, item formats (e.g., constructed response vs. multiple-choice), and item types (e.g., explicit vs. implicit). As such, it remains unclear as to what effect the presence of visual information has on listening tests generally, and whether item types are equally affected. Several researchers in the video-mediated listening test literature (e.g., Batty, 2015; Ockey, 2007; Wagner, 2002) have suggested that a video effect is more likely to manifest in listening items requiring inference, especially those within social contexts, as nonverbal communication features prominently there. The present research investigates this suggestion by way of a mixed-methods quasi-experimental study employing many-facet Rasch measurement (MFRM) to compare the difficulty of implicit and explicit L2 listening item types when delivered via video or audio-only.

Previous mixed-methods eye-tracking research by the present author (presented at LTRC 2016) found that the overwhelming majority of examinee attention in video-mediated listening tests is paid to facial expressions, suggesting that the primary difference between audio and video-mediated listening tests is the absence or presence of this nonverbal communicative information. These findings informed the development of an original 20-item video-mediated English listening test. The test was comprised of short conversational scenes featuring tightly scripted and directed nonverbal behavior, and performed by professional actors. Each scene was paired with a single multiple-choice item. Of the items, ten were explicit, requiring only comprehension of the propositional content of the related scene, and ten were implicit, requiring inference. The 279 Japanese university student participants were randomly assigned to either encounter the odd items in the video condition and the even in the audio, or vice-versa. The test was administered online with the Moodle learning management system’s quiz module. The resulting data were analyzed with MFRM to estimate the size of differences in item and item-type difficulties under the two conditions, and qualitative item analysis was carried out to aid in the interpretation of the quantitative results. The presence of nonverbal communicative cues in the video condition was found to be broadly facilitative, with many p values below .05, but the effect was especially pronounced in the implicit items, with an average of over 1 logit of difference in difficulty between the formats, a difference with a p value below .001. Qualitative analysis of the implicit items revealed likely explanations for the differences, most of which can be understood to be the direct effect of facial expressions.

The presentation concludes with a discussion of the theoretical implications of the findings with regard to validity and a discussion of the practical considerations for test developers considering a transition to video-mediated L2 listening assessment.

---

Friday, July 6
2:35pm – 3:05pm
Venue: Lecture Theatre OGGB 5 (260-051)

Language and Citizenship: A2 Tests and Low-educated Learners
Bart Deygers, KU Leuven, Belgium

This study focuses on a group of test takers who have been underrepresented as a research population in the high-stakes language testing literature (Carlsen, 2017; McNamara & Shohamy, 2008; Tarone & Bigelow, 2012): low-educated, low literate language learners (LESLLA). There are no systematic reviews into testing methods that are appropriate for these learners (Allemano, 2013), no agreed-upon methodological approaches to research this population (McNamara & Shohamy, 2008), and many theories and constructs do not take full consideration of the idiosyncrasies of low-educated learners’ test-taking strategies (Carlsen, 2017). This is remarkable, given the fact that LESLLA learners account for a large proportion of the migrant population (UNESCO, 2014).

In Flanders, Belgium - the research context of this contribution - one third of the migrant population has a LESLLA background. The ministry for integration requires all migrants to prove A2 competence in Dutch on the four skills to fulfill their citizenship requirements. The instrument used to assess A2 language proficiency is a newly developed centralized, computer-based test, but the fitness-for-purpose of this instrument for the group of LESLLA learners has not been investigated. Examining the impact of the new A2 citizenship test on this specific population is the purpose of this study.

First, the LESLLA learners’ performance data will be compared with the larger population (N >1000). The results show that there is a significant difference between the pass probability (p <.000) of the LESLLA population compared to that
of the larger population, and that the test items are biased for educational background. In the second part of the presentation, a qualitative layer will be added to the quantitative findings. In an eye tracking study, we compared the strategies employed by higher and lower educated test takers. The results of this indicate that low-educated learners engage fundamentally differently with test prompts and with the interface of the computer test, compared to their higher educated peers.

This research has considerable implications that go beyond language testing as such. Consequently, after presenting the empirical data, the findings will be discussed in the light of validity, fairness, justice, and human rights. References will be made to insights from within the language assessment literature (ILTA, 2000; McNamara & Ryan, 2011; McNamara & Shohamy, 2008), and to more philosophical sources (Rawls, 2001; Sen, 2010; Young, 2011).

Friday, July 6
2:35pm – 3:05pm
Venue: Case Room 2 (260-057)

The Domain Expert Perspective on Workplace Readiness: Investigating the Standards Set on the Writing Component of an English Language Proficiency Test for Health Professionals
Simon John Davidson, University of Melbourne, Australia

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

The writing task on the OET is a letter of referral, based on a set of provided case notes. 18 doctor participants (all with experience of working as medical educators, GPs or specialists) were recruited to participate in standard-setting workshops designed to elicit decisions about what level of performance on this task was deserving of a passing grade and why. Five OET writing samples representing different performance levels were used in the workshops and an additional 30 take-home samples were issued. To gain further insight into the basis for the standards set, verbal reports in the form of a think-aloud protocol (TAP) were employed with five of the 18 workshop participants. These participants were separately asked to say ‘out loud’ what they noticed while assessing the competence of 12 further samples. A FACETS analysis (Linacre, 2017) was used to calculate new passing standards and compared with current OET cut-scores. In addition, the doctors’ comments were thematically coded and intercoder reliability checks were conducted.

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

The English Impact (EI) project assesses English language capability by measuring the current ability of a targeted sample of the school population using an English language assessment, and evaluating potential through an in-depth analysis of students’ language learning opportunities in and outside the classroom, their language learning motivations and socio-economic backgrounds. The project employs a two-stage cluster design, sampling schools at the first stage and students at the second stage. Typically, 120 schools are randomly sampled from a list of all appropriate institutions. Then 12 pupils at the target age (15.5) are randomly sampled from each school. Due to the complex nature of the sampling approach, the local ministries were heavily involved at this stage of the project in all cases. The project was undertaken in four countries or regions, Bangladesh, Columbia (Bogota), Spain (Madrid) and Sri Lanka. In this paper we focus on the data from the Communidad de Madrid (the Madrid Region) EI project as it is of particular interest, given the significant changes made to the education system there in recent years. The most contentious of these changes has been the introduction of the bilingual schools programme, in which a range of subjects is taught through the medium of English (EMI). Currently, there are over 110 secondary schools (17%), located across a range of socio-economic areas in Madrid, involved in the programme. In spite of criticism in the media and in parliament, the Ministry supported this project, and at no time attempted to influence its outcomes. Because an additional variable was included in this particular study (school type) the sample design was altered to include additional bilingual schools. Statistical modelling indicated that 170 schools should be included in the design for Madrid, including 45 bilingual schools. This final dataset reported on here includes 1,774 pupils from 169 schools.

Following the initial construct validation of the motivational scales using CFA, multi-group comparisons were carried out for three groups of the data (gender, socio-economic status, school type). Comparisons were then reported for the relative levels of motivation expressed for each scale and the relationships between the different areas of motivation. In order to further explore the data, a Latent Variable Mixture Modelling (LVMM) approach was used to detect distinct patterns of motivation reported by participants. This is a person-centered analysis, grouping participants according to a profile of their reported behaviours and achievements, rather than using biodata or social characteristics. The latent classes formed were then used to establish (using a probabilistic statistical model) which of the participants likely belongs to each class. Findings from the two types of analysis were found to offer interesting alternative perspectives on the data. While it was not the focus of the project, the results do appear to offer some consistent support for the bilingual programme, while highlighting significant differences within the population. In this paper we overview the EI project while focusing primarily on the findings from the two sets of analyses and of the implications for policy makers, researchers and test developers.

Looking into Listening: Establishing the Cognitive Validity of the Aptis Listening Test
Kathrin Eberharter, University of Innsbruck, Austria
Franz Holzknecht, University of Innsbruck, Austria
Benjamin Kremmel, University of Innsbruck, Austria
Gareth McCray, Keele University, UK
Matthias Zehentner, University of Innsbruck, Austria
Eva Konrad, University of Innsbruck, Austria
Carol Spöttl, University of Innsbruck, Austria

Although listening comprehension is central to language learning, it is perhaps the least understood skill and research into the cognitive processes underlying the listening construct is still sparse. However, understanding the cognitive processes listeners employ when trying to make sense of auditory signals should be a central focus in listening.
assessments. This study investigates the processing of test-takers while completing the British Council’s Aptis Listening Test, in order to discern whether the empirically measured cognitive processing matches with that intended by the test designers and embedded in the item specifications. Specifically, it examines whether test-takers’ cognitive processes and types of information used in correctly responding to items correspond to those targeted by the specifications at the different CEFR levels targeted by the Aptis test.

Data on participants’ cognitive processing was collected on 25 four-option multiple choice Aptis items (CEFR levels: 7 A1, 7 A2, 6 B2, and 5 B2 items) via stimulated recall and, innovatively for a listening study, eye-tracking. Two primary methodological drivers of this study were to explore the utility of eye-tracking for the determination of cognitive processing during listening tests both by analysis of eye traces and also as a prompt for generating rich stimulated recall data. Eye-tracking data were collected from 30 participants (16 female and 14 male) and stimulated recall data was gathered and analysed from 16 of the 30.

The amount of time fixating on a response option was modelled in a mixed effects linear regression based on various explanatory variables, including location of the response on the screen, intended CEFR level of the item, and test-taker’s listening and reading ability. The stimulated recall data was coded based on Field’s (2013) model of cognitive processes in listening and the type of information used for answering the item as specified in the Aptis test specifications (O’Sullivan & Dunlea, 2015).

The quantitative analysis of the eye-tracking metrics revealed interesting results. As expected, test-takers fixated the response options of higher-CEFR-level items longer than those of lower-level items. Response options higher up the screen were fixated longer than those lower down, and better readers focused on response options longer than poorer readers. Triangulating this with the analysis of the stimulated recall protocols, it was found that Aptis items largely tapped into the cognitive processes intended by the test developers, but that differences between the CEFR levels in relation to the intended cognitive processes could be more pronounced. Combining eye-tracking with stimulated recall provided novel insights into the cognitive processing during listening that could not be gathered through other data collection methods. For example, the finding that the location of a response option on the screen has such a dramatic impact on the amount of fixation, and thus processing, requires further study. We will thus argue that further work is needed to ensure that the effect of response order does not affect the difficulty of an item and introduce unwanted bias into the test taking process.

Friday, July 6
3:10pm – 3:40pm
Venue: Lecture Theatre OGGB 5 (260-051)

Test Fairness: Examining Differential Functioning of the Reading Comprehension Section of the GSEE in China across Graduation Time

Shangchao Min, Zhejiang University, China
Lianzhen He, Zhejiang University, China

Numerous empirical studies have been conducted to investigate test fairness by examining item-level differential item functioning (DIF) (e.g., Koo et al., 2014) and test-level factorial invariance across different subgroups (e.g., Trace et al., 2017), yet very few efforts have been devoted to examining the relationship between DIF, differential bundle functioning (DBF) and differential test functioning (DTF) (e.g., Kim & Jang, 2009), although the detection of DBF and DTF is particularly important as decisions are in general made at the higher level of bundles or tests. In addition, although a plethora of research have examined DIF across gender (e.g., Aryadoust et al., 2011), ethnicity (e.g., Davidson, 2004), language backgrounds (e.g., Elder, 1996) and academic backgrounds (e.g., Fidalgo et al., 2014), little is known about test fairness for test takers across graduation time. This is an important concern for the Graduate School Entrance English Exam (GSEE) in China, where the test-taker population consists of both new graduates who are in the final year of their undergraduate programs, and former graduates who graduated one or more years ago. The GSEE is a high-stakes English test designed and administered by the National Education Examinations Authority (NEEA) for the purpose of selecting candidates for Master’s programs in educational and research institutions in China. The purpose of this study, therefore, was to examine item-level DIF, testlet-level DBF, and test-level DTF in terms of test-takers’ graduation time in the reading comprehension section of the Graduate School Entrance English Exam (GSEE) in China. The datasets included 6,000 test takers’ item-level responses to 6 five-item testlets, namely, 6 sets of passages with 5 items pertaining to each passage. Both DIF and DBF were examined by using Poly-SIBTEST and IRT-LRT, and DTF was investigated with multi-group confirmatory factor analyses (CFA). The results indicated that
although none of the 30 items exhibited statistically and practically significant DIF across graduation time at the item level, 1 testlet was consistently identified as favoring new graduates over former graduates by the two DBF procedures. Content analysis revealed that the differential performance on this testlet across new and former graduates may be attributed to their differential access to learning opportunities and differential background knowledge of the topic. Nonetheless, the DBF does not manifest itself at the overall test score level to produce DTF. We also found, by multi-group CFA, factorial invariance in the reading comprehension section across new and former graduates, substantiating the absence of test-level DTF. The findings revealed that the relationship between item-level DIF and test-level DTF is a complicated issue with the mediating effect of testlets in testlet-based language assessment. Additionally, by examining differential functioning across graduation time, this study has bridged a gap in the published DIF literature while also contributing to the fairness investigations of the GSEEE.

Friday, July 6
3:10pm – 3:40pm
Venue: Case Room 2 (260-057)

Exploring the Effect of Classroom-based Assessments on Medical Students’ EMP (English for Medical Purposes) Learning
Lan Luo, Zhejiang University, China

Research on classroom-based ESP (English for Specific Purposes) assessment is nascent and many issues deserve further investigation. The quality of the assessment procedures is one of the most vexing issues. In the ESP classroom in particular, the standards for reliability and validity were considered vague and malleable. Theoretical frameworks need to be adopted to ensure the quality of the classroom-based ESP assessments, and therefore to facilitate student learning. This study highlights the importance and necessity of ESP testing and aims to investigate the effect of classroom-based assessments on medical students’ EMP learning. With the collaboration between medical professionals and language testing researcher, five classroom-based EMP tests were designed and implemented into the full-blown research with the aim to explore their effectiveness in promoting medical students’ EMP skills as well as the extent to which they have in fostering students’ motivation and learner autonomy in EMP learning.

Students from the EMP course were divided into two groups at random, with one group as the experimental group and the other the control group. T-test and correlation analysis indicated that the two groups had significant differences neither in English proficiency (based on a reading comprehension test taken at the beginning of the course), nor in professional ability (based on students’ GPA in compulsory medical courses). Both groups participated in the pretest and posttest questionnaire as well as the final exam, but only the experimental group completed all five classroom-based EMP tests at the end of each corresponding lesson. The feedbacks were given back to each student in the experimental group through an e-portfolio tool (Seesaw App) to help them track their learning process. Students’ behavior in class were all videotaped and recorded through classroom observation checklists by three observers. Furthermore, follow-up student interviews provided in-depth understanding of medical students’ perspectives towards EMP assessments and EMP learning.

Results revealed that the students in the experimental group performed significantly better (p<0.05) than the control group in medical term translation. No significant differences were found between the two groups regarding the main factors of learning motivation and learner autonomy, but when each questionnaire item was examined alone, the experimental group experienced more significant increases in items associated with intrinsic motivation and self-regulation. The students from the experimental group also showed more positive attentiveness, involvement and attitude in class, and had more positive experience in EMP learning. The study not only explored and reported the development and validation of classroom-based EMP assessments, but also revealed its positive effect on students’ learning. It is believed that this study would shed some lights on the future research of classroom-based ESP assessments as well as contribute to the reform of EMP assessments and EMP learning.
Friday, July 6
3:10pm – 3:40pm
Venue: Case Room 3 (260-055)

Community-based Participatory Research Applied to Language Assessment
Lia Plakans, University of Iowa, USA

This conceptual research presentation proposes Community-Based Participatory Research (CBPR) as an approach to research in language testing. CBPR is used in the field of public health and involves equal involvement of community members in research at all stages from conception to application. This level of involvement with those on the use side of tests is not common in language assessment. However, research is generally undertaken from a practical stance or as Fulcher (2015) describes through pragmatic realism, focusing on tests and testing contexts with applied results. Given this epistemological stance, adopting CBPR seems like a promising path for language assessment research. Given the ever-changing contexts for language assessment use, having embedded and contextualized research approaches benefit the field with research from partnerships that foundationally includes experts in the contexts and the changes therein.

The unique defining feature of CBPR is the researcher-community partnerships that bring different but equally valuable expertise to a problem with the end goal of action to benefit the community (Isreal et al., 2001, 2012; Minkler & Wallerstein, 2008). While the consequences of test use have long been recognized in the assessment process and validation (Kane, 2006; Messick, 1989; Shepard, 2007), community members are generally designated as test users not collaborators. CBPR in language testing would move test users’ role more centrally in the research process as partners. The conceptual and procedural foundations for CPBR will be detailed in this presentation.

