38th Annual Language Testing Research Colloquium
June 20–24, 2016
Palermo, Sicily (Italy)

LANGUAGE CONSTRUCTS, CONTEXTS, AND CONTENT IN CLASSROOM AND LARGE-SCALE ASSESSMENTS

ILTA
INTERNATIONAL LANGUAGE TESTING ASSOCIATION

UNIVERSITÀ DEGLI STUDI DI PALERMO

21
Università degli Studi di Palermo 1906-2016
PALERMO CITTÀ UNIVERSITARIA
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.


http://www.ets.org/toefl/research
# TABLE OF CONTENTS

Welcome Message from the President 5

Welcome Message from the Chairs 6

The Samuel J. Messick Memorial Lecture 8

The Davies Lecture 10

The Cambridge/ILTA Distinguished Achievement Award 12

Conference Support 14

ILTA 2016 15

Award Winners 16

Colloquium Venues 18

LTRC 2016 Conference Overview 21

LTRC 2016 Conference Schedule 25

Pre-Conference Workshop Abstracts 39

Paper Abstracts: Tue, June 21 49

Paper Abstracts: Wed, June 22 51
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paper Abstracts: Thu, June 23</td>
<td>68</td>
</tr>
<tr>
<td>Paper Abstracts: Fri, June 24</td>
<td>86</td>
</tr>
<tr>
<td>Symposium Abstracts</td>
<td>103</td>
</tr>
<tr>
<td>Poster Abstracts</td>
<td>119</td>
</tr>
<tr>
<td>Work-in-Progress Abstracts</td>
<td>145</td>
</tr>
<tr>
<td>Demo Abstracts</td>
<td>167</td>
</tr>
<tr>
<td>Index of Presenters</td>
<td>178</td>
</tr>
</tbody>
</table>
Welcome to LTRC 2016

Benvenuti in Sicilia for the 38th Language Testing Research Colloquium on the theme of Language Constructs, Contexts, and Content in Classroom and Large-Scale Assessments. This year, the context, at the crossroads of the Mediterranean world, is, appropriately enough, the island home of Archimedes: the grandfather of measurement. The attractions of the hospitable climate, cosmopolitan Sicilian cuisine, and Palermo world heritage site surrounding us at the conference will surely only be displaced by the volume of eureka moments inside the conference hall.

It is an encouraging sign of the growing reach of our organisation that we find ourselves in Italy for the first time and that next year we will cross the Atlantic to another exciting new location: Bogotà in Colombia.

This year’s conference theme expresses the broadening interests of ILTA members with increased attention to the place of assessment in the classroom, alternative intellectual perspectives on theory and practice, and the importance of preserving and reflecting on the history of the development of our field.

Our three featured speakers all touch on these themes. James Pellegrino (The Samuel J. Messick Memorial Lecture Award) discusses instructionally relevant assessments, John de Jong (ILTA/Cambridge Distinguished Achievement Award) looks back over his career and forward to a new global standard, and John Norris (The Alan Davies Lecture) explores the familiar question of the interface with SLA research through new interdisciplinary perspectives. The conference also gives us an opportunity to share our memories of Alan Davies himself, who passed away earlier this year, a founding father of ILTA and an inspirational teacher, mentor and friend to many of us.

In total, we received 368 submissions. At the time of writing, the programme includes forty-five research papers, four symposia, twenty-three posters and twenty works-in-progress. Palermo sees the introduction of an innovative presentation format: the demo. Reflecting the growing role of new technology in language assessment practices, the nine demos in Palermo profile an intriguing range of software tools with the potential to improve assessment practice. Directly before the colloquium, five workshops are featured, one of which will be conducted in Italian—the second time (after Korea) that an LTRC workshop is given in a language other than English.

Of course, constructing the context for and content of a large-scale conference is a feat of engineering that would tax even the great Archimedes himself. We owe special thanks to our hosts at the University of Palermo, the LTRC co-chairs: James Purpura and Monica Barni and to the local chairs, Carla Bagna and Mari D’Agostino. Thanks are also due to the other members of the organizing committee: Veronica Benigno, Jee Wha Dakin, Barbara Dobson, Francesca Gallina, Kirby Grabowski, Giuliana Grego Bolli, Heidi Han-Ting Liu, Lucilla Lopriore, Sabrina Machetti, Paola Masillo, Saerhim Oh, and Lorenzo Rocca. In addition, we are grateful to our student volunteers. And finally, thanks to Erik Voss, who has made an invaluable contribution as webmaster.

Whether you are an ILTA veteran or this is your first time at LTRC, I do hope you will find inspiration and pleasure in Palermo in equal measure.

Best wishes,

Tony Green
ILTA President
Welcome Message from the Chairs

Welcome to LTRC 2016 in Palermo

As the co-chairs of LTRC 2016, we would like to warmly welcome you to Palermo (Palermu in Sicilian; Panormus in Latin; Panormos in Greek; Batarm in Arabic; and Ziz—flower, in Phoenician). The idea to take LTRC to Italy, and specifically to Palermo, was hatched over a beautiful lunch, attended by Giuliana Grego Bolli (Perugia), Lucilla Lopriore (Rome), Marc Silver (Modena), and us (Monica – Siena and Jim – NYC of Sicilian parents) during a language testing meeting in Bologna back in 2012. The semi-innocent question, ‘Why not an LTRC in Italy’ spawned a very lively discussion, made probably even livelier by a glass (or more) of a dry, ruby-red Sangiovese. The Italians all agreed that it was very timely to have LTRC in Italy, especially since language assessments are now increasingly used not only in the Italian educational system on the elementary, secondary, and university levels, but also in Italian society at large—e.g., Italian language testing for immigrants. It was further noted that a conference of this sort was important because there has never been a tradition of formal academic study and research in Italy on the topic of assessment. As a result, more and more tests are produced with little consideration of validity, reliability, impact, or the social consequences of tests. This lack of a formal assessment tradition, therefore, has produced debates that end with opposing views about tests: those in favor and those against. Such debates, however, seldom reflect an understanding of assessment as technically sound and meaningful, providing trustworthy and useful information about examinees, or flawed for any number of reasons, producing dubious information, or debates showing an awareness that tests can be used effectively in decision-making or misused for social, cultural, economic, or political purposes.