To illustrate CBPR in language assessment, the presenter will detail a partnership between a language testing researcher and a K-12 bilingual education (dual language) program. The study began through conversations between the partners about what testing research was most urgent in the school setting. The partners determined a need to evaluate the districts’ recently developed testing system. Through discussions, they narrowed to questions of how the assessments were fulfilling their established purposes. The tests and scores were used in the dual language program to (1) establish learning trajectories for individual students, (2) communicate progress to teachers, students, and parents, and (3) justify the ongoing support of the dual language program to the local school board (an advisory committee of local citizens). With these goals in mind, the partners created a plan to study the assessments by reviewing assessment documentation together and analyzing two years of student test data.

The study will be presented in light of insights for the use of CBPR as a research framework. The approach holds promise in sharing the processes of assessment literacy development with community partners, maintaining the pragmatic-driven nature of language testing research, and building partnerships for sustained long-term study of tests in use. While inherently messy, community-based research reflects the challenges community members face in test use. It also affords the field a foray into investigation of changes in contexts of assessment use. The session will conclude with challenges of this approach in language testing and a model for sustaining CBPR work over time.

Friday, July 6
3:10pm – 3:40pm
Venue: Case Room 4 (260-009)

The Impact of Connected Speech on Item Difficulty in Second Language Listening Assessment
Jonathan William Trace, Keio University, Japan
James Dean Brown, University of Hawaii at Manoa, USA

Naturally occurring connected speech is predictable, patterned, and most importantly ever-present in the language that L2 listeners encounter in the world. Yet traditional second language listening assessments remain fixated on constructs of comprehension that are rooted in the learners’ ability to recognize features of written language rather than natural speech. While unscripted passages that contain connected speech are becoming more commonplace in L2 assessment contexts, the unit of analysis and measurement remains centered upon the comprehension of individual words and phonemes. Such boundaries, however, are characteristic of written speech, and are seldom as clear cut in naturally occurring speech. Where written speech has visible shapes and spaces between words to delineate the separation and connection of words and letters, aural speech is reliant upon other factors to communicate meaning. Among these characteristics are features of intonation such as pitch, speed, and stress length.
While previous studies have begun to examine the degree to which second language learners are aware of and able to parse connected speech, the impact of these different features on item difficulty in listening assessment remains unexplored.

The current study sets out to do just this, by looking at the function of listening items across both connected and non-connected speech specific measures. Data were gathered on three 50-item dictation tests administered to 182 second language university students entering a North American academic English university program. Two scripted connected speech dictations—one based on an informal dialogue between two people and one a narrative passage—were designed to specifically target awareness of common instances of connected speech. In addition, items from an existing academic, scripted dictation were also analyzed for comparison purposes. Each of the 150 items were quantified based on a number of vocabulary-based variables (i.e., length, frequency, collocation) as well as connected speech characteristics (i.e., pitch, stress length, intensity). These variables were then entered into a principle components analysis and structural model to examine their relationship and influence on item difficulty, as determined by Rasch logit measures. Based on the results, item difficulty is influenced differently in different connected speech assessment conditions, which can have a meaningful impact on how items that—either intentionally or otherwise—incorporate natural speech are interpreted for second language listening purposes. Hopefully, this study will help keep the discussion about the need to include connected speech in both second language teaching and assessment moving forward.
Demo Presentations

Wednesday, July 4
11:40am - 12:10pm
Venue: Case Room 3 (260-055)

Diagnosing Linguistic Problems in Chinese Students’ Academic Writing in English: Specifying the Target Constructs for Assessment Design
Qin Xie, The Education University of Hong Kong, China

Defining and specifying target constructs have been a longstanding obstacle for the purposeful development of diagnostic assessment instrument. This paper reports the innovative procedures we adopted to identify and specify the target constructs for developing assessment tools, which can diagnose problems in the English academic writing of undergraduate students. The method harnessed the power of meta-analysis, manual and computer-assisted error analysis and corpus linguistic tools, which, we believe, bear originality and methodological rigor for test development and validation.

To start with, systematic meta-analyses were conducted to survey and analyze the linguistic errors investigated by existing studies in the literature (between 1988 and 2015). This meta-analysis generated a list of errors which either frequently appeared in Chinese students’ academic writing in English or (not of high frequency but) were considered to be serious by target readers. From the list of high-frequency and grave errors, those that could be reliably detected by existing computational technology were excluded. The final list contains 25 linguistic errors at lexical and syntactical levels. A detailed error-tagging manual was developed to guide manual error tagging. Manual error tagging was conducted by three raters subsequently with 387 essays selected from an in-house written corpus of over 1000 untimed and researched academic essays written by first-year undergraduates in Hong Kong. Rigorous measures were adopted to monitor and control inter-coder reliability and coding validity during the multi-stage coding process.

Finally, error tags were extracted and examined using the software AntConc. Statistics of the error tags were computed for error prevalence and frequency in order to validate the error list regarding their appropriateness for the target student population. The empirically generated rank-ordered error list was also compared with the one generated from the meta-analysis conducted earlier to explore and understand the differences. Target constructs were further specified based qualitative examination of the linguistic properties of the error-tagged sentences and the sub-categories under each error code. This textual analysis prepared rich and authentic linguistic materials for developing test items to diagnose the target linguistic problems in students’ writing.

Besides detailed description and demo of the methods and techniques adopted and their benefits, the issues and difficulties encountered will also be discussed and shared. Towards the end of this presentation, sample diagnostic items developed from this approach will be presented. Potential applications, as well as the further direction of research, will be discussed.

Thursday, 5 July 2018
1:50pm - 2:20pm
Venue: Lecture Theatre OGGB 4 (260-073)

The Duolingo English Test
Burr Settles, Duolingo, USA
Cynthia M. Berger, Duolingo and Georgia State University, USA
Masato Hagiwara, Duolingo, USA

Language assessment in the twenty-first century is being challenged by the large number of test takers globally, the high-stakes nature of standardized language tests, and the need to offer assessment instruments that are equally fair, valid, and accessible (Chalhoub-Deville, 2010). Computer adaptive tests (CATs) are well-poised to meet this challenge, due to their ability to provide uniformly precise scores for examinees of all levels, rather than just the average level (Thissen & Mislevy, 2000), and to deliver significantly shorter tests while maintaining this higher level of precision (Weiss & Kingsbury, 1984). Internet-based CATs are particularly promising today, given that nearly half of the world’s population now lives with Internet access (ITU, 2017). This new cohort of potential examinees represents a wide range of diverse economic backgrounds and geographically remote locations not well-served by traditional test instruments.
This demo presentation describes the Duolingo English Test (DET), an Internet-based general English language assessment which is available to language learners, employers, and educational institutions worldwide. The DET combines on-demand delivery with remote proctoring and the following innovations: (1) a statistical machine learning approach to creating criterion-referenced assessment constructs based on the Common European Framework of Reference, (2) modern interactive item formats that can be generated in large quantities and automatically calibrated for a more secure item bank, and (3) an efficient CAT protocol that uses novel item scoring algorithms. Because each test administration is recorded for security purposes, each score report also includes a writing sample and video interview of the examinee, providing important supplemental qualitative information for high-stakes decision-makers. Ultimately, we believe that the design and delivery of tests like the DET can offer more accessible, accurate, and practical options for global language assessment in the twenty-first century.

In the demo, we overview the design and development process for the DET, including construct operationalization, objectives, and current evidence for reliability, validity, practicality, and impact. We walk through the DET test items, scoring procedures, and rationales for their development and evidence of their predictiveness. We also present several case studies of how the DET is already being in high-stakes settings for both typical and underserved populations, and conclude by discussing future directions, challenges that the DET has encountered, and ongoing efforts to improve both test security and usefulness to examinees and institutions alike.

Thursday, 5 July 2018
5:00pm - 5:30pm
Venue: Case Room 2 (260-057)

Tuning into the Listeners’ Brain: Providing Neurocognitive Evidence for Validity through the Near-Infrared Spectroscopy (fNIRS) Technology
Vahid Aryadoust, National Institute of Education, Singapore; Nanyang Technical University, Singapore
Chiao-Yi Wu, Nanyang Technical University, Singapore

Language learning and assessment researchers have widely used quantitative and qualitative methods to investigate behavioral data of language learners and test takers. Although these research methods have extended our knowledge of language learning and assessment, there is a dearth of research on the neurocognitive mechanisms that correlate with such behavioral data. The goal of this project is two-fold: first, it introduces the concept of “neurocognitive evidence for validity” which has been developed by the presenters. The concept of validity has been founded upon test content, test takers’ cognitive processes, and to a higher extent on psychometric analysis. However, we propose that this concept needs to be updated and the methods to investigate validity should be integrated with new technologies that can provide information about the neurocognitive dimensions of language assessment.

To address this gap, the second goal of the project is to demonstrate the application of high-density functional near-infrared spectroscopy (fNIRS) to investigate the neurocognitive processes involved in language assessment. fNIRS is an advanced technique to measure human brain activities through hemodynamic responses (i.e., changes in oxygenated and deoxygenated hemoglobin that are induced by neural activities). We demonstrate the main steps of using fNIRS which include connecting the optodes (optical sensors) and the device as well as the software to record data. We illustrate the relevant software to develop proper “events” and “blocks” to make a “paradigm” for brain imaging. The software chosen for this project is further able to synchronize other sources of data such as the eye-tracking data with the listening test, and the fNIRS data. Next, we present relevant software packages which are used for various pre-processing steps (data trimming, filtering, etc.) followed by specific statistical analysis of fNIRS signals to identify the brain regions with hemodynamic effects.

Finally and as a case in point, we use fNIRS to measure hemodynamic responses in the pre-frontal areas and explore whether cognitive functions during listening assessment are correlated with brain activities. We also discuss that, as fNIRS is non-invasive and portable, it is feasible to conduct language assessment experiments in different settings (e.g., classroom or lab) without disrupting students’ assessment experience. To our knowledge, this project is the first study using advanced brain imaging techniques in language assessment. Another innovation of the project is the introduction of neurocognitive evidence for validity which is discussed in further detail in the presentation.
Determining Seat Time for a Shortened Section of a High-stakes Standardized Test
L.D. Nicolas May, Cambridge Michigan Language Assessments, USA
Patrick M. McLain, Cambridge Michigan Language Assessments, USA
Stephen P. O’Connell, Cambridge Michigan Language Assessments, USA
Gad Lim, Cambridge Michigan Language Assessments, USA

In high-stakes standardized testing contexts, practical factors such as item count and seat time are important concerns. When Michigan English Test (MET) stakeholders conveyed feedback that both the number of items and duration of the test were onerous, the matter was investigated. After confirming from multiple sources of evidence that the concerns were well-founded and that a shortened version of the test would maintain high reliability and content validity, the test developers faced the challenge of determining an appropriate new seat time that avoids undue speededness – a practical concern for which there is little information publicly available.

Test speededness occurs when some test takers are unable to attempt all items within the allotted time (Bejar, 1985). Speededness can result in inaccurate reliability estimates and may weaken validity claims by potentially introducing construct irrelevant variance, test bias, and construct underrepresentation (Lu & Sireci, 2007). However, given practical constraints for high-stakes standardized tests, some speededness is inevitable. Therefore, the goal of the MET shortening project was to avoid undue test speededness.

Educational measurement literature provides methods to account for and estimate speededness (Oshima, 1994; Bolt et al., 2002) and even precisely models time limits for unspeeded tests (van der Linden, 2011), but the language assessment literature appears to provide little guidance beyond mentioning seat time as an important consideration in designing tasks and planning administrations.

In order to avoid undue speededness on the shortened MET grammar and reading section, a study was conducted to determine the time threshold for target test takers to finish the section unspeeded. Using response-time data from trial computer-based administrations, two seat times were chosen to be trialed: 60 minutes and 75 minutes. The two seat times for the 50-item grammar and reading section were trialed with a single administration approach on a sample of the test taker population (n = 929). Methods for determining speededness of either seat time include analyses from classical test theory and Rasch theory, as well as a number of speededness indices collected by Lu & Sireci (2007).

In addition, results from a small internal administration with high-proficiency English users and responses from test-taker and proctor questionnaires were triangulated with the quantitative results to determine an appropriate seat time. Considering the shortage of accounts detailing how to minimize language test speededness, this poster aims to share with the language assessment community one methodological approach for determining an appropriate seat time for a high-stakes standardized language test.

Who Makes the Call? A Contested Stakeholder Narrative on the Development of Rater Training Materials
Mark Chapman, WIDA, USA
David MacGregor, WIDA, USA
Carsten Wilmes, WIDA, USA

Developers of large-scale assessments often have multiple competing stakeholder voices vying for attention. During the design and development of the assessment and supporting ancillary materials, these different voices can shape the creation of content and influence how that content is disseminated. This poster reports on the development of a speaking rater training policy (Brown, 2003; Fulcher, 2003; Luoma, 2004) and describes how different groups of stakeholders attempted to shape the narrative of the content and implementation of the new training materials.

In 2015, the WIDA Consortium’s ACCESS for ELLs 2.0 (hereafter ACCESS), a large-scale English language proficiency test administered to students in K-12 public schools in 39 states across the US, was delivered online for the first time, after having been a paper and pencil test for several years. Although many states made the shift to testing online, a large
number of school districts opted to retain a paper administration mode, resulting in potential issues of score comparability between students who take the test online and those who take the paper test. These issues of score comparability have become high-priority concerns for the Technical Advisory Committee (TAC) that oversees ACCESS test program psychometric and policy decisions, as score comparability issues constitute a threat to test validity (Bachman and Palmer, 2010). A specific directive from the TAC was to ensure that training materials provided to raters for the paper and online speaking tests are parallel, both in scope and content. However, implementation of this directive has proven challenging given the different groups of raters who score the paper and online speaking tests. The online speaking test is scored by a central pool of raters who work for a scoring vendor, while the paper speaking test is administered and scored by local raters who are often language teachers.

The supplementary training materials provided to local speaking raters were reviewed by several state education agencies. Reviewers were generally appreciative of the supplementary training materials; however, several expressed concern about the burden imposed from an expanded rater monitoring policy. The supplementary materials provided specific guidance on when, how, and who should be monitored, along with a rater reliability calculator and a guide explaining how to use the calculator. States reported a range of difficulties they expected to face implementing this guidance, with specific concerns raised over the monitoring of specific percentages of speaking tests. In response to this feedback, WIDA revised how these supplementary materials would be disseminated. Rather than release the supplementary materials via the online training course used directly by local raters, states have requested that they be given local control over the materials and coordinate a rater monitoring policy consistent with state and district resourcing and priorities.

The poster describes this contested narrative, highlighting the assessment context and the central issue of change driven by technological enhancement and stakeholders’ responses to such changes. The poster describes different stakeholder perspectives on program needs and how those differences are resolved between test developers, advisory committees, state education agencies, and educators.

Managing Change: Developing an English Proficiency Certification System for Graduation in Cuban Higher Education
Ivonne de la Caridad Collada Peña, University of Informatics Sciences, Cuba
Claudia Harsch, University of Bremen, Germany
Pedro Castro Alvarez, University of Informatics Sciences, Cuba
Santiago Jorge Rivera Pérez, Ministerio de Educación Superior, Cuba

Acknowledging the importance of English as international and academic language, the Ministry of Higher Education (MES) in Cuba is currently promoting a paradigmatic change in the teaching and learning of the language, including the adaptation of the Common European Framework of Reference (CEFR) as internationally recognized proficiency framework. In 2015, the Ministry of Higher Education released a new policy for undergraduate students, requiring English proficiency at CEFR level B1+ as a prerequisite for graduating at Cuban universities. A teaching and certification system is needed to ensure that students acquire the necessary level of proficiency and that Cuban language centres can reliably and validly certify their students’ proficiency at the transition point of graduation.

We will report on setting up a sustainable system for teacher training and test development, which is accompanied by plans for research studies. Partners in the endeavor are the Ministerio de Educación Superior de Cuba, the Universidad de las Ciencias Informáticas, the British Council Cuba and UK, and the University of Bremen, Germany. Previous capacity building projects provided by the Office for University Cooperation of the Flemish Interuniversity Council and supported by ILTA, as well as other initiatives have also contributed to the initial preparation of teachers from different Cuban universities.

The changes introduced by the Ministry have created the need for teacher capacity building with regard to adapting and implementing the CEFR for the Cuban context. Furthermore, language testing and assessment requires immediate attention, since practical experience as well as theoretical underpinning is needed in order to develop such a certification system. As part of this process, and specifically for approaching testing and assessment, a Cuban network of teachers has been created with the support of ILTA. Teachers in this network are to be trained to gain expertise in all assessment-related matters. They are not only designated to work on the development of the new certification system, but they will also act as trainers in their local regions, to enhance assessment literacy amongst their colleagues.

The approach taken in the project is collaborative and dynamic, with face-to-face training phases where international trainers from the University of Bremen and the British Council UK visit Cuba, and online working phases where the teachers collaboratively develop assessment materials and receive feedback from each other and from the
international trainers. We are planning an online collaborative space for teachers and trainers; meanwhile the team is working via email and other means of communication. This is particularly relevant in a context where traveling and face-to-face meetings are challenging to arrange. Hence, we will also focus on the role of national and international networking, on our endeavors to initiate and maintain a sustainable training and development system, and on the role of online collaboration in this process.

Overall, the project has the following long-term objectives:

- Developing a certification system for English in Cuban higher education.
- Developing assessment literacy amongst language teachers in Cuban HE.
- Pursuing a joint research program accompanying the development of assessment literacy and the certification system.

**Towards Improved Clinical Communication: An Overview of the Revised Version of the Occupational English Test (OET)**

Gad S. Lim, Cambridge Assessment English, UK
Brigita Seguis, Cambridge Assessment English, UK
Barbara Ying Zhang, Cambridge Boxhill Language Assessment, Australia

In an age when a fair proportion of the world’s population lives outside their country of birth, the issue of migration continues to occupy a prime position on the political, educational and economic agendas across the globe. In most of the developed world there is a rising trend of migration for employment purposes, which means that having relevant migrant selection mechanisms in place is both crucial and imperative.

One area where such mechanisms are of paramount importance is healthcare. While proof of relevant professional qualifications is the first step in ensuring that a health professional can safely practice medicine in a given country, the second crucial step is evidence of a high level of language proficiency, which can only be assessed by relevant language tests.

The Occupational English Test (OET) is an international language test specifically designed to assess the language communication skills of healthcare professionals who seek to register and practise in an English-speaking environment. While originally developed in the 1980s, OET has been subject to continuous research, validation and evaluation to ensure it reflects the demands from a growing range of stakeholders and remains relevant in a constantly changing healthcare environment. The aim of this poster presentation is to provide an overview of planned changes of the test, to give an outline of some of the studies that informed the revision process, and explain the rationale behind the changes that were implemented.

The revised test, due to be launched in late 2018, is based on the findings from a number of research projects (e.g. Lim 2016; Knoch et al. 2017) and will incorporate several improvements, such as the introduction of indigenous criteria for the writing and speaking tests. Other ways of broadening construct coverage include the addition of professional-to-professional interaction (alongside professional-to-patient interaction) in the listening test, which will allow to test the understanding of more complex ideas and arguments at C1 level. Likewise, texts from broader healthcare-relevant genres will be included in the reading section to target understanding at the higher proficiency level. The changes will also enable the use of the Rasch model in scoring the test.

In addition to enhancing the test’s content, the revised test will also introduce improved more granular way of reporting test outcomes to accommodate a greater variety of test uses. The presentation will explain the rationale behind the changes implemented and the introduction of a new scoring scale, as well as consider the impact of these modifications on the test stakeholders.

References:
Language Assessment Education in a Globalized Era - A Cross-cultural Comparative Study on Language Testing and Assessment Courses
Wenjing Yao, University of Auckland, New Zealand

China and New Zealand, located in opposite hemispheres, vary across a range of dimensions, one of which is the orientation of assessment in its educational system. The former is known for an exam-oriented and highly centralized educational system whereas the latter, like most of its western counterparts, features a less exam-dominated educational environment. Against this backdrop, this study probes into the post-graduate level language testing and assessment courses (LTAC) during the preparation of language teachers in the contexts of New Zealand and China. Its intention is to identify the similarities and differences between these language assessment courses, understand how they are constructed and implemented, and investigate how they foster and establish assessment literacy for language teachers within disparate social-cultural contexts and educational environments. Altogether, ten semi-structured interviews were conducted with course lecturers (four from New Zealand and six from China), and interview data were analysed with inductive thematic analysis. Overall, despite the disparate social-cultural educational environments, a great number of similarities were identified in this preparation process. When designing LTAC for assessment preparation: 1) instructors in both nations contextualized the course by taking students’ prior assessment experience and possible future career into consideration; 2) instructors laid a strong emphasis on classroom-based assessment and formative assessment; 3) instructors hold similar definition of assessment literacy, which was set as core competence for language teachers. When delivering the courses instructors share the following similarities: 1) a large majority of instructors adopt a practice-based approach to teach the seemingly complex assessment theory and principles; 2) They preach what they teach by delivering a variety of assessment tasks, 3) though instructors advocate classroom and formative assessment, language testing still remains a large proportion of the course---for one thing, testing remained the main teaching content; for another, most of the assessment tasks were mainly on tests development or evaluation rather than formative assessment procedures; 4) instructors were challenged by restricted time to teach and students’ limited ability either to think critically or to teach reflectively. Nevertheless, there are also some differences identified: 1) NZ instructors were facing increasing number of international students from multi-lingual and multi-cultural background (though most of them were from Asian countries) while they were learning in an English-speaking country, whereas, Chinese instructors teach in a monocultural context where English is a foreign language; 2) NZ instructors placed a stronger focus on national policy of assessment than their Chinese counterparts and they relied on policy as an assessment yardstick; 3) NZ instructors were equipped and facilitated with on-line teaching technology—their course was a combination of on-line and on-site learning for students; 4) Chinese instructors who emphasized on non-testing assessment were making a compromise with the national dominate testing-culture by teaching students formative use of tests, which was not the case for NZ instructors.

A Re-envisioning and Re-validation of an Assessment Framework for a Large-scale Language Proficiency Test: Participation and Impact of Stakeholders
Margo Gottlieb, WIDA, USA
Mariana Castro, WIDA, USA
Tim Boals, WIDA, USA

Summative English language proficiency testing for English language learners (ELLs), those students for whom English is an additional language, has been mandated in the United States as part of federal legislation since 2002. For over a decade, ACCESS for ELLs (ACCESS) has been the large-scale secure K-12 English language proficiency test of the WIDA consortium, a collective of 39 states and entities devoted to the linguistic and academic advancement of ELLs. It is the primary measure in the U.S. for determining ELLs’ annual growth in their English language development as well as serving as a criterion for student reclassification from language support and as an indicator for state-level accountability. Historically ACCESS has been anchored in and aligned with its K-12 English language development standards. The principal validity claim for the test has rested on coverage of WIDA’s five English language development standards with its four language domains (listening, speaking, reading, writing), and its six levels of English language proficiency. Its validity use argument, based the test scores of over a million ELLs each year, has centered on the usefulness of these data for multiple stakeholders across the educational spectrum (Westerlund, 2014; Molle, 2013). Since 2015 with the inauguration of ACCESS 2.0 on-line and paper forms, WIDA has been re-envisioning its standards framework to better represent and capture current language theory, the language development process and their
translation into practice. As a result, changes in its summative assessment framework will occur that affect how the test is conceptualized and operationalized in the future. The major questions that the consortium faces are, 1. “How does WIIDA maintain the validity of test score inferences from ACCESS as it undergoes change in the representation of its standards and assessment frameworks?” and 2. “How are stakeholders involved in this change process and how do they contribute to the continuity of ACCESS test score interpretation?” This poster presentation depicts the role of multiple stakeholder groups, including national language and content experts, teachers, administrators, state educational agencies, and test developers, in the development and re-validation of the WIDA standards and assessment frameworks. It illustrates how research centered on stakeholder participation at strategic points in the project has helped enhance the assessment framework as it moves to its next iteration. A timetable with key events and review points will graphically depict stakeholder involvement; additionally, it will include the incremental decisions made along the way by each stakeholder group. Specifically, the poster will highlight changes in the WIDA standards framework and their direct impact on the assessment framework within the context of stakeholder involvement. Based on extensive stakeholder input and feedback on the underlying theoretical rationale and components of the assessment framework, it is anticipated that the changes will be fully integrated into WIIDA, institutional-wide, and ACCESS 2.0 specifications, test design and delivery at the onset of the upcoming decade.

Assessing Comprehension of Conversational Implicature: Comparing Two Item Formats
Stephen P. O’Connell, University of Maryland, College Park, USA

The ability to understand conversational implicature is often linked to the upper levels of language proficiency frameworks (e.g., the CEFR, ACTFL), but how understanding implicature is assessed using multiple-choice items is an under-researched area—despite how common the practice is. Building on work from the philosophy of language (Grice, 1975) and pragmatics (Sperber & Wilson, 2013), researchers such as Bouton (1988, 1994, 1999), Roever (2006, 2013), and Taguchi (2005, 2008, 2009) have made significant contributions to understanding how different types of implicature result in different levels of difficulty for learners of English. But whether the ability to comprehend conversational implicature can be separated from general listening comprehension, and whether that ability is truly “higher level” are ongoing questions.

The effect of inferencing on item difficulty has been investigated for some time (e.g., Freedle & Kostin, 1996; Kostin, 2004), but while many of these studies have found that inferencing increases difficulty, they have typically been conducted with existing test forms, rather than with bespoke tests of implicature (as with Bouton, 1988). This may be one reason why questions remain about implicature (or general inferencing) as the source of item difficulty and whether understanding implicature has psychological reality as a separate subskill. This study attempts to address these questions by analyzing the results of a 60-item listening comprehension test that was administered to 255 non-native speakers of English whose proficiency ranged from low intermediate to low advanced (i.e., B1 to C1 on the CEFR). 30 of the test’s items were designed to assess conversational implicature (as conveyed in short dialogues) and 30 were designed to assess global non-implicature understanding. Details on the development of the test (which followed standardized-testing best practices) will be provided.

All 60 items were administered in multiple-choice and short constructed-response formats and scored using the Rasch partial credit model to allow for scoring the dichotomous (multiple-choice) and polytomous (constructed-response) data simultaneously. The comparability of item difficulty across the two formats and the relation between the two sub-skills will be discussed, with a focus on the fact that the subset of participants who were independently measured as CEFR C1 language users did perform in a significantly different manner (as indicated by logistic regression) from CEFR B1 and B2—although it appears that higher general language proficiency, rather than greater ability with implicature was the determining factor in this difference.

In addition, qualitative analyses of participants’ constructed responses will be summarized, focusing on the patterns of interpretation of the tested implicature that the participants made. Specifically, the results from this study appear to provide evidence that high intermediate and low advanced test-takers can provide the keys that test developers expect, thus rebutting a common critique of the multiple-choice method of testing, particularly when multiple-choice items are designed to assess implicature.

The broader implications of the validity of using either a multiple-choice or constructed-response format to assess comprehension of conversational implicature will also be discussed.
Mixed-methods Development of a Test-taker-oriented Writing Rating Scale
Beverly Baker, University of Ottawa, Canada
Maryam Homayounzadeh, Shiraz University, Iran
Angel Arias, University of Ottawa, Canada

Most high-stakes language proficiency tests provide supplementary materials to help candidates to familiarise themselves with the structure and content of the test. With respect to performance assessment tasks such as writing, transparency is enhanced with the provision of a test-taker-oriented rating scale to clearly communicate the criteria of quality performance. As Messick (1994) reminds us, “…not only should students know what is being assessed, but the criteria and standards that constitute good performance should be clear in terms of how the performance is to be scored…” (p. 16). Such a rating scale should ideally be distinct from the one used by raters themselves. As Knoch (2009, 2011) suggests, any rubric designed for rater use in proficiency or placement decisions might not be appropriate for a test-taker audience for self-assessment activities.

However, very little research has been done on the use of test-taker-oriented rating scales for self-assessment related to test preparation. The study reported here presents a mixed methods approach to the development of a test-taker-oriented rubric for the writing section of a high-stakes proficiency test. A first draft of the scale was created with reference to the original scale as well as to guidelines on the creation of diagnostically oriented scales (Knoch 2009, 2011). The draft was reviewed by the original developer of the scale as well as an experienced operational grader, for advice on whether the same criteria were being addressed with the same approximate weighting. It was also reviewed by multiple test takers for feedback on comprehensibility of language and perceived usefulness. Next, approximately 100 previously graded essays at varying score levels were selected and some content of the essays was extracted in order to provide concrete examples of “clichés,” “awkward language,” and other elements mentioned in the scale. This was done to enhance the diagnostic properties of the scale.

Finally, these essays were graded by approximately 20 student volunteers who represent the test taker population. K-Means clustering was applied to the data to explore whether the rater oriented and test taker oriented ratings differ (i.e., a single cluster implying that the two rating scales were used in a similar way, and more than one cluster implying that the rating scales were used in a different way).

This study responds to the theme of this conference, as this test is used for high-stakes mobility purposes: credentialing requirements for internationally trained professionals as well as university admissions for international students. It is therefore incumbent upon us to be transparent about expectations and to provide useful tools to support test preparation.

Using Response Time Data to Understand Test Performance for Adolescent Learners of English as a Foreign Language
Venessa Manna, Educational Testing Service, USA
Hanwook Yoo, Educational Testing Service, USA

The use of test-taker response processes to aid in score interpretation has recently taken on increased importance in the field of assessment and measurement due to the spread of computer-delivered assessments. Ercikan & Pellegrino (2017, p.2) define test taker response processes as “the strategies and behaviors of test takers when they interpret assessment situations and establish the solutions as presented through verbalizations, eye movements, response times, or computer clicks”. Response process data allows investigation of the processes that test-takers follow and the outcomes they produce, thus providing insight into observed test scores within and across test-taker groups. This study utilized test-takers’ response times (RT) to gain insights into the test performance of two groups of adolescent learners of English as a Foreign Language (EFL). The groups studied were 13-15 year olds (middle school) and 16-18 year olds (high school). Specifically, we addressed the following research questions:

1: What are the response time patterns across the four language modalities (reading, listening, speaking and writing) for two groups of adolescent learners of EFL?
2: What is the relationship between RT patterns and test performance across two groups of adolescent learners of EFL?

As second language acquisition is influenced by the learning environment, the study is limited to performance of English language learners in middle school and high school in Japan. Specifically, this study used data from an administration of the TOEFL Junior® Comprehensive test (TJC) in Japan in the fall of 2015. The TJC test is not based on any specific curriculum and measures the academic and everyday English skills of students ages 11 and older in non-English-speaking countries (Educational Testing Service, 2015). The test consists of four sections: Reading, Listening, Speaking, and Writing.
There are two phases to the analyses. First, descriptive statistics of RTs were reviewed by test section in an effort to better understand the patterns in RT across the test-taker groups. Second, investigation into the relationship between RT and test taker proficiency across the subgroups was conducted using an item-level psychometric model that allows for joint estimation of speed and item parameters using both RTs and response data (Fox, Klein Entink, & van der Linden, 2007). The R-package LNIRT, was used to estimate the joint RT model parameters. Overall, results indicate that average RT is similar for the two groups, but there is variation by item. This is also supported by the similarity across groups with respect to the top five items with the shortest and longest RT. The results also suggest that RTs were not necessarily indicative of item difficulty.

References

Assessment of Medical Terminology
Simon David Boynton, University of Hong Kong, China
Laura Wakeland, University of Hong Kong, China
Michelle Reyes Raquel, University of Hong Kong, China

Most medical textbooks teach and assess medical terminology knowledge using ‘a body systems approach’ where vocabulary is categorized and presented according to the major systems of the human body (e.g., the cardiovascular system, the digestive system). Although this gives medical students a comprehensive and medically-oriented coverage of medical terminology, it relies heavily on memory work and does not equip students with the skills to comprehend medical vocabulary that they have never encountered before. A ‘linguistic approach’, on the other hand, relies on lexical analysis where students are taught medical terms through an understanding of the internal structure of medical terms (i.e., morphological structure of medical terms). The approach is strategy-oriented and encourages inductive and interactive learning.

This presentation introduces the test designed to assess the medical terminology knowledge of undergraduate medical students following this linguistic approach. This test aims to (1) assess students’ knowledge of words in the word part list given in the English course, and (2) determine if they can transform medical terminology to non-specialist term and vice versa by breaking down medical terminology to its linguistic components (i.e., root, prefix, suffix). The test given to all students has six sections where students are asked to answer 60 discrete items or text-based items. Test items require students to either provide or define a medical word, provide the correct grammatical form of the words, and identify features of pronunciation such as problematic sounds, syllable, and stress patterns. The test is administered to four groups of medical students divided according to their medical specialization (Nursing, Chinese Medicine, Biomedical Sciences, Pharmacy). Rasch analysis of results show how different groups of students performed on the test, as well as items that students generally found easy and/or difficult. To determine how students answer items, a small group of students were asked to answer a shorter test with different items, and asked to think-aloud while they are doing the test. Students were also interviewed as to their perceptions of this linguistic approach to learning and testing medical terminology.