During our Bologna conversation we also tried to find a location for the conference. Naturally, all the Italians proposed their own university or city as the best! However, this restrained volcano soon led, fortunately, to a spontaneous proposal to take LTRC to the island of Sicily, famous for her architecture, bearing witness to the aesthetics of the Phoenetician, Carthaginian, Greek, Roman, Byzantine, Arab, Norman, German, French, and Spanish cultural kaleidoscope; also famous for her multilayered cultural tapestry, emotionally held distinct after three thousand years under the thumbs of invaders, and for her complex and intermingled gastronomy referred to by some as the culinary crossroads of the Mediterranean—and one we will all experience. Finally, important to linguists, we had a proposal to go to an island, renowned for its alive but almost secretive language (Sicilianu) with its colorful metaphors, pithy proverbs, songs, and a lexical diversity that embodies the linguistic history of the Mediterranean, where in the 13th century, the Sicilian School of Poetry was acclaimed for its lyricism and idolized image of courtly love, and recognized as such by Dante in his De Vulgari Eloquentia (On Eloquence in the Vernacular). With all this, how could we not take LTRC to Palermo, the rough-appearing, cacophonous, seemingly confusing, always interesting, and mostly charming warren of twisting streets, with names like Via Terra delle Mosche (Land-of-the-Flies Street) or Via delle Sedie Volanti (Flying-Chairs Street), and with ineffaceable images of a once vibrant city on a hurried rebound?
The decision to go to Palermo was, indeed, a challenge (and at times a crazy idea) for all of us to try to organize a conference far from our homes. But, in the end, we hope you will agree with our decision and will enjoy this beautiful city.

So, during this week in Palermo, we wish for you an intellectually meaningful and stimulating LTRC, a warm reawakening of long-held friendships, a blossoming of new relationships, and a joyful, cannoli-sated stay in Palermo. Remembering the words of the Sicilian novelist, Leonardo Sciascia, in *The Council of Egypt* (1963)...

The Abbott ventured out of the house. He has himself driven in a carriage through the city. It was one of those iridescent Palermo mornings, the clouds a shelving deep blue and russet. He rejoiced in the sun, the air, the warm Norman stone, the red Arabic cupolas, the aroma of seaweed and lemon in the market...

**Jim Purpura** and **Monica Barni**
A Framework for Defining and Evaluating the Validity of Instructionally Relevant Assessments

Numerous developments in the fields of cognitive psychology, educational psychology, the learning sciences, and educational measurement have begun to reshape the field of educational assessment. Among these is the realization that assessment is a process of reasoning from evidence that needs to be guided by theories, models, and data on the nature of knowledge representations and the development of competence in typical domains of classroom instruction. Other key understandings include the multiple purposes assessment may serve in the educational system, including formative, summative, and evaluative functions, and that the design of any assessment must be optimized for the intended purpose. Coinciding with these developments is contemporary thought that the validity of a given assessment should be construed as an argument consisting of claims related to an assessment’s intended interpretive use and the forms of evidence that provide the warrants in support of those claims.

Much of the literature on assessment validity has concentrated on assessments designed for large-scale and often high-stakes uses in the educational system including various policies related to accountability. In contrast, increasing attention is being paid to the need for systems of assessments, with an emphasis on the design and use of assessments intended to function much closer to the processes of teaching and learning. Many such assessments are tied more closely to detailed analyses of the instructional domain (e.g., reading, mathematics, science, history) and theories, models and data regarding student knowledge representations and how those change over time as a consequence of instruction.

Despite a significant increase in efforts to frame discussions of assessment in terms of domain-based theories and models of cognition and learning, and the use of assessment in the classroom to support various teaching and learning functions, there is a paucity of discussion about the meaning of validity for such assessments. This presentation provides a framework for considering multiple aspects of validity that are applicable to assessments intended to function at the classroom level to support ongoing processes of teaching and learning. The goal is to explore the concept of validity as it applies to such assessments, including the nature of the types and forms of evidence needed to construct a validity argument for such assessments. The framework encompasses ideas related to cognitive validity, instructional validity, and inferential validity. In addition to describing each of these
components of validity we also consider the multiple forms of evidence that might be relevant to support claims about any given assessment’s validity as related to each of the three components. Examples will be provided of application of the evidentiary reasoning framework and the presentation will conclude with a consideration of issues regarding application of the validity analysis framework, including the need for careful and consistent application of this logic in the design of integrated curriculum, instructional and assessment resources. Such resources are critical to support teachers and students in attaining the deeper learning goals intended by the Common Core Standards in Mathematics and English Language Arts and the Next Generation Science Standards.

James W. Pellegrino is Liberal Arts and Sciences Distinguished Professor and Distinguished Professor of Education at the University of Illinois at Chicago. He also serves as Co-director of UIC’s interdisciplinary Learning Sciences Research Institute. His research and development interests focus on children’s and adult’s thinking and learning and the implications of cognitive research and theory for assessment and instructional practice. He has published over 300 books, chapters and articles in the areas of cognition, instruction and assessment. His current research is funded by the National Science Foundation and the Institute of Education Sciences. He has served as head of several National Academy of Sciences study committees, including co-chair of the Committee on Learning Research and Educational Practice, and co-chair of the Committee on the Foundations of Assessment which issued the report Knowing What Students Know: The Science and Design of Educational Assessment. Most recently he served as a member of the Committee on Science Learning: Games, Simulations and Education; as chair of the Committee on Defining Deeper Learning and 21st Century Skills; and co-chair of the Committee on Developing Assessments of Science Proficiency in K-12. He is a past member of the Board on Testing and Assessment of the National Research Council and a lifetime member of the National Academy of Education. He currently serves on the Technical Advisory Committees (TAC) of several states as well as the TACs of the SBAC, PARCC, DLM, and NCSC consortia of states funded under the USDOE Race to the Top assessment initiative.
Reframing the SLA-Assessment Interface: ‘Constructive’ Deliberations at the Nexus of Interpretations, Contexts, and Consequences

Applied research in the fields of second language acquisition and language testing (should) share much in common. At stake for both are concerns with understanding how and the extent to which languages are learned, as well as what aspects of individuals (identity, aptitude, etc.) and of contexts (physical, social, etc.) shape language learning and use. These overlapping interests have been recognized, for example by Davies (2007): “The role of language testing research as an activity of applied linguistics is to further our understanding of language learning and illuminate the still uncharted space of language use” (p. 30). However, although important forays at the interface have occurred (e.g., Bachman & Cohen, 1998), to date the potential synergistic outcomes of collaborative SLA-assessment research remain largely unfulfilled. By and large, Davies’ (1999) comment in the Dictionary of Language Testing still reflects the status quo: “Attempts have been made to bring work in SLA and language testing together but so far without much success” (p. 176).

In this talk, I will suggest that a handful of high-priority interpretive problems, which go to the heart of ‘understanding language learning and illuminating language use’, have coalesced very recently to the extent that constructive collaboration between the disciplines has become an intellectual requirement. Such problems, common to assessments used in L2 research as well as applied testing, include: (a) the confusing relationship between L2 development, proficiency, and performance; (b) the complicated interaction of task design with learner cognition and language use; and (c) the important roles played by complex constructs like interactional or intercultural competence, as well as the so-called ‘CALF’ variables. Improved construct definition and operationalization are key collaborative steps in resolving these and similar problems (Norris & Ortega, 2003, 2012), and I will highlight recent work by SLA-assessment researchers who have drawn on cross-disciplinary knowledge and methods to do so.