Using Survival Analysis to Investigate Test Performance for Learners of English as a Foreign Language
Lora F. Monfils, Educational Testing Service, USA
Venessa F. Manna, Educational Testing Service, USA

Educational and psychological assessments are typically designed for a target population defined by one or more overarching characteristics, but there is often considerable diversity within the target population that may encompass test takers from diverse demographic, socio-cultural, and educational backgrounds. Moreover, within a given target population, there may be variation in test taking patterns reflecting test use and associated stakes. This study investigated the use of survival analysis (SA) as a means to gain insight into factors associated with achieving designated score criteria on a test of English as a foreign language. Specifically, this study addressed the following research questions:
1. What does SA reveal about the patterns in time to passing, that is, to achieving a designated score level?  
2. What does SA reveal about the relationship between test-taker characteristics and pass/fail patterns?  

To address these questions, this study used data from three administrations years (2014-2016) of the Test of English as a Foreign Language™ (TOEFL® iBT), developed by Educational Testing Service. Designed to measure the ability to use and understand English at the university level, the test consists of four sections: Reading, Listening, Speaking, and Writing. Scores are reported for each section and for the total test. Two levels of passing were defined as reaching CEFR levels B2 and C1 on the total score. 

First, test taking histories were reviewed overall and by demographic subgroups to better understand patterns related to number of times testing, time between retests, and associated score changes for repeaters. Second, SA was conducted using the Cox proportional-hazards regression model to model time to passing at each level as a function of test-taker characteristics, including age, gender, native language (Indo-European or Non-Indo-European), exposure to English (time studying English, time spent in a content class taught in English, time living in a country where English is the main language), reason for testing, and repeater status.

Preliminary results indicated that 81% of the test takers in our sample tested one time in the study period. At first testing, 74% scored at or above B2, and 37% scored at C1. For those who did not score at or above the B2 or C1 levels at first testing, the likelihood of achieving the designated score levels increased with repeated testing, and decreased mildly with age, and both effects are somewhat greater for the C1 criterion in comparison to B2. The covariates with the largest positive association with scoring at both the B2 and C1 levels were language type, time studying English, and taking TOEFL for graduate or post-graduate admission. The covariates with the largest negative association were taking TOEFL for admission to secondary school or admission to a 2-year college. Although more study is needed, these preliminary results are promising in terms of using SA to add to our understanding of patterns of test performance for repeat test takers, in general and in particular as pertains to the international EFL context.

**Faking Low Proficiency for Legal Advantage: Is ‘Standardized’ Testing Possible?**

Margaret van Naerssen, Independent Consultant, USA

A common question in court involving non-native speakers is “How do you know the defendant wasn’t faking a lower language proficiency to avoid possible legal charges?” This concept paper is focused on a performance phenomenon of intentionally underperforming one’s skills in order to deceive. In forensic contexts the deception might be for a legal advantage regarding the admissibility of evidence. This performance has popularly been called “faking” or “malingering”. In this paper, after an introduction to the phenomenon referred to as faking, the presenter asks the audience to imagine being challenged in court on whether the possibility of faking had been considered.

Migrating populations are moving into different legal systems, sometimes encountering serious legal issues, compounded by language and cultural factors. Also, local populations might be suspicious of these newcomers. Thus, linguists/Language Assessment Experts (LAES) are asked, in some legal systems, to become more involved in doing language assessments, linking proficiency to legal questions, and testifying in court. Like migrating populations, some linguists/LAEs might find this new “country” messy and unpredictable, but also ripe with interesting challenges. One challenge linguists/LAEs face, when testifying in court about the language proficiency of non-native speakers, is the possibility of faking.

Sociolinguist Roger Shuy has urged linguists in forensic settings to always look at both sides of the legal question when examining language evidence. In cases involving non-native speakers, one might be looking at whether a defendant was performing at his/her truthful level of proficiency. Two linguists/ Language Assessment Experts (LAES), Eggington and van Naerssen, have independently been refining its label, its scope, and possible strategies for detecting faking of language proficiency in forensic contexts.

Faking is a matter of individual performance involving modification of one’s communication, not a single definable linguistic feature. Five factors are briefly mentioned that might contribute to successful modification of one’s language for workplace advancement, artistic performance, and even espionage. Several strategies are then briefly described for possible detection of faking in individual legal cases.

Once in court, linguists/LAEs need to know the court rules about experts testifying and about admissibility of evidence. In US courts (and maybe in others), then come the challenges of claims, by an opposing attorney, including intentional traps, frequently grounded in false assumptions. Participants are encouraged to imagine implications when the questioner can try to limit responses to Yes/No. (And is there a way out?)

Did you consider that the defendant might be faking? Can you, with absolute certainty, detect, faking?
Is there research on faking?
Was the language test you used designed to detect faking?
Did you talk with the test development center about modifying their test for this case?
Is it possible to develop a “standardized” test for detecting faking?
If not, then “wouldn’t you agree that your assessments are not scientific and therefore, shouldn’t be admitted in court?!’

Self-assessment (SA) of Foreign Language (FL) Literacy Abilities among Adolescent Chinese Learners of English
Huan Liu, Washington University in St. Louis, USA
Cindy Brantmeier, Washington University in St. Louis, USA

Reading and writing in a foreign language is not easy for young learners as both involve complex cognitive and social processes (Koda, 2005; Berndardt, 2010; Grabe & Kaplan, 1996; Silva & Matsuda, 2001). Young learners have to make efforts to internalize these processes while at the same time “deal with the cognitive demands of early literacy, either in their first or the target language, or in both” (Weigle, 2002). Considering those intricacies, assessment for young learners needs to highlight assessment for learning. Assessment should be a tool to build up their ability, monitor learning process, and promote positive attitudes to learning (Weigle, 2002). Self-assessment (SA), “procedures by which learners themselves evaluate their language skills and knowledge” (Bailey, 1998), serves this goal by raising self-awareness of learning (Oscarson, 1997), promoting learning autonomy (Bulter & Lee, 2010), and engaging learners to self-assess in an interactive and low-anxiety way (Bachman & Palmer, 1996; Brantmeier, 2002). To date, few studies focused on SA among young learners, particularly in English learning context in China where high-stake exams dominate. The study attempts to substantiate current research by exploring how adolescent Chinese EFLs self-assess their reading and writing abilities, and the relationship between SA scores and subsequent literacy performance. 106 adolescent Chinese EFLs (ages 12-14) at 7th Grade in a middle school in China were recruited. Participants completed the following instruments: SA Reading Questionnaire (11 items, Cronbach’s α = .80), SA Writing Questionnaire (14 items, Cronbach’s α = .88), Reading Comprehension, Writing Task, and a Demographic Questionnaire. SA Questionnaires were criterion-referenced and asked participants to indicate how they would rate their English reading and writing ability in each situation presented in each item (5-point Likert Scale). Reading Comprehension was consisted of two passages, with each followed by three task types: Free Recall, Multiple Choice, and Sentence Completion. Writing Task asked participants to write a story about a picture presented. Jacobs et al.’ (1981) criterion was used for rating by two trained raters (Cronbach’s α = .90)

R Software was used for data analysis. Spearman’s rho (ρ) correlation analyses indicated that SA reading was significantly related to free recall (ρ = .41, p < .0001), sentence completion (ρ = .46, p < .0001), and multiple choice (ρ = .43, p < .0001). The correlation between overall reading comprehension and SA reading was also significant (ρ = .51, p < .0001) with large effect size (Cohen, 1988). Pearson correlation revealed a significant relation between writing production and SA writing (r = .30, p < .01) with medium effect size. Independent sample t-test showed that, for reading ability, participants self-rated their ability to make predictions and ability to make connections between the texts significantly lower. For writing ability, participants self-rated their ability to keep readers in mind, ability to write an opening paragraph, and ability to use metaphors significantly lower. The study concluded that SA could be used as a tool for adolescent Chinese EFLs to self-identify their strengths and weaknesses in language learning. Pedagogical implications were discussed.

Developing and Validating Tasks to Assess English Core Competencies and Tracking Students’ Development in Senior High School
Shaoqian Luo, Beijing Normal University, China
Yumei Zhang, Beijing Normal University, China

The newly revised National Senior-high-school English Curriculum Standards (SECS) in China is in response to the whole-person educational goal proposed by the Ministry of Education (MOE, 2014). The primary aim for the revision of SECS is to construct core competencies that students are supposed to acquire from English learning. The competence-based SECS includes four competencies: language competence, cultural awareness, thinking capacity and learning ability (Cheng, 2017; MOE, 2018).
At present, the discussion on how to teach and assess core competencies has become nation-wide among policy makers, researchers and classroom teachers. Tests, especially those high-stake ones, have been playing a powerful role in the implementation of educational initiations. However, it remains as a challenge to assess a student’s cultural, thinking or learning ability especially when it comes to a request for standardization. It seems like testing the untestable (Paran & Sercu, 2011).

This current research aims to develop and validate tasks for assessing students’ core competencies, which constitutes part of a larger longitudinal study[1] on senior high school students’ English development of core competencies. The poster mainly focuses on the initial validation results for one prototype task (a reading circle task) developed. In the research, two experts in language teaching and assessment and five senior high school English teachers were asked to make judgements about the competencies that could be elicited from each item. A consensus of the constructs and the scoring rubric was then reached through reexamination and discussion on the definitions of the competencies and a pilot scaling of students’ responses. Then responses from ninety target test takers were rated for a quantitative analysis on the task validity. The results yield promising validity of this task in testing students’ core competencies and provide preliminary empirical insights into the identification of the core competencies as distinct test constructs and their interaction in task completion.

Furthermore, the design of the follow-up longitudinal study will be reported. The study plans to track about 300 students from different districts in Beijing across two years using the tasks validated and classroom observations. The data collected will be analyzed with ANOVA and multi-group growth curve modeling. The findings aim to unveil features of students’ development of core competencies and thereby shed empirical light on the rationality of the descriptors for different levels of the four core competencies prescribed in the curriculum standard.

[1] The research is funded by the Beijing government: Beijing’s 13th five-year plan key project (CADA17076).

Linking the Global Test of English Communication (GTEC) to CEFR Levels
Masashi Negishi, Tokyo University of Foreign Studies, Japan
Yoji Kudo, Tamagawa University, Japan
Yasuko Okabe, Center for Entrance Examination Standardization, Japan
Yuko Kashimada, Benesse Corporation, Japan
Mika Hama, Benesse Corporation, Japan
Yuko Umakoshi, Benesse Corporation, Japan

The Japanese government is moving towards replacing current college English entrance exams with 4-skill English examinations developed by private organizations. They have already put forward the required conditions that these 4-skill English examinations must meet, one of these being that the exam should be aligned with CEFR levels and that the organizations creating these tests must reveal how exactly they were able to achieve this. This session focuses on how the makers of one of these 4-skill tests known as the Global Test for English Communication (GTEC) conducted their standard setting when aligning the GTEC score system with CEFR levels.

The GTEC is a 4-skill English test developed by Benesse Corporation designed to assess English communicative abilities, of which its test-takers are comprised mainly of high school level students, with many junior high school students taking the test as well. The GTEC is a paper-and-pencil based test for assessing listening, reading and writing skills and uses a tablet PC to assess speaking skills. The test was developed using Item Response Theory, or IRT for short. The score range of each section of the GTEC ranges from 0 to 320. The four sections are added to create the GTEC total score, which exists on a range from 0 to 1280. There are about 1,000,000 test-takers each year. The standard-setting for the GTEC was conducted in 2016 by a six-expert-panel—each expert with a diverse linguistic background and sufficient knowledge of CEFR. Members of this panel were also joined by some of the item developers for the GTEC where relevant, in order to assist with the overall process. The data used for this standard-setting was derived from actual answers to test items by test-takers from operational GTEC tests administered in 2016. The method used for standard-setting was a “contrasting-group method” (Council of Europe, 2009), for speaking and writing. By using IRT data, the test-takers’ answers were ranked according to performance and score, after which each member of the panel analyzed the data in order to determine where to assign the corresponding cut-scores of each CEFR level from A1 through C1. There were three steps in deciding cut-scores:
1. It was confirmed that each member of the panel had the same understanding of the English ability of people at the border between two CEFR levels.
2. The panel was divided into three groups of two experts where each group had a discussion to reach an agreement on cut-scores. Each two-person group received the additional company of one of Benesse Corporation’s item developers, in order for the item developers to provide their support or assistance when necessary.
3. All members of the panel then had an additional discussion to reach a final agreement.
In conclusion, the method and process worked well for determining the GTEC’s CEFR level alignment. There were some discrepancies in judgements between groups, but the panel was able to reach a final agreement by referring to and discussing those performances which stood near the cut-scores.

The Test-takers’ Perception: An Investigation of the Source-text Influence on Integrated Writing Performance
Youxia Sun, East China Normal University, China
Xiaoxian Guan, East China Normal University, China

It is probably uncontroversial that learning to write effectively from external sources is an important academic skill. Recognizing this fact, major standardized English language exams (e.g., TOEFL iBT, PTE-Academic, NMET, DELNA) have been designed to assess writing performance via sourced-based writing tasks. However, there has been relatively little research on exploring learner performance on source-based writing tasks in relation to the individual and external factors that affect performance (Delaney 2008). Previous studies, though being few in number, have identified prior knowledge and experience influential to learners’ source-based writing performance (see Shi 2012; Cumming et al. 2016). They yet have been limited in the examination of prior knowledge and experience (such as learners’ content knowledge, disciplinary background, and educational experience), with no specific predications concerning learners’ perception of source texts. Given a recent change of writing tasks in National Matriculation English Test (NMET) in some provinces of China, it is imperative to measure the possible source-text influence to better inform task design and writing instruction. The present study, therefore, offers an approach that attempts to investigate the influence of test-takers’ perception of source texts on their source-based writing performance in the context of a reading-into-writing task.

The data collected for the present study included an argumentative reading-into-writing writing test and an expository independent writing test that 371 Chinese undergraduate students simultaneously took, as well as the scores rated analytically by experienced raters. The independent writing scores were taken as measure of test-takers’ writing abilities, while the integrated writing scores as measure of test-takers’ reading-into-writing abilities. The design of test-takers’ perception of source texts was informed by these students’ questionnaire responses surveyed directly after the two writing tests. The source-text perception measure consisted of ten individual characteristics that stemmed from previous studies (see Polio & Glew 1996; Powers & Fowles 1998; Skehan 1998; Lee 2008; He & Shi 2012; Cho, Rijmen & Novák 2013) and were tailored by pilot studies. Three latent source-text characteristics were extracted from the ten variables by factor analysis, i.e., source-text knowledge (test-takers’ knowledge of the source text), source-text difficulty (perceived difficulty of the source text) and source-text identification (test-takers’ identification with the source text). Substantial multilevel linear modeling (MLM) analyses were conducted to observe the possible relationship among writing ability, source-text characteristics perceived by test takers and their reading-into-writing performance. The obtained MLM results showed that writing ability played an important role in explaining the variation in reading-into-writing abilities, and only source-text knowledge was significantly related to reading-into-writing abilities. Structural equation modeling (SEM) analysis was used to confirm the relationship between test-takers’ source-text perceptions and their writing performance. The SEM results suggested that test-takers’ source-text knowledge had a weak yet positive effect, source-text difficulty a weak yet negative effect, and source-text identification a negligible effect. These results demonstrate the complex interactions between test takers and source texts, and may shed light on the nature of integrated writing.

Strengthening Beginning Chinese Heritage Language Teachers’ Assessment Literacy: Conceptions and Practices
Jiani Yun, University of Auckland, New Zealand
Mary Hill, University of Auckland, New Zealand
Martin East, University of Auckland, New Zealand

Research in the last two decades has explored the mediating factors contributing to teachers’ assessment literacy, such as knowledge and skills within assessment literacy, training efficacy, assessment conceptions and practices, and contextual influences on teachers’ assessment literacy. However, some of factors still need to be further explored with reference to specific research contexts. For example, research into assessment training has found that beginning teachers have limited time to learn assessment during teacher education (Earl & Timperley, 2014; Stiggins, 2002). Furthermore, instead of having deep understandings of the spirit of assessment, many novice teachers have only
superficially learnt practical assessment tools during professional development programs (Earl and Timperley, 2014; Gearhart et al., 2006). It is unlikely that they will be assessment literate, if they do not have the opportunity to reveal their existing conceptions of assessment, address the limited aspects of their knowledge base, and gain feedback on their assessment practice (Hill et al., 2017).