However, adequate validity arguments and supporting evidence are in and of themselves insufficient. That is, given that language tests are inherently “social facts” (Davies, 1999)—even those employed for research/theoretical purposes—their warranted use is further challenged by both the contexts within which they occur and the consequences that ensue for individuals and communities. For one example, the rising prominence of learning-oriented assessment calls for much
more than just ‘good measurement’. Such assessments are deployed as intentional and integral components of L2 educational designs, thereby begging the question of their relationship with instructed language acquisition, and their use impacts educational programs and stakeholders alike. In this and similar cases, I also argue that SLA and language testing researchers must engage much more deliberately with the ‘big picture’ of how assessments are actually used in context, a fitting heuristic for which is validity evaluation (Norris, 2008). I will provide examples of researchers engaging in validity evaluation that attempts to deal with new construct spaces and heretofore unrecognized consequences, specifically as these evolve in contemporary L2 instructional environments, and I will point to future work that is sorely needed at the SLA-assessment interface.

John Norris is a Principal Research Scientist at ETS. His interests include educational assessment, program evaluation, language pedagogy (task-based language teaching in particular), and research methods. He has taught language and applied linguistics courses, and consulted on assessment, evaluation, curriculum development, and teacher education projects, in Belgium, Brazil, Germany, Japan, Spain, and around the U.S. Prior to joining ETS, he taught at Georgetown University and the University of Hawai‘i, and he served as assessment specialist at Northern Arizona University. At Georgetown University, he was the founding director and principal investigator for the federally funded “Assessment and Evaluation Language Resource Center”. John’s most recent books explore the topics of outcomes assessment (Student learning outcomes assessment in college foreign language programs), evaluation (Innovation and accountability in language program evaluation), and research methods (Improving quantitative reasoning in second language research). He has served as chair of the TOEFL Committee of Examiners and founding chair of the International Consortium on Task-Based Language Teaching. John speaks German, Spanish, Portuguese (and a little Japanese and French), and he is an avid runner/hiker/surfer.
Modelling language competence into a global framework:
Taking the past into the future

At this occasion of receiving the 2016 Distinguished Achievement Award, I would like to talk about my lifetime in testing, highlighting where my development and that of our field progressed in parallel.

In 1977 I started my testing career at CITO, the Dutch national institute for educational measurement, as a coolie item writer. In 1981 I attended the AILA conference in Lund, my first international conference. I hadn’t dared to send in a proposal, but I met Alan Davies and mentioned my 1979 internal research report on validating listening comprehension tests and he encouraged me to present my work at future conferences. Criticism on this report came from CITO’s psychometric department: I had compared changes in test scores using p-values (proportion correct) and I should have applied latent trait theory or at least transformed the p-values using Fischer’s z. I was advised to read Best Test Design by Wright and Stone. That really got me started on scaling tests designed for different levels and assessing very basic European language skills for workers in international transport leading to the definition of Breakthrough in the LangCred project. In 1982 I reported on this work at international measurement conferences where participants were open to latent trait theory. However, at the Language Testing Forum in Reading, UK, I was met with a downright rejection of unidimensionality and the Rash model. I found myself in two different professional networks.

Interest in defining increasing functional language proficiency originated in 1950 in the USA with the ILR scale developed by the Foreign Service Institute. In Europe, the Threshold level (1975) was to become the cornerstone of the Common European Framework of Reference for Languages (CEFR) and in Australia David Ingram developed the Australian Second Language Proficiency Ratings (ASLPR, 1984, later renamed as ISLPR). Among these the CEFR has undoubtedly become the most widely used. However, researchers and practitioners have pointed to a number of essential limitations of the CEFR.

In this talk I will disentangle some misunderstandings about the CEFR, discuss the limitations of the framework and focus on the psychometric definition of being at a level. One of the weaknesses of the CEFR, for example, is that its presentation of levels invites its interpretation as a set of separate and discrete levels and
therefore misrepresents the real pace and nature of language learning, which is in fact a continuous process albeit with occasional hiccups. Another limitation is that its set of descriptors is very sparse, severely restricting the applicability of the CEFR in the many different contexts of language use.

I will then provide arguments for the usefulness of a global standard, as a basis for creating a common rationale for instructional and testing materials, and for learning, teaching, and assessment practices in general. I will argue that in the current publishing panorama of CEFR-aligned learning, teaching, and testing resources, only a small percentage of them clearly reflect an operationalization of the CEFR levels.

John H. A. L. De Jong, Senior Vice President Global Assessment Standards with Pearson and Professor of Language Testing at VU University Amsterdam. John graduated in General Linguistics, French and English languages from Leiden University and obtained a Ph.D. in Educational Measurement from Twente University. He has almost 40 years’ experience in language testing and educational measurement. He has published widely on language assessment and educational measurement and has specialised in empirical scaling and the development of internationally standardised reporting scales of language proficiency. Before starting his career in language testing John had been teaching French for seven years in secondary schools. He joined Pearson in September 2006 where apart from directing language testing programmes he is also Programme Director for developing the Frameworks for PISA 2015, 2018 and PISA for Development. He has been involved in consultancy with many countries such as Azerbaijan, China, Italy, Japan, Kazakhstan, Malaysia, Mexico, Turkey, Peru.
LTRC Organizing Committee 2016

Co-Chairs
- Monica Barni, University for Foreigners of Siena
- James E. Purpura, Teachers College, Columbia University

Local Chairs
- Carla Bagna, University for Foreigners of Siena
- Mari D’Agostino, University of Palermo

Other Members of the Organizing Committee
- Veronica Benigno, Pearson
- Jee Wha Dakin, ETS
- Barbara Dobson, CamLA
- Francesca Gallina, University for Foreigners of Siena
- Kirby Grabowski, Teachers College, Columbia University
- Giuliana Grego Bolli, University for Foreigners of Perugia
- Heidi Han-Ting Liu, Teachers College, Columbia University
- Lucilla Lopriore, University of Rome
- Sabrina Machetti, University for Foreigners of Siena
- Paola Masillo, University for Foreigners of Siena
- Lorenzo Rocca, University for Foreigners of Perugia

Webmaster
- Erik Voss, Northeastern University

LTRC 2016 Program Book Designer
- OL3 srl

LTRC 2016 Program Preparation Support
- Jee Wha Dakin
- Francesca Gallina
- Kirby Grabowski
- Paola Masillo

LTRC 2016 Session chairs
- Poster Chairs: Barbara Dobson & Lucilla Lopriore
- WIP Chairs: Barbara Dobson & Lorenzo Rocca