In addition to the assessment development challenges, many teachers of Chinese in NZ secondary schools have Chinese heritage cultural backgrounds with assessment conceptions that may differ from the outcome-based approach to assessment used in NZ secondary schools. Given these cultural differences and the challenge of learning to be assessment literate for the NZ secondary school context, the study proposed here will examine novice Chinese heritage language teachers’ assessment conceptions and practices in the New Zealand secondary school context. A case study approach will be employed. Four novice Chinese language teachers who finished teacher education in 2016 and started to teach Chinese language or English as a second or other language (ESOL) in New Zealand secondary schools will be invited to participate. Their development of assessment conceptions as well as practices before registration (normally two years) will be documented by interviews and classroom observations within their first two working years. Data will be firstly collected at the end of the school academic year in 2017, and then collected in the beginning, middle and end of the academic year of 2018 respectively. All the data will be analysed progressively after each round of data collection. The intention is to gain understandings of how novice Chinese heritage language teachers adapt their existing conceptions and practices within New Zealand’s assessment regime. It is hoped that the findings will inform teacher preparation and professional development for language teachers, especially those from Chinese heritage backgrounds.

Working towards an English Language Test for Young EFL Learners in Shanghai: The Role of Stakeholders

Jufang Kong, Zhejiang Normal University, China
Mingwei Pan, Guangdong University of Foreign Studies, China
Jason Fan, Fudan University, China

Test stakeholders refer to the individuals or groups who are affected by the development and implementation of a test, including, for example, test takers, parents, teachers and test score users. Stakeholders’ perceptions can affect the public acceptance of a test, and by extension, its consequential validity (Messick, 1989). Therefore, it is recommended as good testing practice that stakeholders’ voices be examined and taken into consideration in the process of test development (e.g., AERA, APA & NCME, 2014). Young learners are children who are in primary or elementary school between the ages of approximately five and twelve (McKay, 2006). In Shanghai, China, few English language tests are currently available, targeting young EFL learners. A recent survey of these tests indicates that they are flawed in design with limited validity evidence to support score interpretations and uses (Pan & Kong, 2017). As such, Shanghai Center for Research in English Language Education (SCRELE), a research organization under Shanghai Municipal Education Commission, sets out to develop an English language test known as SCRELE Elementary, targeting young EFL learners in Shanghai.

The present study made a cross-sectional review of the existing primary school exit or secondary school entrance English test papers in China on the basis of a survey of extant literature on the characteristics of young EFL learners in terms of age, interest and motivation, social and personal growth, and cognitive style as well as a thorough examination of the educational documents published in the past decade in Shanghai. Discrepancies were identified among the cognitive and psychological characteristics of young EFL learners, the requirements stated in local educational documents as well as the design of the current English language tests assessing young EFL learners in China. Based on the preliminary findings, SCRELE Elementary was developed. Focused interviews were conducted with 15 experienced English teachers from two local primary schools and 8 English language testing experts respectively. Their insights into the current SCRELE Elementary were collected centering on its test specifications, test task characteristics, test score report and so on.

In the next step, the study will continue to collect the sentiments of other stakeholders concerning SCRELE Elementary, including test-takers, parents and test score users, adopting multiple research methods (e.g., questionnaires, interviews and focus-groups). In addition, a small sample trial of SCRELE Elementary will be carried out and cognitive validity evidence of the test will be tentatively elicited and collected. This study, as work towards an English language test for young EFL learners in Shanghai, is bound to be a significant exploration of an improved design of assessing young EFL learners in the changing context of Shanghai, better equipped with stakeholders’ perceptions.
1) An Exploration of Scenario-based Assessment of Speaking for Argument-building

Jorge Luis Beltran Zuniga, Teachers College, Columbia University, USA

One of the goals of research in L2 assessment is developing tests for which the operational construct actually reflects features of the TLU domain. Numerous efforts have been made in order to better reflect the needs of 21st century language test users, with some approaches to assessment highlighting the potential of technology in reimagining L2 assessment. For instance, game-based assessment (e.g. Attali & Arieli-Attali, 2015) and scenario-based assessment (e.g. Sabatini & O’Reilly, 2013; O’Reilly & Sheehan, 2009) aim to take advantage of the technological features that allow for more authentic representations of language use, which become particularly relevant given the current impact of computer-based systems on education and assessment.

In scenario-based assessment, task sequencing and sampling is contingent upon an overarching goal and theme, which provides an opportunity to contextualize a test in a purposeful, learning-oriented fashion, (Sabatini & O’Reilly, 2013). Thus, this approach represents a promising alternative for the assessment of integrated speaking ability. For example, the use of simulated peers, multiple turn tasks, and branched elicitations through decision-making tasks in alignment to an overarching goal would help authenticate construct representation in semi-direct tests of speaking ability.

When it comes to the coverage of communicative purposes in semi-direct assessment of speaking ability, one communicative function that has been readily explored in the context of writing, but has not been examined in such depth in the context of speaking ability, is the elaboration and defense of an argument. Nonetheless, certain real-life tasks require that language learners display their ability to build and defend an argument, for example, class discussions or debates. In light of this, it was determined that the development and analysis of a scenario-based test that guides students through the argumentation cycle (Song, Deane, Graf, & van Rijn, 2013) could provide useful information in terms of the functionality and measurement qualities of such test, and would help evaluate the effectiveness of the tasks to elicit argument-building language.

Therefore, this paper aims to report preliminary findings on the following issues: a) whether the test successfully elicits argumentative language, b) whether the use of multi-turn items enhances construct representation for the semi-direct test (e.g. whether test takers display understanding of an audience through simulated peers), and c) whether a rubric including an appropriateness component is adequate to score the data. To answer these questions, the following analyses will be conducted: qualitative analysis and Multi-Faceted Rasch Measurement (MFRM). Data collection is to be held in Spring of 2018. In addition to preliminary findings, limitations and possibilities for future research will be discussed.
2) Leveraging Eye Tracking for Gathering Validity Evidence Based on Response Processes
Ruslan Suvorov, University of Hawaii at Manoa, USA

Recent technological advances have created a fertile ground for innovative changes not only in language test development (such as the use of automated essay scoring and automated speech evaluation), but also in validation research. One technology that has spawned considerable interest among language testing researchers in the past few years is eye tracking (Bax, 2013; Bax & Chan, 2016; Brunfaut & McCray, 2015). Eye-tracking technology appears to hold a particularly strong potential for enhancing research on the processes underlying L2 learners’ responses to test items. While the importance of process-oriented approaches to validation, which entail gathering validity evidence based on the analysis of individual response processes, is widely acknowledged (e.g., AERA, APA, NCME, Joint Committee on Standards for Educational & Psychological Testing, 2014; Messick, 1995; Wu & Stone, 2016), few studies have utilized eye tracking for process-oriented validation research.

To address this gap, this work-in-progress presentation will introduce an ongoing research project that leverages eye-tracking technology to examine the types of test-taking strategies that language learners use when responding to 4-option multiple-choice items and 4-option true-false items (which display one option at a time rather than all four options at once) in an L2 academic listening test. Motivated by the author’s recent study (2017, forthcoming), in which L2 learners’ use of test-wiseness strategies were found to contribute to construct-irrelevant variance and have a statistically significant effect on the observed scores for multiple-choice items from the Michigan English Test (MET), this study intends to investigate if a 4-option true-false item format is less conducive to the use of test-wiseness strategies than a 4-option multiple-choice item format.

This study uses the convergence model of the data triangulation design (Creswell & Plano Clark, 2007), which entails gathering and analyzing test score data, eye-movement data, and verbal report data from 40 participants who are non-native speakers of English at a public university in the Pacific region. The data are collected with one participant at a time during a 2-hour individual session in a lab equipped with a remote eye-tracker GazePoint GP3 HD (150 Hz data sampling rate) that is used to record participants’ eye movements while they are taking an academic listening test. Immediately following the test, verbal report data are elicited via cued retrospective reporting using participants’ eye-movement recordings as a stimulus. Verbal report data will be transcribed, coded, and analyzed in NVivo to identify test-taking strategies used by L2 learners. Eye-movement data will be analyzed using visual scanpath analysis and converged with the verbal report data. Inferential statistics will be calculated to determine the extent to which the use of test-wiseness strategies affects observed scores for two item formats in the test. The researcher will seek input from the audience on the approach for analyzing the data to determine (a) the types of test-taking strategies used by L2 learners during their interaction with 4-option multiple-choice items and 4-option true-false items and (b) the extent to which two different item formats are conducive to the use of test-taking strategies that introduce construct-irrelevant variance.

3) Investigating Picture-based Prompt Comparability in an In-house Writing Test
Yali Shi, Zhejiang University, China

For security concerns, using parallel writing prompts across test administrations is common practice in large-scale language tests. Lack of prompt comparability would result in unjustified variances in test performance, which would jeopardize not only validity and reliability but also fairness of the test. Therefore, it is crucial for test developers to gather empirical evidence from different aspects to substantiate the validity argument that the different tasks measure the same intended construct, and the inferences and decisions thereby made are meaningful and impartial. The past three decades have witnessed a number of empirical studies on the comparability of writing prompts used in specific assessment contexts (e.g., Brossell, 1983; Liu, 1997; Schaeffer et al., 2001; O’loughin & Wiggleworth, 2003; Brelend et al., 2004; Lim, 2010; Barkaoui & Kouzu, 2012; Cho et al., 2013), among which those of picture-based prompts are very few (Bae & Lee 2010). The results so far generated have been inconclusive or even contradictory. Furthermore, most studies approached the issue through the perspective of either raw scores alone or together with prompt difficulty judgment by test-takers or experts. While some studies included text analysis on various levels, few took a comprehensive perspective which integrated analysis of scores, text as well as prompt. To address the limitation, the present study aims to conduct a systematic investigation of the comparability of five picture-based prompts rotated across writing subtest administrations of an in-house English language proficiency test in one major university in China. Data comes from 500 test-takers based on the stratified random sampling across five prompts of all those who sat the writing subset, with altogether 1,000 writing scores of double marking by four raters. The instruments used consist of five picture-based prompts, 500 writing scripts, responses given by four raters and 15 test takers to interview questions involving their perceptions of prompt difficulty. A three-pronged approach is taken for
the comparability investigation. First, a prompt analysis is performed not only by an objective angel based on Weigle (2002)’s writing task dimensions, but also by a subjective view in the format of raters’ and test takers’ difficulty judgment. Second, a three-facet Rasch analysis is conducted on writing scores. Third, a systematic text analysis of writing scripts is to be conducted through both human coding and automated tools in terms of grammatical, discourse and sociolinguistic features, in order to examine whether the prompts elicit comparable responses. Prompt analysis reveals that the five prompts are comparable along a series of dimensions except subject matter and visual stimulus features. Score analysis results show that the five prompts exhibit significantly differential difficulty, the effect of which on scores is nevertheless not large enough to have practical significance. Both coding of raters’ and test takers’ interview data and text analysis of writing scripts are being processed. The results of the study will provide prompt comparability evidence from the above three aspects for the validity, reliability and fairness of the writing subtest, and meanwhile will shed light on picture-based prompt design and test development of EFL writing.

4) Towards Change in Language Assessment Literacy Development: Designing an Instrument to Assess the LAL of EFL Teacher Educators

Salome Villa Larenas, Lancaster University, UK

Arguably, in times of movement, transition, and change, it is vital that key language assessment stakeholders possess a multidimensional level of language assessment literacy (Pill & Harding, 2012), including language assessment knowledge and skills, awareness of principles that guide good practice, and the ability to place these within broader social and political contexts.

So far, language assessment literacy (LAL) research has primarily focused on the knowledge, skills and practices of language teachers, and a range of other stakeholders (e.g., Hasselgreen et al., 2004; Pill & Harding, 2012). However, to date, no studies have looked into detail in the LAL of those who prepare pre-service teachers for their future profession, i.e. teacher educators. Also, the dominant methodology in existing LAL research has been the survey approach, eliciting perceptions and self-reported knowledge to gain insights into LAL needs (e.g., Berry & O’Sullivan, 2016; Vogt & Tsagari, 2014). While past LAL research has certainly been valuable, scholars have started to call for an understanding of LAL as a more complex construct – beyond needs and self-reported knowledge – intertwining stakeholders’ beliefs with their knowledge base and with contextual constraints (e.g., Levy-Vered & Nasser-Abbu, 2015). At the same time, no validated framework exists which specifies what the LAL knowledge base of stakeholders such as teacher educators should be, nor do there seem to exist validated instruments to measure the LAL of this stakeholder group. Hence, Xu and Brown (2017) have argued for the need to establish LAL constructs, standards and guidelines, to enable the construction of LAL assessment measures.

This Work-in-Progress paper will report on a study aiming to explore the LAL of teacher educators in terms of their theoretical knowledge base, skills and practices. The first stage of the study involved the development of a LAL Framework for teacher educators, inspired by Xu and Brown’s (2016) Teacher Assessment Literacy in Practice framework (TALiP). Then, informed by Mertler and Campbell’s (2005) Assessment Literacy Inventory from the field of general education, and based on the teacher educator LAL framework developed in stage One, a Language Assessment Literacy Inventory (LALI) for teacher educators was created. The aim of this instrument (LALI) – an online ‘test’ consisting of language assessment scenarios for which the teacher educator needs to select the most suitable solution from a set of options – is to measure language teacher educators’ knowledge as specified in the LAL Framework. During the Work-in-Progress session, the development of the LAL framework and the Language Assessment Literacy Inventory will be described. Additionally, initial results from the administration of the LALI to 50 teacher educators in Chile will be shared, to shed light on their LAL as well as on the quality of the LALI instrument. It is hoped that the LAL framework and the LALI instrument will help gain an understanding of the LAL of teacher educators, and by extension their potential to drive a multidimensional level of LAL in future generations of teachers – many of whom will operate in contexts of movement, transition, and change.

5) The Impact of Raters’ Use of an Analytic Rating Scale on Language Advising

Stephanie Rummel, University of Auckland, New Zealand
Morena Botelho de Magalhães, University of Auckland, New Zealand
Judit McPherson, University of Auckland, New Zealand

The Diagnostic English Language Needs Assessment (DELNA) is used at the University of Auckland to help identify the Academic English needs of students following admission in order to direct them to appropriate support (Read, 2015).
The assessment is conducted in two stages: a screening to filter out more proficient users and a diagnosis. The diagnosis includes reading, listening and writing components. After students complete the diagnosis, most are asked to come in for a 30 minute or 1 hour advisory session with a language adviser. The DELNA writing component is rated by trained raters using an analytic rating scale. As our assessment is used for diagnostic purposes, the analytic scale measures a number of different aspects of writing and provides detailed feedback to students regarding their strengths and weaknesses (Weigle, 2002). Currently, the scale includes nine traits clustered in three categories: coherence and academic style (text organisation, cohesion inside text and academic tone), content (description of data, reasons for trends observed, expansion of ideas), and form (sentence structure, grammatical accuracy, and vocabulary). Each trait is divided into six band levels ranging from four to nine. As raters rate, they are to fill out a marking sheet while referring to graded level descriptors for each trait. There is space on the marking sheet for raters to award a band for each of the nine traits, along with room for them to comment on each trait and provide ticks for correct uses of cohesive devices and referencing. They are also asked to provide crosses for incorrect uses of grammar and vocabulary and language impacting academic style, such as personal pronouns, contractions and informalities. Language advisers then discuss the marking sheet with the student during the advisory session to provide a detailed overview of their linguistic strengths and weaknesses.

Because the rating scale plays an integral part in the DELNA advising process, questions have emerged regarding the quality and helpfulness of the raters’ comments. For the current study, a selection of rating sheets with detailed comments from experienced raters will be analysed and coded by two independent researchers. The initial codes will identify which comments are considered valuable by language advisers in that they allow advisers to provide constructive feedback related to specific aspects of students’ writing such as grammatical forms, development of ideas, and academic style. Further coding will then take place to establish themes regarding features such as specificity and clarity that make a comment valuable or not valuable. It is hoped that once themes have been identified, training procedures can incorporate the findings and enable raters to utilise the scale to its full potential so that the information provided is valuable to advisers and students.

6) **Using Virtual Reality (VR) for Language Assessment**

Jayoung Song, Rice University, USA
Wei-Li Hsu, Rice University, USA

For the past decades, computer technology has been increasingly used in all aspects of language testing, including test design and development, test administration, scoring, and analysis in order to accurately measure one’s second language speaking ability (Chapelle & Douglas, 2006). One of the recent technologies is Virtual Reality (VR), where users can experience immersive environment with VR goggles. The purpose of the present study is to explore if a VR could be a suitable platform for second language assessment. It specifically compares Korean as a foreign language (KFL) students’ performance in a VR and a face-to-face test. A total of 30 KFL students enrolled in a large language institution in a southwestern part of the United States participated in the study. After receiving instruction with VR videos where students practiced speaking with pre-recorded VR videos, students took an oral exam both with a VR and in a face-to-face modes. The order of the two modes among the 30 participants were counter-balanced. Data was drawn from students’ oral test scores, a survey asking their perceptions towards the two testing modes, and interviews. The findings will discuss students’ scores and perception toward this new technology use in assessment. The preliminary analysis suggests no statistical significance between the two modes, regarding the test scores and perception. However, the analysis of interview indicates that students’ preference to a certain mode for different reasons. The preliminary findings suggest the potential of VR technology in distance-learning to elicit similar language production in role-play tasks. It would be beneficial to have feedback on scoring to examine more refined differences between the two modes.