Abstract Reviewers
- Lyle Bachman
- Jayanti Banerjee
- Khaled Barkaoui
- Tineke Brunfaut
- Micheline Chalhoub-Deville
- Yeonsuk Cho
- Ikkyu Choi
- Martyn Clark
- Deborah Crusan
- Alister Cumming
- Larry Davis
- John H.A.L. de Jong
- Nivja de Jong
- Dan Douglas
- Catherine Elder
- Tim Farnsworth
- Evelina D. Galaczi
- Ardeshir Geranpayeh
- April Ginther
- Tony Green
- Giuliana Grego Bolli
- Liz Hamp-Lyons
- Luke Harding
- Ching-Ni Hsieh
- Yo In’nami
- Talia Isaacs
- Eunice Eunhee Jang
- Dorry M. Kenyon
- Mkyung Kim Wolf
- Ahyoung (Alicia) Kim
- Ute Knoch
- Yong-Won Lee
- Constant Leung
- Gad Lim
- Lorena Llosa
- Sari Luoma
- Margaret E. Malone
- Tim McNamara
- Fumiyo Nakatsuhara
- Barry O’Sullivan
- Gary Ockey
- Spiros Papageorgiou
- Aek Phakiti
- Lia Plakens
- India Plough
- John Read
- Lorenzo Rocca
- Nick Saville
- Yasuyo Sawaki
- Jonathan Schmidgall
- Rob Schoonen
- Youngsoo So
- Charles Stansfield
- May Tan
- Veronika Timpe Laughlin
- Dina Tsagari
- Carolyn Turner
- Alan Urmston
- Alistair Van Moere
- Erik Voss
- Elvis Wagner
- Sara Cushing
- Jill Wigginsworth
- Xiaoming Xi
- Yoshinori (Josh) Watanabe
- Guoxing Yu

Volunteers
- Martina Bellinzona
- Jorge Beltrán
- Martina Mannu
- Yuna Patricia Seong
- Michelle Stabler-Havener
- Fred Tsutagawa

Language Constructs, Contexts, and Content in Classroom and Large-Scale Assessments
ILTA Executive Board
2016

President: Anthony (Tony) Green, University of Bedfordshire, UK
Vice-President: Catherine (Cathie) Elder, University of Melbourne, Australia
Secretary: Margaret E. Malone, Center for Applied Linguistics, USA
Treasurer: Jayanti Banerjee, Worden Consulting LLC, USA
Immediate Past President: Dan Douglas, Iowa State University, USA

Members at Large
• Kirby Grabowski, Teachers College, Columbia University, USA
• Eunice Eunhee Jang, OISE, University of Toronto, Canada
• Beverly Baker, University of Ottawa, Canada
• Talia Isaacs, University of Bristol, UK

LTRC 2016 Co-Chairs
• James E. Purpura, Teachers College, Columbia University, USA
• Monica Barni, University for Foreigners of Siena, Italy

ILTA Nominating Committee 2016
• Chair: Mikyung Kim Wolf, ETS, USA
• Diane Schmitt, Nottingham Trent University, UK
• Dina Tsagari, University of Cyprus, Cyprus
• Erik Voss, Northeastern University, USA

Award Committees
Cambridge/ILTA Distinguished Achievement Award
• Chair: Tim McNamara, University of Melbourne, Australia
• Evelin Galaczi, Cambridge English Language Assessment, UK
• Kirby Grabowski, Teachers College, Columbia University, USA
• Tony Green, University of Bedfordshire, UK

Sage/ILTA Book Award
• Chair: Annie Brown, Australian Council for Educational Research, Australia
• Slobodanka Dimova, University of Copenhagen, Denmark
• Jayanti Banerjee, Worden Consulting LLC, USA
• Ute Knoch, University of Melbourne, Australia
• Beverly Baker, University of Ottawa, Canada

ILTA Student Travel Awards
• Chair: Tony Green, University of Bedfordshire, UK
• Eunice Eunhee Jang, OISE, University of Toronto, Canada
• Beverly Baker, University of Ottawa, Canada
• Talia Isaacs, University of Bristol, UK

ILTA Workshops and Meetings Award
• Chair: Tony Green, University of Bedfordshire, UK
• Eunice Eunhee Jang, OISE, University of Toronto, Canada
• Beverly Baker, University of Ottawa, Canada

Lado Award (to be awarded at LTRC)
• Chair: Maryam Wagner, Carleton University, Canada
• Dorry M. Kenyon, Center for Applied Linguistics, USA
• Benjamin Kremmel, University of Nottingham, UK
• Talia Isaacs, University of Bristol, UK

ILTA Best Article Award 2014
• Chair: Claudia Harsch, University of Bremen, Germany
• Kirby Grabowski, Teachers College, Columbia University, USA
• Luke Harding, Lancaster University, UK
• Elvis Wagner, Temple University, USA
• Paul Deane, ETS, USA
**Award winners**

**Samuel J. Messick Memorial Lecture Award**
James Pellegrino, Learning Sciences Research Institute, University of Illinois at Chicago
Lecture title: *A Framework for Defining and Evaluating the Validity of Instructionally Relevant Assessments*

**The Davies Lecture Award**
John Norris, ETS
Lecture title: *Reframing the SLA-Assessment Interface: ‘Constructive’ Deliberations at the Nexus of Interpretations, Contexts, and Consequences*

**Cambridge/ILTA Distinguished Achievement Award**
John H. A. L. de Jong, Senior Vice President Global Assessment Standards, Programme Director PISA 2015 & 2018 Framework Development Pearson, Professor of Language Testing at VU University Amsterdam
Lecture title: *Modelling Language Competence into a Global Framework: Taking the Past into the Future*

**Robert Lado Memorial Award**
To be announced at the LTRC 2016 banquet.

**ILTA Student Travel Awards**
Woranon Sitajalabhorn, University of Melbourne
Sharon Yahalom, University of Melbourne

**2014 ILTA Best Article Award**
Scott Crossley, Amanda Clevinger, and Youlin Kim, Georgia State University.

**2016 Jacqueline Ross TOEFL Dissertation Award**
Kimberly Woo, New York University, *Examining the construct of oral social language in the English language proficiency assessment of young ELLs.* (Supervisor: Lorena Llosa)
2015 Jacqueline Ross TOEFL Dissertation Award
Hans Rutger Bosker, Utrecht University, NLD, The processing and evaluation of fluency in native and non-native speech. (Supervisors: Nivja H. de Jong and Hugo Quené)

2016 TOEFL Outstanding Young Scholar Award
Tineke Brunfaut, Lancaster University

2015 Caroline Clapham IELTS Masters Award
Hyunjin Kim, University of Bristol, UK, Teachers’ voices in the decision to discontinue a public examination reform. (Supervisor: Talia Isaacs)

Highly commended:
Saeede Haghi, University of Warwick, UK (Supervisor: Claudia Harsch)

Sage/ILTA Book Award


2015 TIRF Doctoral Dissertation Grant Awardees in Language Assessment
Bai Ying, University of Melbourne, Australia
Jookyoung Jung, University College London, UK
Xu Yueting, University of Hong Kong, China
Yi Mei, Queen’s University, Canada
Piazza S. Antonino, 1
(Workshops)

Palazzo Steri
Palazzo Chiaramonte
(Opening Session)

University of Palermo
(Colloquium)
Villa Tasca  
(Wine Tasting)

Orto Botanico  
Via Lincoln, 2  
(Conference dinner)

Language Constructs, Contexts, and Content in Classroom and Large-Scale Assessments
<table>
<thead>
<tr>
<th>Hotels</th>
<th>Distance/Time (Hotel → Shuttle)</th>
<th>Distance/Time (Hotel → Colloquia)</th>
<th>Distance/Time (Hotel → Opening Session)</th>
<th>Distance/Time (Hotel → Workshops)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hotel Porta Felice</td>
<td>400 meters 5-minute walk</td>
<td>3 kilometers 38-minute walk</td>
<td>280 meters 3-minute walk</td>
<td>1.4 kilometers 18-minute walk</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Piazza Marina shuttle is recommended</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grand Hotel Piazza Borsa</td>
<td>400 meters 4-minute walk</td>
<td>2.5 km 32-minute walk</td>
<td>450 meters 5-minute walk</td>
<td>950 meters 12-minute walk</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Piazza Marina shuttle is recommended</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hotel Quinto Canto</td>
<td>700 meters 8-minute walk</td>
<td>2.2 km 28-minute walk</td>
<td>800 meters 10-minute walk</td>
<td>750 meters 9-minute walk</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Piazza Marina shuttle is recommended</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NH Hotel Palermo</td>
<td>900 meters 11-minute walk</td>
<td>3 kilometers 38-minute walk</td>
<td>850 meters 11-minute walk</td>
<td>1.2 kilometers 16-minute walk</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Piazza Marina shuttle is recommended</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Monday, June 20: Pre-Conference Workshops
(Piazza S. Antonino, 1)

<table>
<thead>
<tr>
<th>Time</th>
<th>Workshop</th>
<th>Room</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 - 16:30</td>
<td>Workshop 1, Day 1: Interaction in L2 Assessment: A Conversation Analytic Perspective</td>
<td>Room 109</td>
</tr>
<tr>
<td></td>
<td>Workshop 3, Day 1: Multi-Level Modeling in SLA &amp; Language Assessment</td>
<td>Library</td>
</tr>
<tr>
<td></td>
<td>Workshop 4, Day 1: Scenario-Based Language Assessment in Educational Settings: Theoretical Foundations, Prototype Examples &amp; Future Applications</td>
<td>Room 101</td>
</tr>
<tr>
<td></td>
<td>Workshop 5, Day 1: Understanding &amp; Using Standard Setting in LT: A Practical Foundation</td>
<td>Room 108</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:30</td>
<td>Coffee break</td>
</tr>
<tr>
<td>12:30</td>
<td>Light lunch</td>
</tr>
<tr>
<td>16:30</td>
<td>Coffee break</td>
</tr>
</tbody>
</table>

### Tuesday, June 21: Pre-Conference Workshops
(Piazza S. Antonino, 1)

<table>
<thead>
<tr>
<th>Time</th>
<th>Workshop</th>
<th>Room</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 - 16:00</td>
<td>Workshop 1, Day 2: Interaction in L2 Assessment: A Conversation Analytic Perspective</td>
<td>Room 109</td>
</tr>
<tr>
<td></td>
<td>Workshop 2, Day 1: Methods &amp; Techniques for Large-Scale Italian Test Construction/Metodi e Tecniche per la Costruzione di Test</td>
<td>Room 104</td>
</tr>
<tr>
<td></td>
<td>Workshop 3, Day 2: Multi-Level Modeling in SLA &amp; Language Assessment</td>
<td>Library</td>
</tr>
<tr>
<td></td>
<td>Workshop 4, Day 2: Scenario-Based Language Assessment in Educational Settings: Theoretical Foundations, Prototype Examples &amp; Future Applications</td>
<td>Room 101</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:30</td>
<td>Coffee break</td>
</tr>
<tr>
<td>12:30</td>
<td>Light lunch</td>
</tr>
<tr>
<td>16:00</td>
<td>Coffee break</td>
</tr>
</tbody>
</table>
### Tuesday, June 21: ILTA EAB Meeting
(Piazza S. Antonino, 1)

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>12:00 – 17:00</td>
<td>ILTA Pre-Conference EAB Meeting</td>
<td>Room 105</td>
</tr>
<tr>
<td>15:30 – 17:00</td>
<td>Newcomers Session</td>
<td>Room 101</td>
</tr>
</tbody>
</table>

### Tuesday, June 21: LTRC 2016 Opening Sessions
(Palazzo Chiaramonte-Steri)

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>15:00 - 17:30</td>
<td>Registration</td>
<td>Balcony</td>
</tr>
<tr>
<td>17:30 – 18:00</td>
<td>Welcome to LTRC 2016</td>
<td>Sala Magna</td>
</tr>
<tr>
<td>18:00 – 18:30</td>
<td>Plenary</td>
<td></td>
</tr>
<tr>
<td>18:30 – 19:00</td>
<td>Plenary</td>
<td></td>
</tr>
<tr>
<td>19:00 – 21:00</td>
<td>Opening Reception</td>
<td>Neoclassico</td>
</tr>
</tbody>
</table>

### Wednesday, June 22: LTRC 2016
(University of Palermo, Scuola Politecnica - Faculty of Engineering)

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 – 9:00</td>
<td>Registration</td>
<td>Aula Magna Lobby</td>
</tr>
<tr>
<td>9:00 – 9:55</td>
<td>The Davies Lecture</td>
<td>Aula Magna</td>
</tr>
<tr>
<td>10:00 – 10:30</td>
<td>Parallel sessions</td>
<td>Aula Magna Rooms 6, 7, 8</td>
</tr>
<tr>
<td>10:30 – 11:00</td>
<td>Coffee break</td>
<td></td>
</tr>
<tr>
<td>11:00 – 11:30</td>
<td>Parallel sessions</td>
<td>Aula Magna Rooms 6, 7, 8</td>
</tr>
<tr>
<td>11:35 – 13:00</td>
<td>Symposia</td>
<td>Rooms 7, 8</td>
</tr>
<tr>
<td>13:00 – 14:30</td>
<td>Lunch break</td>
<td></td>
</tr>
<tr>
<td>13:00 – 14:30</td>
<td>LAQ EAB Meeting</td>
<td>Room 6</td>
</tr>
<tr>
<td>14:30 – 15:00</td>
<td>Parallel sessions</td>
<td>Aula Magna Rooms 6, 7, 8</td>
</tr>
<tr>
<td>15:05 – 15:35</td>
<td>Parallel sessions</td>
<td>Aula Magna Rooms 6, 7, 8</td>
</tr>
<tr>
<td>15:35 – 16:00</td>
<td>Coffee break</td>
<td></td>
</tr>
</tbody>
</table>
## Thursday, June 23: LTRC 2016

(University of Palermo, Scuola Politecnica - Scuola Politecnica - Faculty of Engineering)