7) **Development of a Framework for the Assessment of Interactional Competence: A Case of Spanish Language Learners in Trinidad & Tobago**

Romulo Guedez-Fernandez, The University of the West Indies, Trinidad and Tobago

This study investigates the assessment of speaking performance of a three-year degree Spanish programme at the university level in one Anglophone Caribbean country. Spanish plays an important role, both in high school education and as a desired communication tool/asset. Students seeking to be admitted into this programme must possess a level of proficiency equivalent to the A2+ level of the CEFR. Classroom instruction aims at facilitating students’
development of grammatical, sociolinguistic, strategic and interactional competences. Some inconsistencies in terms of the constructs of content and interactional competence in the assessment of the oral trait were identified from a pilot study. These findings suggested that the current rating criteria used to assess speaking performance were to be empirically examined. This would allow researchers to identify any sources of discrepancy between the context of teaching of the Spanish language degree programme and the assessment of peer-to-peer speaking performance. As a result, it has been suggested that more specific descriptors for the assessment of the constructs of content and interactional competence should be formulated. This research focuses on the construct of interactional competence. The very nature of interactional competence is the co-construction of the actions and utterances during a particular discursive practice. Thus, interactional competence, which implies individual knowledge and abilities in using socio-cultural rules of discursive practices for interaction with other participants, is not restricted to the single individual language user which has been the focus of previous communicative models (Young 2000, 2011, 2013; He and Young 1998). Participants (n=116, year one (n=44), year two (n=34) and year three (n=38)) are Spanish majors and minors who are enrolled in CEFR levels B1, B2 and C1 Spanish language courses. Participants previously had attended formal Spanish language instruction for at least (7) seven years in high school. The methods of data collection that will be used in this investigation include: a) transcribed video-recordings of peer-to-peer performances in conversation tests, b) verbal recall with candidates’ videotaped tests, c) verbal protocol with raters on the criteria that is used to assess candidates’ performances in the speaking test, and d) semi-structured interviews with test-takers. Peer-to-peer test performances will be analysed using conversation analysis. Factors associated with candidates’ performance during the test and their performance in the classroom will be drawn from both candidates’ and raters’ perspectives. Scores awarded to candidates will be analysed using Multifaceted Rasch (MFRM) analysis for detecting and correcting rater variability. The investigation makes a case for the development of an empirical and theoretical knowledge-based framework for the assessment of the interactional competence construct in this context. These findings are expected to offer guidance for enhancing the quality and accountability of the Spanish undergraduate programme assessment.

8) Searching for the Yardstick: Measuring the "Readiness" to Incorporate Sources in L2 Academic Writing
Izumi Watanabe-Kim, International Christian University, Japan

Past studies concerning academic writing assessment most frequently measure proficiency by quantifying lexical variation, fluency, accuracy, and complexity. Since these measures do not consider other critical elements that distinguish academic writing from other forms of writing practice (e.g., referencing and citing sources), it is vital that L2 language educators in higher education monitor and assess the ability to use and misuse (i.e., plagiarism) sources in academic writing.

Most research on plagiarism focuses on either: (a) attitudes towards plagiarism via surveys and interviews (Sutherland-Smith, 2005; Li and Casanave, 2012; Wheeler, 2009; Thompson, Morton, and Storch, 2013; Currie, 1998), and/or (b) degrees of plagiarism through analysis of writing samples (Li and Casanave, 2012, Currie, 1998; Pecorari, 2003, 2006; Storch 2009, 2012). Possible causes of plagiarism have also been discussed. Some argue that non-Western L2 writers’ acts of plagiarism can be attributed to their culture and perception of learning (Currie, 1998; Barker, 1997), while others claim that plagiarism is caused by lack of language proficiency and experience in academic writing (Kirkland and Soudners, 1991; Li and Casanave, 2012; Wheeler, 2009). Nevertheless, researchers seem to uniformly agree on the need to develop more effective instruction and measures to assess the ability to use sources appropriately in academic writing (Pecorari, 2003; Storch, 2012).

To this end, this study is a work-in-progress project to develop a way to assess and measure students’ "readiness" to undertake academic research writing at the exit point of an undergraduate EAP program. First, drawing upon the literature, an operational definition of plagiarism was sketched out. Next, an assessment instrument was developed to include: (1) yes/no items regarding when to cite, (2) error identification items to assess the knowledge of the mechanical aspects of citations, and (3) items relating to acceptability of a variety of writing samples that may or may not exhibit plagiarism.

The pilot test was given to 96 Japanese university students enrolled in a required writing course, where they received instructions on how to incorporate sources in academic writing. The descriptive statistics of this pilot study indicate that three sections of the test measured discrete skills. The post-test session reveals that although instruction seems to be effective in helping students acquire the mechanical skills of citation, many appeared to intuitively answer questions regarding ethics and academic integrity. It also became apparent that reading comprehension seems to interfere with their judgments of the acceptability of source use.
Finally, this presentation aims to initiate a discussion of how to assess the complex process of using sources in academic writing, and to receive feedback on how best to proceed in the next stage: the refinement of test items to increase reliability and validity. More specifically, it seeks to ask: 1) if the test is measuring what it intends to measure, and if so, 2) whether it is reflective of the knowledge and skills required to incorporate sources in academic writing.

9) An Investigation into the Cognitive Validity of the Speaking Section of the Vietnam’s Standardized Test of English Proficiency
Huu Nguyen, Ministry of Education and Training, Vietnam

This study in progress is designed to craft a validation argument for the Vietnam’s Standardized Test of English Proficiency (VSTEP) in relation to the speaking’s cognitive processing. For such an aim, the cognitive processing models applied in the language domain of speaking are studied. Validation evidence will then be elicited via various methods of survey questionnaires, focus group interviews, and stimulated recall, which are used to study the extent to which the cognitive process that the VSTEP test-takers are supposed to have in test conditions resembles the one in non-test conditions. These methods are designed for the different processes that contribute to the target construct of the test, which are reflected in the input, the test format, and the test tasks (Field, 2013). The results of the study are expected to provide implications for the commissioning and calibration of the VSTEP speaking tasks and rating scale. The VSTEP which was developed within the scope of the National Foreign Languages 2020 Project of the Ministry of Education and Training of Vietnam (the NFL P2020) is a measure of English language proficiency of B1 to C1 levels. The mission of the NFL P2020 is to enhance foreign language education, especially English education, in Vietnam to prepare for the economic growth of the country. The four-skill English language test VSTEP was among the outcomes of the NFL P2020, important and critical to the accomplishment of the said mission of the NFL P2020.

10) Considerations for Assessment in the Multicultural Tertiary EFL Context: The Case of Arab EFL Students
Özlem Yalçın Çolakoğlu, Bahçeşehir University, Turkey
Merve Selçuk, Altinbas University, Turkey

Due to the political problems that the middle-east region has been dealing with for the last years, Turkey has become a widely preferred country for people seeking shelter or looking for other countries to continue their higher education. According to the Council of Higher Education (CoHE) in Turkey, universities all around the country have received around 40,000 students from Arabic speaking countries in the last academic year and this figure continues to grow. This increasing demand has caused changes in the student population of English Language Preparatory Programs, not only in terms of incoming nationalities but also in the variety of first languages (L1) with their orthographic differences, learner types, attitudes and most importantly their needs. When compared to the main population of Turkish EFL learners, correlation analysis reveals that there is a significant difference in Arab learners’ speaking and writing performances indicating that they are more proficient in their speaking skill whereas they lag behind in writing. Normally, similar results are expected in both productive skills since “well-developed oral language skills in English are associated with better writing skills in English” (Geva, 2006 p.14). When taking a proficiency exam, usually consisting of a single cut-off score, at the end of the program, Arab learners’ strengths in other skills compensate for their weakness in writing, which eventually promotes them to their faculties without having acquired the necessary written proficiency in English. Therefore, EFL/ESL language programs welcoming these students should take this problem into consideration as writing is crucial for academic success. Thus, this longitudinal study, based on the data collection of a full academic year, is undertaken to test the hypothesis that Arab university students coming to Turkey to study at the tertiary level have weaker writing skills compared to their speaking performance. Quantitative and qualitative data are being collected from 250 Arabic-speaking English as a foreign language learners, enrolled in the English language preparatory program at a foundation university in Istanbul, at four different proficiency levels, classified by the Common European Framework of Reference (CEFR) as basic and independent users, to look at the relationship between their writing and speaking test scores. In order to identify the reasons why Arab EFL learners perform differently in their productive skills, text analysis of rhetorical patterns and other features of writing will be conducted to identify the problematic areas in their language development. The study aims to provide implications for EFL classrooms and to raise EFL teachers’ awareness regarding the needs of their culturally diverse students. More importantly, with the discussion of the data, the study seeks to encourage policy makers,
administrators and all other stakeholders involved in the language testing process to reconsider important decisions such as; whether having separate cut scores for the different skill sections in the proficiency exam would benefit the students more than having one single cut score as it would prevent the faulty passing of less proficient students to their faculties.

11) Enacting the Assessment Use Argument for a Large-scale Academic English Language Proficiency Test: The Question of Construct
Rosalie Grant, WIDA, USA
Mark Chapman, WIDA, USA

This work-in-progress session addresses one aspect of the assessment use argument of a large-scale, high-stakes English language proficiency test. The test is used to evaluate the academic language development of students who migrate to the United States or who grow up in a household where languages other than English are used. Test scores are used at students’ transitional educational moments, as they enter English language schooling for the first time and also as a criterion for when students are ready to exit English language support services. The study examines the construct validity of the internal structures of ACCESS for ELLs 2.0 through item level structural equation modeling. This technique enables hypotheses to be tested about relationships between item responses and hypothesized latent constructs.

The test is designed to assess English Learners’ (ELs’) academic English language proficiency (ELP) in the U.S. K-12 public school context, as described in the WIDA document, *English Language Development Standards, Grades 1-12*. Five academic Language Standards (Social and Instructional Language, Language of Language Arts, Language of Social Studies, Language of Science, and Language of Math) each addressing four language domains (Listening, Reading, Speaking, and Writing) reflect the “language expectations for ELs represented within progressive levels of (academic) language proficiency (WIDA, 2012)”.

The research question asks whether there are distinct language factors underpinning the responses of ELs in Grades 1-12 to the online version of ACCESS. The hypothesis is that ELs’ responses can be explained by four first-order language domain factors (Listening, Reading, Writing, and Speaking), two second-order language factors (Receptive and Productive) and a third order language factor (Academic Language). This hypothesis is grounded in the construct of the assessment, as defined in the English Language Development Standards.

Analyses are based on the 2015/16 test population of 1,148,000 students who took the online test, across 38 states. Each selected response item or constructed response task was designed to assess a particular ELP level within a particular domain and academic language content standard. Grade-specific models are being developed for three cohorts, each reflecting students’ responses to items designed for a particular ELP level – high, medium and low. During the modeling process, particular attention is paid to: (a) the conceptual soundness of the model; (b) factor loadings; (c) RMSEA, CFI/TLI, AIC and BIC goodness-of-fit values; (d) modification indices; and (e) model parsimony. Should an initial postulated model be rejected and a decision made to re-specify and re-estimate the model, as Byrne (2012) recommends, attention is paid to: (i) whether the estimation of the targeted parameter is substantively meaningful; (ii) substantively large modification indices and expected parameter change values; and (iii) whether the refitted model will lead to an over-fitted model.

The work-in-progress session will focus on the aims of the study, the design and construct of the assessment, and the methodology. The presenters hope to receive feedback on the approach and the extent to which it will provide evidence of the construct validity of the assessment, as part of the ACCESS for ELLs 2.0 assessment use argument.

12) Revision of a Pronunciation Rating Scale For a University-based English Oral Assessment: What Do Raters Look for?
Li Liu, Renmin University of China, China

Pronunciation scales have long been shown to be highly problematic to design and implement. Challenges include the consistency in descriptors and clear trajectories and so on (Harding, 2013; Isaacs, 2014). The study reports the revision of a pronunciation rating scale of *Test of Oral Proficiency in English*, a university-based English oral assessment in Renmin University of China. It explores how teacher raters view the original and revised rating scales in general and how they use the revised scale to award scores to students at different level bands. Pronunciation is specifically examined in the TOPE by a reading aloud task. The original rating scale is analytical while the revised one is holistic.
In the first phase of the study, questionnaires were employed to elicit perceptions from 34 experienced raters on both the old and new rating scales, including 1) the problems they encountered when using the rating scale descriptors and 2) their confidence about judging a candidate’s use of different features covered in the rating scale descriptors and 3) the features they focused on when distinguishing between levels. Candidate speaking samples are all collected and videotaped from live tests of TOPE. For the second phase, stimulated verbal protocol will be conducted for raters to reflect on the salient features in informing their scoring judgments.

Initial findings included: raters did not develop a consistent interpretation of assessment criteria and thus mainly take segmental features into consideration. And they tend to prefer the revised rating scale using holistic descriptors to the old one using analytical scoring method. In addition, there is significant rater variability regarding interpretation of rating features. Follow up studies will be conducted using acoustic analysis tools to examine candidate speech samples at various band levels to provide another sources of evidences. The study provides insights into definition of the construct, further revision of the descriptors and underlines the need for rater training.

13) Rater-cognition in Progress: An Institutional Change Study
Sukran Saygi, Middle East Technical University, Turkey

This study aims to scrutinize the role of cognitive attributes in decision-making of teachers of English as a foreign language when rating writing. When they interpret and apply rating criteria, the raters become the arbiters of the test construct as expressed in the scores they assign (McNamara, 1996). However, as a highly complex concept, rater judgment is influenced by numerous factors (Lumley, 2002). Understanding rater cognition and investigating the effects of rater attributes on scores or raters’ mental processes during rating would be of great importance to limit the effect of this complex and often unpredictable nature of human judgement.

This work-in-progress presentation will outline a study designed to investigate the cognitive processes involved in rating written English. Data for the study is collected at a preparatory school of Middle East Technical University, Turkey. At the institution, there has been a curriculum renewal project going on since 2015. As part of this project, the existing institutional proficiency exam, which also serves as the exit exam for the foundation year, was officially replaced with the updated one as of December 13, 2017. Therefore, the EFL teachers employed at the institution need to rate new writing tasks using new rating scales, which means they have been experiencing different rating processes. Analysing these processes will guide the test-developers and the school administration for the possible future actions to be taken, including further rater training sessions. In the light of these, the study sets out to answer the following questions: (i) What behaviours do the participant EFL raters attend to while they assess student essays when they use an unfamiliar set of criteria? (ii) How do they process the new rating scales? (iii) How can these processes and behaviors be related to their personal and professional backgrounds?

During the first phase, 5 teachers with varied experience levels rate a set of writing performances (N=15) and produce concurrent think-aloud protocols. Audio recording of these protocols are transcribed. The transcribed interviews are analysed following the decision-making behaviors of raters suggested by Cumming et al. (2001). Following the initial analysis of the protocols, face-to-face interviews are conducted in order to investigate the influence of background factors on the raters’ behaviors. In the second phase, the same procedure is repeated after the administration of another exam in June 2018. Here, the researcher tries to understand if the same raters change their behaviours or their views of the rating scale after they have got more familiarised with the new scale after scoring student essays as part of their teaching practices during the semester.

All in all, this study seeks to contribute to the field’s understanding of rater behaviour involved in rater judgment. This presentation will present results from the first round of data collection.

14) Investigating the Design of Assessments of Writing Ability for Transition from Education to Work
Alan Urmston, The Hong Kong Polytechnic University, Hong Kong, China
Michelle Reyes Raquel, The University of Hong Kong, Hong Kong, China
Felicia Fang, The Hong Kong Polytechnic University, Hong Kong, China

The Graduating Students’ Language Proficiency Assessment (GSLPA) is an assessment of written and spoken workplace English, developed at one university in Hong Kong by a team of language assessment experts and launched as a graduation requirement at that university in 2000. The Assessment was redeveloped in 2007, utilising the best of available technology for assessment delivery, scoring and reporting. Since then, the Assessment has been taken by
over 20,000 graduating students. In 2013, the Human Resources Office of the University requested a test of written English proficiency to help them make decisions on recruitment and promotion of non-academic staff and after careful consideration it was decided that the GSLPA Written Language Test could be used for this purpose. The decision was based on the premise that an assessment designed to predict how well graduating students might be able to use English in the workplace would also be a valid measure of the same abilities expected by people working in non-academic positions in a higher educational institution. The new assessment (known simply as the Language Proficiency Test, or LPT in the University) has been running since 2014 and has been taken by over 2000 candidates. This work-in-progress reports on a longitudinal investigation into the appropriacy of this assessment instrument for its purpose as the LPT. Typically, it is difficult to build a test-use argument (Bachman, 2005) for exit assessments as once they “exit” the institution or stage of education/employment they are in, it can be impossible to track them down to see whether in fact the assessment was able to predict future performance in language-use tasks. As the test takers will remain within the university and as such be accessible to researchers, this unique situation gives us this chance. Specifically, we are looking to answer the following research questions:

1. Do the LPT and GSLPA candidates interact in the same or different ways with the writing tasks they have to address in the assessment?
2. Does performance on the writing tasks predict performance in the workplace in an administrative setting?
3. What can the research tell us about the design of exit/entrance tests at points of transition?