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 – 9:00</td>
<td>Registration</td>
<td>Aula Magna Lobby</td>
</tr>
<tr>
<td>10:00 – 16:00</td>
<td>Book Exhibit &amp; Sponsors Tables</td>
<td>Classroom Building Hall 1st floor</td>
</tr>
<tr>
<td>9:00 – 9:55</td>
<td>The Samuel J. Messick Memorial Lecture</td>
<td>Aula Magna</td>
</tr>
<tr>
<td>10:00 – 10:30</td>
<td>Plenary</td>
<td>Aula Magna</td>
</tr>
<tr>
<td>10:30 – 11:00</td>
<td>Plenary</td>
<td>Aula Magna</td>
</tr>
<tr>
<td>11:00 – 11:30</td>
<td>Coffee break</td>
<td></td>
</tr>
<tr>
<td>11:30 – 12:00</td>
<td>Parallel sessions</td>
<td>Rooms 5, 6, 7, 8</td>
</tr>
<tr>
<td>12:05 – 12:35</td>
<td>Parallel sessions</td>
<td>Rooms 5, 6, 7, 8</td>
</tr>
<tr>
<td>12:35 – 14:00</td>
<td>Lunch break</td>
<td></td>
</tr>
<tr>
<td>12:35 – 14:00</td>
<td>LT EAB Meeting</td>
<td>Room 6</td>
</tr>
<tr>
<td>14:00 – 14:30</td>
<td>Plenary</td>
<td>Aula Magna</td>
</tr>
<tr>
<td>14:30 – 16:00</td>
<td>Symposium</td>
<td>Room 7</td>
</tr>
<tr>
<td></td>
<td>Demos</td>
<td>Room 8</td>
</tr>
<tr>
<td>16:00 – 16:30</td>
<td>Coffee break</td>
<td></td>
</tr>
<tr>
<td>16:30 – 17:00</td>
<td>Parallel sessions</td>
<td>Room 5, 6, 7, 8</td>
</tr>
<tr>
<td>17:05 – 17:50</td>
<td>WIPs</td>
<td>Classroom Building Hall upper level</td>
</tr>
<tr>
<td>17:55</td>
<td>Group Photo</td>
<td>On stairs in front of Aula Magna</td>
</tr>
</tbody>
</table>
Friday, June 24: LTRC 2016
(University of Palermo, Scuola Politecnica - Scuola Politecnica - Faculty of Engineering)

8:00 – 9:00  Registration  Aula Magna Lobby
10:00 – 16:00  Book Exhibit & Sponsors Tables  Classroom Building Hall 1st floor
9:00 – 10:30  Symposium  Aula Magna
10:30 – 11:00  Coffee break  Aula Magna
11:00 – 11:30  Parallel sessions  Room 5, 6, 7, 8
11:35 – 12:05  Parallel sessions  Rooms 5, 6, 7, 8
12:10 – 12:40  Parallel sessions  Rooms 5, 6, 7, 8
12:40 – 13:30  Lunch break
13:30 – 15:30  ILTA Business Meeting  Aula Magna
15:30 – 16:00  Plenary  Aula Magna
16:05 – 16:35  Parallel sessions  Room 5, 6, 7, 8
16:35 – 17:00  Coffee break
17:00 – 17:30  Plenary  Aula Magna
17:35 – 18:35  The Cambridge/ILTA Distinguished Achievement Award  Aula Magna
18:35 – 18:45  Closing session  Aula Magna
20:30  Banquet  Orto Botanico

Saturday, June 25
10:00 – 13:30  Palermo Walking Tour
09:00 – 21:00  Excursion to Segesta – Trapani – Mondello

Sunday, June 26
10:00 – 12:30  Palermo Street Art Tour

Language Constructs, Contexts, and Content in Classroom and Large-Scale Assessments
### Monday, June 20: Workshops (S. Antonino)
Centro Linguistico e Scuola d’Italiano per Stranieri – Università di Palermo - Piazza S. Antonino, 1, Palermo

<table>
<thead>
<tr>
<th>Time</th>
<th>Workshop 1</th>
<th>Workshop 2</th>
<th>Workshop 3</th>
<th>Workshop 4</th>
<th>Workshop 5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Carsten Roever and Gabriele Kasper</td>
<td>Steven J. Ross</td>
<td></td>
<td>John Sabatini, Tenaha O’Reilly, and James E. Purpura</td>
<td>Dorry M. Kenyon and David MacGregor</td>
</tr>
<tr>
<td></td>
<td>Room 109</td>
<td>Library</td>
<td></td>
<td>Room 101</td>
<td>Room 108</td>
</tr>
</tbody>
</table>

### Tuesday, June 21: Workshops (S. Antonino)
Centro Linguistico e Scuola d’Italiano per Stranieri – Università di Palermo - Piazza S. Antonino, 1, Palermo

<table>
<thead>
<tr>
<th>Time</th>
<th>Workshop 1</th>
<th>Workshop 2</th>
<th>Workshop 3</th>
<th>Workshop 4</th>
<th>Workshop 5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Carsten Roever and Gabriele Kasper</td>
<td>Roberto Ricci</td>
<td>Steven J. Ross</td>
<td>John Sabatini, Tenaha O’Reilly, and James E. Purpura</td>
<td>Dorry M. Kenyon and David MacGregor</td>
</tr>
<tr>
<td></td>
<td>Room 109</td>
<td>Room 104</td>
<td>Library</td>
<td>Room 101</td>
<td>Room 108</td>
</tr>
<tr>
<td>Time</td>
<td>Event</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>----------------------------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12:00 – 17:00</td>
<td>ILTA Pre-Conference Executive Advisory Board Meeting</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Room 105</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15:30 – 17:00</td>
<td>Newcomers Session (Sponsored by Language Learning and Testing Foundation)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Charlie Stansfield and Margaret E. Malone</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Room 101</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Tuesday June 21: LTRC Opening Sessions**
Palazzo Chiaramonte-Steri (Sala Magna) - Università di Palermo - Piazza Marina, 61, Palermo