The research adopts a multi-method approach: first, by looking at how the different task versions in the Assessment allow for the characteristics of writing ability to be demonstrated by employees/potential employees of the University utilizing methods of Many-Facet Rasch Analysis; second, by talking to experienced raters to gain their insights into these characteristics; third, by tracking the progress of these employees in the University after taking the Assessment; and finally, through an evaluation of the usefulness of the Assessment for its purpose by its major stakeholders, the Human Resources Office. It is anticipated that the research can throw light on the issues of the design of assessments at transition points, such as the exit from education and the entrance to the workplace.

15) Assessing the Strategic Competence of Young Language Learners for Oral Communication: The Development of a Self-assessment Scale
Jun Wang, Shanghai Jiao Tong University, China

In language research, the importance of Strategic Competence (SC) has gained a substantial recognition nowadays. There is an increasing number of strategy-based research; many educational policies and curricula include using strategies as a part of the language learning goals. Following such a trend, this study aims to develop an assessment instrument specifically for assessing SC. Young Language Learners (YLLs) refers to students who learn the target language “during the first six or seven years of formal schooling” (McKay, 2006, p. 1). Studies have confirmed the existence of a positive relationship between applying strategies and being ‘good’ learners (see Macaro, 2006, for a review). Learning the target language along with the development of SC could help YLLs both use the target language in an efficient and appropriate matter and acquire it more effectively.

This study selects self-assessment as the target instrument to assess YLLs’ SC. The reason is mainly twofold: (a) strategy questionnaire is widely used in strategy-based research, and studies have proved its usability on the elicitation and evaluation of learner strategies (see Cohen, 2011; Oxford, 2011); (b) alternative assessment is argued to be more suitable for YLLs assessment by being conducive to children’s emotional wellbeing. As one type of the alternative assessment, self-assessment shares this merit, and studies have proved its feasibility for assessing YLLs (e.g., Butler & Lee, 2006, 2010; Dan, 2002).

Thus, the overall purpose of this study is to conceptualize the construct of YLLs’ (Chinese young learners of English aging from 9 to 12) SC for oral communication and operationalize it into a self-assessment scale. To fulfill this purpose, this study applies a combination of intuitive, qualitative and quantitative approaches. The procedure of scale construction includes: (a) a review of literature to set up a descriptive scheme for the target construct, (b) a document analysis among relevant materials to collect descriptors, (c) a series of simulated recall interviews and classroom observations to collect strategy-using data from the target population, and (d) a round of expert consultation to revise the drafted scale. The validation procedure includes: (e) a pretested large-scale self-assessment survey to examine the construct model and to investigate the scale quality through advanced statistical methods (i.e., SEM & Rasch analysis), and (f) a series of semi-structured interviews to explore end-users’ (i.e., teachers & students) attitudes and perceptions. The above (a) and (b) procedures have completed the preliminary analyses. A 2-week pilot study for the
classroom observation was conducted in May 2017 (Grade 3-6, 14 classes). The results show abundant applications of oral strategies during classroom activities. Corresponding to a rapidly globalized world, the objective of English education in China has gradually shifted from knowledge accumulation to effective communication, and the importance of being strategic learners and users is increasingly recognized. Following the wave of such a change, this study aims to develop an empirically-grounded self-assessment instrument to assess YLLs’ SC. The findings may yield positive influence on strategy education for YLLs in China, and provide useful implications for strategy-based research and assessment instrument development practices.

16) Development and Validation of a Diagnostic Rating Scale for an EFL University Writing Classroom
Apichat Khamboonruang, University of Melbourne, Australia

In recent years, increased attention has focused on classroom-based assessment to explore how best to generate information that can support language learning and teaching and ultimately promote learners’ language development during an ongoing language classroom (e.g., Elder, 2017; Jang, 2012; Knoch & Macqueen, 2017; Lee, 2015; Turner & Purpura, 2016). However, few studies have investigated to what extent a classroom-based assessment can best be implemented in an intact classroom over time. This mixed-methods PhD research sets out to develop and validate a diagnostic rating scale (DRS) that can be implemented across expository writing genres to provide specific, detailed, and targeted information regarding students’ writing strengths and weaknesses and learning progress, and to facilitate teachers in continually improving students’ writing ability over an ongoing EFL writing course in the Thai university setting. In this research, the DRS criteria are informed by theoretical, empirical, and intuitive sources. Kane’s argument-based approach to validation (Kane, 1992, 2006, 2013, 2016) is used as the theoretical framework for the development and validation of the DRS.

To collect qualitative and quantitative data to support the DRS development and validation, this mixed-methods research employs a multiphase exploratory sequential design (Creswell & Plano-Clark, 2011, 2017; Galaczi & Khabbazbashi, 2016), where two types of data are accumulated over three study phases. Participants include five teachers and around 82 English-major undergraduates from four classrooms in two writing courses in Mahasarakham University, Thailand. In Phase 1, the DRS was designed and developed based on theoretical, empirical, and intuitive sources. In Phase 2, the DRS was piloted by the teachers who evaluated the DRS, provided feedback for the DRS revision, and applied the DRS to assess prior student writing samples, produced in the context. In Phase 3, the DRS will be administered in the four classrooms, where the teachers will use the DRS to assess their students’ essays and the students will use the DRS to self-assess their own essays several times throughout the courses. The concurrent think-aloud and post-interview methods will be employed to elicit the teachers’ diagnostic rating behaviours. Interviews will also be used to investigate the teachers’ perception of the DRS quality and usefulness, the students’ self-assessment behaviours, and the students’ perception of the DRS usefulness and quality. The quantitative data will be analysed using Classical Test Theory, Generalizability Theory, and Many-Facets Rasch Model approaches and the qualitative data will be analysed following the qualitative content analysis.

The key findings of this research are expected to (1) provide guidance for researchers and test developers to better understand and characterise the construct of EFL writing for classroom diagnostic assessment by drawing on several theoretical and contextual sources, (2) illuminate whether Kane’s argument-based validation approach will be viable for validation of an ongoing assessment in the classroom, and (3) serve as an example of language assessment research that employs multiple data sources, Kane’s argument-based validation approach, and a mixed-methods research design to develop and validate language assessment tools in general and classroom diagnostic assessment instruments in particular.

17) Diagnosing Listening and Reading Skills in EFL Context: Cross-mode Stability and Variability
Lidi Xiong, Zhejiang University, China
Shangchao Min, Zhejiang University, China
Min Wang, Xi’an Jiaotong University, China

Cognitive diagnostic models (CDMs) are now receiving more attention for its crucial feature of identifying strengths and weaknesses of target learners, but the actual application is still far from widespread, with most researches focused on model comparison (e.g. Lee & Sawaki, 2009a; Lee & Sawaki, 2009b) and reading skills (e.g. Jang et al., 2015; Li et al., 2016). Listening seems to be an under-researched skill and much less research has examined the two
receptive skills at the same time or tapped into the relationship between the fine-grained attributes across modes. Although some empirical research has indicated that listening and reading share a considerable amount of common variance despite the separability (e.g. Song, 2008; Zeelant & Schmitt, 2012), most of them stopped at claiming listening and reading are similar or different, failing to present clearly how the subskills operate in the comprehension process under the two modes and how the different subskills contribute to the overall comprehension. To address this gap, this study attempts to: 1) diagnose the strengths and weaknesses of undergraduate students’ EFL listening and reading ability; and 2) examine to what extent the attributes shared by EFL listening and reading operate similarly or differently under the two modes and how those different subskills contribute to the overall comprehension. The preparation phase of the project consists of test and Q-matrix design. The instrument we used is a high-stakes in-house EFL exit test, which means undergraduates at this university need to pass the test to get their bachelor’s degree. In terms of attributes definition, we referred to previous studies, expert analysis as well as the verbal protocol analysis. Six attributes shared by the listening and reading test were identified, listed as follows: (A1) understanding words, (A2) understanding syntactical structures, (A3) extracting explicit information, (A4) summarizing main ideas, (A5) making inferences, and (A6) identifying attitudes and purposes. Our data stem from the aforementioned large-scale test administered in a prestigious university in China, with over 8000 test takers annually. Because of the multi-matrix design, a subsample of 917 students was chosen for initial analysis. In analyzing phase, G-DINA package from R was utilized since we found G-DINA best fitted the data compared with other diagnostic models like DINO, DINA, LLM, ACDM and it has also been empirically proved to be reliable for diagnostic purpose (De La Terre, 2011). We have obtained results indicating test-takers’ mastery of each attribute, their latent classification and across-group mastery differences, through which we could build detailed learning profiles, addressing strengths and weaknesses of individual learners and facilitating teaching. The paper will proceed to explore operating patterns of attributes across modes using structural equating modeling (SEM) to see whether the same attributes function differently across modes and use data from different test forms to cross-validate the results. This study not only provides individualized feedback to students for their listening and reading using CDA, but also adopts a new approach to study the relationship between the two receptive skills at a more fine-grained level.

18) Integrating Language Proficiency into Academic Readiness Screening

Peppi Taalas, University of Jyväskylä, Finland
Anna Grönlund, University of Jyväskylä, Finland

This paper explores a work in progress development project on admissions procedures for International Master’s degree programs. Our focus is especially on integrating language proficiency into the concept of academic readiness. The aim is not to create yet another language test but instead a more comprehensive, diagnostic approach to both screen and predict the prospective students’ skill set and preparedness for academic study. The results also inform syllabus planning and provide preliminary information on the individual needs of the prospective students. Our previous study has shown that the current language proficiency requirements and the ways of validating that proficiency in the Master’s level admissions processes do not predict the skills required for academic study. The importance of researching the development of selection mechanism has been highlighted in literature to better understand and identify the combination of factors that are predictive of academic performance (see Cliff and Hanslo, 2009). This has led us to look for an alternative approach in which language skills are embedded in the comprehensive academic skill set and assessed in that context in the IMDP application process. College or academic readiness, a concept typically used in North American context, is a multifaceted and also rather fuzzy concept comprising numerous factors internal and external to the classroom environment. It is usually described to consist of four components: key cognitive strategies, academic knowledge and skills, academic behaviors and contextual skills and awareness (Conley, 2007). As a concept academic readiness is related to academic literacy. It seems that language proficiency does not figure in the concept as such whereas writing is mentioned as one of the overarching components of the academic skills set. Our novel approach brings this idea into the higher education and Master’s level admissions context with special focus on academic language skills. In our pilot study we have expanded the prevailing admissions procedures to include a structured process to screen the academic readiness of applicants. This procedure consists of three stages planned in cooperation with language experts and IMDP staff. The applicants shortlisted for interviews are given an online pre-task which is time-limited and based on a journal article and a popular article. The task output is free-form and the applicants are expected to be able to complete it only with the given prompts. In the interviews the applicants are required to reflect on completing the pre-task in addition to the more typical questions on motivation, academic background etc. The completed pre-tasks and the taped interviews are assessed jointly by the language experts and the IMDP staff, and the results inform the final student selection. After two rounds of piloting this approach in the
admissions process our preliminary impressions are optimistic that this approach brings added value to the admission process, the syllabus design and the planning of language elements in the curriculum.

19) Investigating the Spoken Production of Young Learners Aged 13-15
Gwendydd Caudwell, Cambridge Assessment English, UK

This presentation outlines a study, which aims to contribute to exploring the broad question of how we can assess the speaking skills of 13-15 year olds at B2 level of the CEFR. At this age, they are still in the process of developing from a child into a young person, both mentally and physically, and their world-knowledge is still rather limited. This may mean that testers cannot use test tasks and topics developed for Adults, as well as the rating scales developed based on the characteristics of adult performances expected at this CEFR level.

This study comprises of two main phases aimed at identifying differences in spoken performance of Adults and Teenagers and therefore how ultimately that may affect future considerations in tasks and rating scales designed for this age group. The first is to examine what different characteristics teenager performance exhibits compared to those of adults on a B2 level, computer-based, test task. Complexity, Accuracy and Fluency (CAF) measures along with coding of particular characteristics identified in the study are used to compare 30 samples each of 4 different groups. These are both native and non-native speaker teenagers and adults. These groups are chosen to not only compare between adult and teenage performance but also to investigate whether native and non-native speaker teenagers exhibit the same characteristics or approach to the task. The aim being to identify particular characteristics as related to the social, cognitive and linguistic developments of teenagers rather than to language learning.

The second phase examines which characteristics contribute most to the overall score through multiple regression analysis and stimulated recall with raters. This aims to investigate whether any particular characteristics identified in phase one, affect the score and therefore should be reflected accordingly in the scales.

Initial findings from phase one data analysis will be discussed. Key highlights of this study are exhibited in the ways in which different characteristics are captured using CAF measures and how certain measures may need to be further refined in order to better capture them. This broadens our understanding of how teenagers perform in spoken tasks and ultimately highlights some key considerations in task design for this age group in order to ensure that they can perform to the best of their abilities. This study also contributes to the larger question of whether this age group are capable of demonstrating all performance elements as described in the CEFR. And therefore, whether a young learner CEFR should be constructed taking these differences into account.

20) Assessing machine translation in use: Trustworthiness and safety of Google Translate
Translations in Obtaining Surgical Consent
Susy Macqueen, Australian National University, Australia
Christine Phillips, Australian National University, Australia
Tina Campbell, Australian National University, Australia
Zhengdao Ye, Australian National University, Australia
Johanna Rendle-Short, Australian National University, Australia
Katrina Anderson, Australian National University, Australia
Diana Slade, Australian National University, Australia

Healthcare contexts are linguistically and culturally diverse places, and many of the interactions within them occur under stressful circumstances. Obtaining consent prior to surgical procedures, for example, is an especially complex and high-stakes medicolegal interaction which is likely to be difficult to comprehend for people who are consenting in a second language in which they have limited proficiency (Fink et al., 2010; Sudore et al., 2006). As comprehension is critical in informed consent procedures, there is a chance that clinicians will reach for convenient tools such as Google Translate (GT) for assistance, despite evidence that it may mistranslate complex medical concepts (Patil & Davies, 2014). It is important that translation tools are evaluated regarding their suitability for this purpose in terms of trustworthiness and safety. However, the assessment of machine translation is not typically considered in its interactional context, nor is it carried out using theories and methods drawn from the fields of language assessment, instead tending to be based on automated evaluations of overlap with reference translations and/or under specified criteria for human evaluations.