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>15:00 – 17:30</td>
<td>Registration</td>
</tr>
<tr>
<td></td>
<td>Palazzo Chiaramonte-Steri, Balcony</td>
</tr>
<tr>
<td>17:30 – 18:00</td>
<td>Welcome to LTRC 2016</td>
</tr>
<tr>
<td></td>
<td>James E. Purpura and Monica Barni, Welcome from the Co-Chairs</td>
</tr>
<tr>
<td></td>
<td>Mari D’Agostino, Welcome and Introduction of Rector of the University of Palermo</td>
</tr>
<tr>
<td></td>
<td>Rector Frabrizio Micari, Welcome from the Rector of the University of Palermo</td>
</tr>
<tr>
<td></td>
<td>Patrizia Ardizzone, Welcome from President of the University Language Center</td>
</tr>
<tr>
<td></td>
<td>Carla Bagna, Introduction of the evening sessions</td>
</tr>
<tr>
<td></td>
<td>Palazzo Chiaramonte-Steri, Sala Magna</td>
</tr>
<tr>
<td>18:00 – 18:30</td>
<td>Plenary</td>
</tr>
<tr>
<td></td>
<td>The Dynamics of Test Impact in the Context of Australia’s Skilled Migration Policy: RE-Interpreting Language Constructs and Scores</td>
</tr>
<tr>
<td></td>
<td>Kellie Frost</td>
</tr>
<tr>
<td></td>
<td>Palazzo Chiaramonte-Steri, Sala Magna</td>
</tr>
<tr>
<td>18:30 – 19:00</td>
<td>Plenary</td>
</tr>
<tr>
<td></td>
<td>Young Learners’ Processes and Rationales for Responding to Different Types of Self-assessments</td>
</tr>
<tr>
<td></td>
<td>Yuko Butler</td>
</tr>
<tr>
<td></td>
<td>Palazzo Chiaramonte-Steri, Sala Magna</td>
</tr>
<tr>
<td>19:00 – 21:00</td>
<td>Opening Reception (Sponsored by Cambridge English)</td>
</tr>
<tr>
<td></td>
<td>Palazzo Chiaramonte-Steri, Neoclassico</td>
</tr>
</tbody>
</table>

**Wednesday, June 22: LTRC**
(University of Palermo, Faculty of Engineering Engineering classrooms & Aula Magna)
Scuola Politecnica e Aula Magna Ingegneria - Via delle Scienze 7 - Palermo

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 – 9:00</td>
<td>Registration</td>
</tr>
<tr>
<td></td>
<td>Aula Magna Lobby</td>
</tr>
<tr>
<td>10:00 – 16:00</td>
<td>Book Exhibit &amp; Sponsors Tables</td>
</tr>
<tr>
<td></td>
<td>Classroom Building Hall (1st floor)</td>
</tr>
</tbody>
</table>
### Times Event

<table>
<thead>
<tr>
<th>Times</th>
<th>Event</th>
</tr>
</thead>
</table>
| 9:00 – 9:55 | **The Alan Davies Lecture** (Sponsored by the British Council)  
Reframing the SLA-Assessment Interface: ‘Constructive’ Deliberations at the Nexus of Interpretations, Contexts, and Consequences  
*John Norris*  
Aula Magna |
| 10:00 – 10:30 | **4 Parallel Sessions**  
Exchanging the Validity of a Performance-Based Oral Test for University Lecture Certification  
*Slobodanka Dimova*  
Aula Magna  
Room 6 |
| 10:00 – 10:30 | **4 Parallel Sessions**  
Examining the Validity of a Performance-Based Oral Test for University Lecture Certification  
*Dimova*  
Room 6 |
| 10:00 – 10:30 | **4 Parallel Sessions**  
Examining L2 Learning and Learning Outcomes as a Result of Embedded Classroom Assessments  
*James Purpura*,  
*Jorge Beltrán*,  
*Heidi Han-Ting Liu*, and  
*Michelle Stabler-Havener*  
Room 7 |
| 10:30 – 11:00 | Coffee Break |
| 11:00 – 11:30 | **4 Parallel Sessions**  
Assessing Low Educated Learners – Implications for the Test Construct for Large-Scale Assessment  
*Cecilie Hamnes CarlSEN*  
Aula Magna  
Room 6 |
| 11:00 – 11:30 | **4 Parallel Sessions**  
Maximizing Feedback for Learning: Investigating Language Learners’ Differing Cognitive Processing when Receiving Computer-Based Feedback  
*Maggie Dunlop*  
Room 7 |
| 11:00 – 11:30 | **4 Parallel Sessions**  
 Validation of an LSP Test Aligned with the CEFR  
*Joan Jamieson*,  
*Daniel Isbell*, and  
*L. D. Nicolas May*  
Room 8 |
| 11:00 – 11:30 | **4 Parallel Sessions**  
Validation of a Rating Scale for the Assessment of Summary Writing within an Argument-based Framework  
*Woranon Sitajalabhorn*  
Room 7 |
| 11:00 – 11:30 | **4 Parallel Sessions**  
Investigating the Predictive Validity of a High-Stakes French Admission Test  
*Romain Schmitt* and  
*Shahrzad Saif*  
Room 8 |
2 Symposia

Organizers: Beverly Baker and Gillian Wigglesworth
Assessment of Indigenous Languages: Examining the Cases of Australia and Canada

Paper 1: Challenges in the assessment of indigenous languages (Gillian Wigglesworth)
Paper 2: The impact of standardised testing on children in indigenous communities (Ute Knoch)
Paper 3: A narrative inquiry into the formative assessment practices of an indigenous language teacher (Joyce Germain and Beverly Baker)
Paper 4: NETOLNEW ‘One mind, one people’: Developing a context-relevant Assessment Tool for adult Indigenous language learners in Canada (Onowa McIvor, Peter Jacobs, and Barbara Jenni)

Discussant: Elana Shohamy

Organizers: Dorry M. Kenyon and David MacGregor
Using a Comprehensive Framework to Integrate Constructs, Contexts, and Content in a Large-Scale Language Assessment

Paper 1: The Need for a Comprehensive Assessment Framework (Dorry M. Kenyon)
Paper 2: Using the Framework to Address Issues Related to Test Consequences, Decisions, and Interpretations (Ahyoung Alicia Kim and Mark Chapman)
Paper 3: Using the Framework to Address Issues Related to Constructs and Content (Jennifer Norton)
Paper 4: Using the Framework to Address Measurement Issues in Assessment Delivery (David MacGregor and Xin Yu)

Discussant: Lyle Bachman

13:00 – 14:30 Lunch Break

Language Assessment Quarterly EAB Meeting
Room 6

14:30 – 15:00 4 Parallel Sessions

After the Entrance Test. A Longitudinal Study of L2 Students’ Experiences at University
Bart Deygers

“I can see that”: Developing Shared Rubric Category Interpretations through Score Negotiation
Valerie Meier, Jonathan Trace, and Gerriet Janssen

Applying a Concept Model of Dynamic Assessment to a Large-Scale Speaking Test
Tzioni Levi

Demo
Using Web-Technology and Spoken Dialogue Systems to Assess L2 Pragmatic Ability
Veronika Timpe Laughlin, Heidi Han-Ting Liu, Jennifer Wain, and Keelan Evanini