This paper reports on work in progress which evaluates the translations provided by Google Translate in obtaining surgical consent. While prior studies of machine translation have evaluated static written texts or sentences (e.g. Taylor et al., 2015), this study considers the efficacy and appropriateness of machine translations within the dynamics
of high-stakes conversations. Role-plays between obstetricians and simulated patients with limited English proficiency were carried out in the process of obtaining surgical consent for Caesarean section. GT was used to translate between English and two typologically disparate languages: Chinese and Indonesian. The role-plays were video-recorded and transcribed, and 'translation episodes' were extracted for evaluation. Insights from role-play participants, translators and health practitioners were sought in the development of criteria for assessment of the translation trustworthiness. The theoretical basis for the assessment approach was derived from assessment practices in translation and interpreting (Eyckmans et al., 2016) and the concept of indigenous criteria (Jacoby & McNamara, 1999). The study is potentially revealing in respect to the assessment of machine translation in the dynamics of spoken interaction in healthcare settings.
## Index of Presenters

<table>
<thead>
<tr>
<th>Name</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anderson, Katrina</td>
<td>120</td>
</tr>
<tr>
<td>Arias, Angel</td>
<td>80, 102</td>
</tr>
<tr>
<td>Aryadoust, Vahid</td>
<td>96</td>
</tr>
<tr>
<td>Bachman, Lyle</td>
<td>33</td>
</tr>
<tr>
<td>Baker, Beverly</td>
<td>80, 102</td>
</tr>
<tr>
<td>Banerjee, Jayanti</td>
<td>62</td>
</tr>
<tr>
<td>Barron, Chris</td>
<td>67</td>
</tr>
<tr>
<td>Batt, Aaron Olaf</td>
<td>87</td>
</tr>
<tr>
<td>Beltran Zuniga, Jorge Luis</td>
<td>109</td>
</tr>
<tr>
<td>Berger, Cynthia M.</td>
<td>95</td>
</tr>
<tr>
<td>Bishop, Kyoungwon</td>
<td>64</td>
</tr>
<tr>
<td>Blood, Ian</td>
<td>58</td>
</tr>
<tr>
<td>Boals, Tim</td>
<td>100</td>
</tr>
<tr>
<td>Bonk, William</td>
<td>58</td>
</tr>
<tr>
<td>Botelho de Magalhães, Morena</td>
<td>111</td>
</tr>
<tr>
<td>Boynton, Simon David</td>
<td>103</td>
</tr>
<tr>
<td>Brantmeier, Cindy</td>
<td>105</td>
</tr>
<tr>
<td>Brown, James Dean</td>
<td>93</td>
</tr>
<tr>
<td>Campbell, Tina</td>
<td>120</td>
</tr>
<tr>
<td>Castro Alvarez, Pedro</td>
<td>98</td>
</tr>
<tr>
<td>Castro, Mariana</td>
<td>100</td>
</tr>
<tr>
<td>Caudwell, Gwendydd</td>
<td>120</td>
</tr>
<tr>
<td>Chapelle, Carol</td>
<td>80</td>
</tr>
<tr>
<td>Chapman, Mark</td>
<td>81, 97, 115</td>
</tr>
<tr>
<td>Chen, Ivy Win-wen</td>
<td>72</td>
</tr>
<tr>
<td>Cho, Yeonsuk</td>
<td>58</td>
</tr>
<tr>
<td>Choi, Ikkyu</td>
<td>53</td>
</tr>
<tr>
<td>Christison, MaryAnn</td>
<td>33</td>
</tr>
<tr>
<td>Cliff, Alan</td>
<td>50</td>
</tr>
<tr>
<td>Clothier, Joshua</td>
<td>73</td>
</tr>
<tr>
<td>Cook, Howard Gary</td>
<td>64</td>
</tr>
<tr>
<td>Cushing, Sara</td>
<td>74</td>
</tr>
<tr>
<td>Czajkowski, Michelle</td>
<td>67</td>
</tr>
<tr>
<td>Dabarera, Carol</td>
<td>76</td>
</tr>
<tr>
<td>Damböck, Barbara</td>
<td>33</td>
</tr>
<tr>
<td>Davidson, Simon John</td>
<td>89</td>
</tr>
<tr>
<td>Dawson-Smith, Mark</td>
<td>33</td>
</tr>
<tr>
<td>De Backer, Fauve</td>
<td>87</td>
</tr>
<tr>
<td>de la Caridad Collada Peña, Ivonne</td>
<td>98</td>
</tr>
<tr>
<td>Deygers, Bart</td>
<td>88</td>
</tr>
<tr>
<td>Dimova, Slobodanka</td>
<td>86</td>
</tr>
<tr>
<td>Dixon, Daniel</td>
<td>33</td>
</tr>
<tr>
<td>Drennan, Laura</td>
<td>50</td>
</tr>
<tr>
<td>Dunn, Karen</td>
<td>70, 90</td>
</tr>
<tr>
<td>East, Martin</td>
<td>107</td>
</tr>
<tr>
<td>Eberharter, Kathrin</td>
<td>90</td>
</tr>
<tr>
<td>Elder, Cathie</td>
<td>48</td>
</tr>
<tr>
<td>Fan, Jason</td>
<td>85, 108</td>
</tr>
<tr>
<td>Fang, Felicia</td>
<td>116</td>
</tr>
<tr>
<td>Friginal, Eric</td>
<td>74</td>
</tr>
<tr>
<td>Frost, Kellie</td>
<td>73</td>
</tr>
<tr>
<td>Galacz, Evelina</td>
<td>59</td>
</tr>
<tr>
<td>Gottlieb, Margo</td>
<td>100</td>
</tr>
<tr>
<td>Grant, Rosalie</td>
<td>115</td>
</tr>
<tr>
<td>Grapin, Scott</td>
<td>74</td>
</tr>
<tr>
<td>Green, Tony</td>
<td>57</td>
</tr>
<tr>
<td>Grönlund, Anna</td>
<td>119</td>
</tr>
<tr>
<td>Gu, Peter</td>
<td>63</td>
</tr>
<tr>
<td>Guan, Xiaoxian</td>
<td>107</td>
</tr>
<tr>
<td>Guedez-Fernandez, Romulo</td>
<td>112</td>
</tr>
<tr>
<td>Hagiwara, Masato</td>
<td>95</td>
</tr>
<tr>
<td>Hama, Mika</td>
<td>106</td>
</tr>
<tr>
<td>Hanlon, Sean</td>
<td>58</td>
</tr>
<tr>
<td>Harding, Luke</td>
<td>45, 48</td>
</tr>
<tr>
<td>Harradine, Owen</td>
<td>48</td>
</tr>
<tr>
<td>Harsch, Claudia</td>
<td>49, 71, 98</td>
</tr>
<tr>
<td>He, Lianzhen</td>
<td>55, 91</td>
</tr>
<tr>
<td>Hill, Mary</td>
<td>107</td>
</tr>
<tr>
<td>Holzknecht, Franz</td>
<td>90</td>
</tr>
<tr>
<td>Homayounzadeh, Maryam</td>
<td>102</td>
</tr>
<tr>
<td>Hsieh, Ching-Ni</td>
<td>82</td>
</tr>
<tr>
<td>Hsu, Wei-Li</td>
<td>112</td>
</tr>
<tr>
<td>Hynninen, Niina</td>
<td>46</td>
</tr>
<tr>
<td>Inbar-Lourie, Ofra</td>
<td>47</td>
</tr>
<tr>
<td>Iwashita, Noriko</td>
<td>62</td>
</tr>
<tr>
<td>Jenkins, Jennifer</td>
<td>47</td>
</tr>
<tr>
<td>Jin, Yan</td>
<td>54, 65, 85</td>
</tr>
<tr>
<td>Jølle, Lennart Jøran</td>
<td>68</td>
</tr>
<tr>
<td>Jun, Henry</td>
<td>83</td>
</tr>
<tr>
<td>Kane, Michael T.</td>
<td>53</td>
</tr>
<tr>
<td>Kashimada, Yuko</td>
<td>106</td>
</tr>
<tr>
<td>Khamboonruang, Apichat</td>
<td>118</td>
</tr>
<tr>
<td>Khan, Kamran</td>
<td>48</td>
</tr>
<tr>
<td>Kim, Ahyoung Alicia</td>
<td>81</td>
</tr>
<tr>
<td>Kim, Hyunah</td>
<td>67</td>
</tr>
<tr>
<td>Knoch, Ute</td>
<td>48, 67</td>
</tr>
<tr>
<td>Kong, Jufang</td>
<td>108</td>
</tr>
<tr>
<td>Kong, Xiaoxiao</td>
<td>62</td>
</tr>
</tbody>
</table>
Konrad, Eva, 90  
Kremmel, Benjamin, 75, 90  
Kudo, Yoji, 106  
Lam, Daniel, 59  
Larson, Elizabeth Jean, 67  
Lau, Clarissa, 78  
Lavagno, Antonella, 71  
Lee, Shinhye, 64, 81  
Lee, Yong-Won, 83  
Lei, Yuyun, 69  
Leung, Constant, 47  
Li, Naihsin, 60  
Liardet, Cassi, 50  
Lim, Gad, 97  
Liu, Huan, 105  
Liu, Jianda, 55  
Liu, Li, 115  
Liu, Xiaohua, 73  
Lo Bianco, Joseph, 39, 45  
Luo, Lan, 92  
Luo, Shaqian, 105  
MacGregor, David, 97  
Macqueen, Susy, 67, 120  
Malone, Margaret, 74  
Manna, Venessa, 102, 103  
May, L. D. Nicolas, 97  
May, Lyn, 59  
McCray, Gareth, 90  
McLain, Patrick M., 97  
McPherson, Judit, 111  
Min, Shangchao, 91, 118  
Monfils, Lora F, 103  
Moss, Pamela, 41  
Müller-Karabil, Anika, 49  
Nakatsuara, Fumiyo, 59  
Negishi, Masashi, 106  
Nguyen, Huu, 114  
Ockey, Gary John, 33, 79  
O’Connell, Stephen P., 97, 101  
O’Hagan, Sally, 67  
Okabe, Yasuko, 106  
O’Sullivan, Barry, 56, 90  
Palmer, Adrian, 33  
Pan, Mingwei, 108  
Papageorgiou, Spiros, 52  
Phakiti, Aek, 37, 109  
Phillips, Christine, 120  
Plakans, Lia, 93  
Plough, India, 62  
Purpura, James, 33, 34  
Raquel, Michelle Reyes, 103, 116  
Read, John, 33, 50  
Rendle-Short, Johanna, 120  
Riazi, Mehdi, 36, 50  
Rivera Pérez, Santiago Jorge, 98  
Rocca, Lorenzo, 49  
Rummel, Stephanie, 111  
Russell, Bruce, 67  
Saville, Nick, 56  
Sawaki, Yasuyo, 70  
Saygi, Sukran, 116  
Schmidgall, Jonathan, 54  
Sebolai, Kabelo, 50  
Seedhouse, Paul, 66  
Seguis, Brigita, 99  
Selçuk, Merve, 114  
Seong, Yuna Patricia, 60  
Settles, Burr, 95  
Sewell, Andrew, 46  
Seyferth, Sibylle, 71  
Shi, Yali, 110  
Shih, Chilin, 69  
Shohamy, Elana, 46  
Skar, Gustaf Bernhard Uno, 68  
Slade, Diana, 120  
Song, Jayoung, 84, 112  
Spiby, Richard David, 70  
Spöttl, Carol, 90  
Sun, Youxia, 107  
Suvorov, Ruslan, 110  
Taalas, Peppi, 49, 119  
Tannenbaum, Richard J., 53  
Tolentino, Florencia, 82  
Trace, Jonathan William, 76, 93  
Tsutagawa, Fred S., 65  
Turner, Carolyn E., 34, 43  
Umakoshi, Yuko, 106  
Urmston, Alan, 116  
Van Moere, Alistair, 58  
van Naerssen, Margaret, 104  
Villa Larenas, Salome, 111  
Vincett, Megan, 67, 78  
Wakeland, Laura, 103  
Wallace, Matthew, 86  
Wang, Jun, 117  
Wang, Min, 118
Wang, Yun, 82
Watanabe-Kim, Izumi, 113
Wattam, Craig, 67
Weideman, Albert, 50
Wigglesworth, Gillian, 73
Wilmes, Carsten, 81, 97
Wu, Chiao-Yi, 96
Wu, Jessica, 60
Wu, Sha, 56
Xie, Qin, 95
Xiong, Lidi, 118
Xu, Jing, 61
Yalçın Çolakoğlu, Özlem, 114
Yan, Xun, 69
Yao, Wenjing, 100
Ye, Zhengdao, 120
Yeom, Soohye, 83
Yoo, Hanwook, 102
Yu, Han, 54
Yun, Jiani, 107
Zehentner, Matthias, 90
Zhang, Barbara Ying, 99
Zhang, Limei, 77
Zhang, Lin, 65
Zhang, Wenxia, 56
Zhang, Yumei, 105
Zhou, Jiming, 78
Zhu, Qiyun, 82
Can’t get enough of LTRC? Our next conference is only 8 months away!

LTRC 2019
41st Language Testing Research Colloquium
4-8 March 2019  Atlanta, Georgia USA
Language Testing and Social Justice

- Convenient all-in-one location at the Decatur Marriott Courtyard
- Easily accessible by public transportation from airport/downtown Atlanta
- Conference venue in walkable district known for restaurants and shopping
- Everything you’ve come to expect from LTRC, including cutting-edge
  pre-conference workshops, inspiring plenaries and papers, and networking
  opportunities
DELNA
Diagnostic English Language Needs Assessment

The University of Auckland's post-entry language assessment

The DELNA Screening is a quick measure of students’ language competence comprising vocabulary and timed reading tasks.

The Screening is:

- computer-based
- easy to administer
- cost effective
- available for licensing

For more information about DELNA or to license the assessment:
www.delna.auckland.ac.nz
delna@auckland.ac.nz

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

Lancaster University has a world-wide reputation for excellence in research on language testing and assessment.

We offer a range of postgraduate degrees delivered on-campus or by distance.

- 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 a summer school in language testing.

Further information:
www.lancaster.ac.uk/linguistics/study

Contact: postgraduatelinguistics@lancaster.ac.uk | Tel: +44 (0)1524 593028 | LU_LanguageTesting
Language Assessment for Classroom Teachers

Lyle Bachman | Barbara Damböck

This book provides teachers with an entirely new approach to developing and using classroom-based language assessments, with step-by-step guidance and written exercises. This approach is based on current theory and practice in the field of language assessment and on an understanding of the assessment needs of classroom teachers.

Purchase this title from UBIQ at ubiq.co.nz
DELTA is an English language proficiency measurement system for post-secondary students studying in English-medium institutions. Through DELTA, students can diagnose their strengths and weaknesses in reading, listening, grammar and vocabulary; track their English language gains; and plan their English language learning.

DELTA is a fully integrated system for:
- test enrolment and scheduling
- test delivery and scoring
- diagnostic reporting and tracking of learner progress
- item banking and test construction
- proficiency and progress monitoring for teachers and administrators

The DELTA diagnostic report and tracking system enables learners to monitor their progress in achieving their targets for improving their English to help them progress in their academic learning.

For more information about DELTA, you can go to www.deltatesting.net or contact the DELTA Team at ecltu@polyu.edu.hk or (852) 34003318.
We are Michigan Language Assessment. A nonprofit joint venture between Cambridge Assessment English and the University of Michigan, we build on a long history of excellence in applied linguistics and language testing to deliver a range of fit-for-purpose English language tests.

**Selected Research Publications**

Valuing the MET speaking test through phraseological analysis: A corpus approach to language assessment.  
*CaMLA Working Papers* (in press) Ann Arbor, MI.  
Michigan Language Assessment

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

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

Fairness and Bias in Language Assessment.  
In D. Tsagari & J. Banerjee (Eds.).  

Investigating Lexico-grammatical Complexity as Construct Validity Evidence for the ECPE Writing Tasks: A Multidimensional Analysis.  
*CaMLA Working Papers* 2016(1) Ann Arbor, MI.  
Cambridge Michigan Language Assessments

Predicting Listening Item Difficulty with Language Complexity Measures: A Comparative Data Mining Study.  
*CaMLA Working Papers* 2014(2) Ann Arbor, MI.  
Cambridge Michigan Language Assessment

The Characteristics of the Michigan English Test Reading Texts and Items and their Relationship to Item Difficulty.  
*CaMLA Working Papers* 2015(2) Ann Arbor, MI.  
Cambridge Michigan Language Assessment

Linking the Common European Framework of Reference and the MET Writing Test.  
*CaMLA* (2014)  
*CaMLA Technical Report* Ann Arbor, MI.  
Cambridge Michigan Language Assessment

A look into cross-text reading items: Purpose, development and performance.  

The Relative Difficulty of Dialogic and Monologic Input in a Second-Language Listening Comprehension Test.  
*Language Assessment Quarterly*, 9(4), 375–397

*CaMLA* (2007)  
*CaMLA Technical Report* Ann Arbor, MI.  
Cambridge Michigan Language Assessment

**Research & Internship Opportunities**

Michigan Language Assessment conducts and supports research to inform test development and revision, to provide evidence of quality and validity for its tests, and to contribute to knowledge in the field of language testing. Visit our website to learn more about summer internships in language assessment and about funding opportunities under the Spaan Research Grant Program.
Fumiyo Nakatsuhara, Chihiro Inoue, Vivien Berry and Evelina Galaczi

Exploring performance across two delivery modes for the same L2 speaking test: Face-to-face and video-conferencing delivery

A preliminary comparison of test-taker and examiner behaviour

available now at ielts.org/research
The Discourse of the IELTS Speaking Test

Interactional design and practice

Paul Seedhouse and Fumiyo Nakatsuhara

hear Paul discuss his findings at LTRC, Wednesday 4 July 11.20-12.10
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.


ETS — Measuring the Power of Learning.

www.ets.org/toefl/research
The fair and secure English test

**PTE Academic** is the leading computer-based English test for international study and migration. Results are accepted globally by thousands of academic programs around the world.

**Supporting research in language assessment**

We are conducting an expanded research program that collaborates with independent external researchers in universities all over the world. Faculty members, doctoral students as well as qualified freelance researchers are invited to submit proposals.

Learn more about the research that underpins **PTE Academic** and our next Research Call opening in November 2018 by visiting [www.pearsonpte.com/research](http://www.pearsonpte.com/research).