Aula Magna Room 6 Room 7 Room 8
<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>15:05 – 15:35</td>
<td><strong>4 Parallel Sessions</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Is the Large-scale Testing of English for Academic Purposes a Misnomer?</strong></td>
<td>Jo Lewkowicz, Constant Leung, and Jennifer Jenkins</td>
</tr>
<tr>
<td></td>
<td><strong>Comparability of Students’ Writing Performance on TOEFL-iBT and Required University Writing Courses</strong></td>
<td>Lorena Llosa, Margaret E. Malone, Anne Donovan, Christopher Van Booven, and Scott Grapin</td>
</tr>
<tr>
<td></td>
<td><strong>Complex Systems Interacting in a Classroom-Based Dynamic Assessment</strong></td>
<td>Clarissa Lau and Eunice Eunhee Jang</td>
</tr>
<tr>
<td></td>
<td><strong>Demo</strong></td>
<td>Test-ify: Software for Assessment Design and Instruction</td>
</tr>
<tr>
<td></td>
<td><strong>Aula Magna</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Room 6</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Room 7</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Room 8</strong></td>
<td></td>
</tr>
<tr>
<td>15:35 – 16:00</td>
<td><strong>Coffee break</strong></td>
<td></td>
</tr>
<tr>
<td>16:00 – 16:30</td>
<td><strong>Plenary</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Diagnosing Development of Reading and Writing in a Foreign Language - A Longitudinal Perspective</strong></td>
<td>Ari Huhta, Charles Alderson, Lea Nieminen, Riikka Ullakanoja, and Eeva-Leena Haapakangas</td>
</tr>
<tr>
<td></td>
<td><strong>Aula Magna</strong></td>
<td></td>
</tr>
<tr>
<td>16:30 – 17:45</td>
<td><strong>Posters</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Chairs: Lucilla Lopriore and Barbara Dobson</strong></td>
<td>Classroom Building Hall (upper level)</td>
</tr>
<tr>
<td></td>
<td><strong>1</strong></td>
<td>Language Assessment Expert, Yes, But Are You a Police Interrogation Expert?</td>
</tr>
<tr>
<td></td>
<td><strong>2</strong></td>
<td>Oral Composition and Mini-debate: Do They Measure the Same Construct?</td>
</tr>
<tr>
<td></td>
<td><strong>3</strong></td>
<td>Assessment of the Speaking Proficiency and the Civic Competence within the Migration Context: A Case Study on a Large Scale Certification</td>
</tr>
<tr>
<td></td>
<td><strong>4</strong></td>
<td>The Interface of Construct, Contexts, and Content in the Revision of a Reading Test</td>
</tr>
<tr>
<td></td>
<td><strong>5</strong></td>
<td>Using Proficiency Test Data to Set Program Goals: Interpretation, Speculation, and Collaboration</td>
</tr>
<tr>
<td></td>
<td><strong>6</strong></td>
<td>Monitoring Assessment of Writing in Large Scale Examinations Context: A Tool to Improve Raters’ Self-Awareness</td>
</tr>
<tr>
<td></td>
<td><strong>7</strong></td>
<td>Improving Quality in Classroom-Based Language Assessment: The Case of a Toolkit Alternative Assessment</td>
</tr>
</tbody>
</table>

**Language Constructs, Contexts, and Content in Classroom and Large-Scale Assessments**
<table>
<thead>
<tr>
<th>Session</th>
<th>Title</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>8)</td>
<td>Testing for Differential Item Functioning with No Internal Matching Variable and Continuous Item Ratings</td>
<td>Michelle Y. Chen, Wendy Lam, and Bruno D. Zumbo</td>
</tr>
<tr>
<td>9)</td>
<td>Effects of Changing TOEFL Cut-off Scores: The Impact of Raising the Bar</td>
<td>Laura Decker, Troy Cox, and Jesse Egbert</td>
</tr>
<tr>
<td>10)</td>
<td>Validating a Scale Revision Project: Results and Implications for the Underlying Model of Validation</td>
<td>Jamie Dunlea, Judith Fairbairn, and Barry O’Sullivan</td>
</tr>
<tr>
<td>11)</td>
<td>The Effects of Test Re-Dos on Learner Performance and Development</td>
<td>Kristen Sullivan</td>
</tr>
<tr>
<td>12)</td>
<td>Comparability of the GEPT – Advanced and the iBT: Content and Constructs</td>
<td>Antony Kunnan and Nathan Carr</td>
</tr>
<tr>
<td>13)</td>
<td>Developing an Electronic Pre-Test for the EPPLE Examination: Test Design, Technical Challenges and Mobile Technology</td>
<td>Douglas Altamiro Consolo and Debora Mieko</td>
</tr>
<tr>
<td>14)</td>
<td>Measuring Progress in English for Tourism in a Hungarian University Context</td>
<td>Zoltán Lukács</td>
</tr>
<tr>
<td>15)</td>
<td>Creating a Computer Delivered, Multimedia Listening Test: Challenges, Choices and the AUA</td>
<td>May Tan, Nancy Powers, Roderick Broeker, and Ruth Lucas</td>
</tr>
<tr>
<td>16)</td>
<td>OLA (Offerta Linguistica di Ateneo) at UniCal: Designing and Implementing a New Test Format</td>
<td>Carmen Argondizzo, Jean Jimenez, and Ian Robinson</td>
</tr>
<tr>
<td>17)</td>
<td>Teacher Assessment Literacy Enhancement: Identifying Contextual Factors</td>
<td>Dina Tsagari, Tony Green, Karin Vagt, Ildiko Csepes, and Nicos Sifakis</td>
</tr>
<tr>
<td>18)</td>
<td>Bridging the Gap between Research and Practice: Reflections from Constructing a Cognitive Diagnostic Reading Assessment</td>
<td>Junli Wei</td>
</tr>
<tr>
<td>19)</td>
<td>Adding an Additional Oral English Language Assessment after University Admission</td>
<td>Erik Vass and Annie Hsu</td>
</tr>
<tr>
<td>20)</td>
<td>Test Development at the Intersection of Context, Construct, and Content</td>
<td>Yasuko Okabe, Catherine Pulupa, Yuko Kashimada</td>
</tr>
<tr>
<td>21)</td>
<td>Study on Comparability of Language Testing in Europe</td>
<td>Nick Saville and Esther Gutierrez Eugenio</td>
</tr>
<tr>
<td>22)</td>
<td>Learning-Oriented Assessment in and out of Class: Text and Context at Work</td>
<td>Angela Fenara</td>
</tr>
<tr>
<td>23)</td>
<td>Developing a Speaking Rubric for the Vietnamese Standardized Test of English Proficiency (VSTEP)</td>
<td>Huu Nguyen, Hoa Nguyen, Tuan Huynh, Phuong Tran, Ha Do, Thao Nguyen, Pham, Duyen Can, Anh Nguyen, and Fred Davidson</td>
</tr>
<tr>
<td>Time</td>
<td>Event</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>19:00 – 21:00</td>
<td><strong>Wine Tasting (Subsidized by Pearson)</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Villa Tasca (Bus to Venue at 18:00; meeting place TBA)</td>
<td></td>
</tr>
</tbody>
</table>

**Thursday, June 23: LTRC**  
(University of Palermo, Faculty of Engineering Engineering classrooms & Aula Magna)  
Scuola Politecnica e Aula Magna Ingegneria - Via delle Scienze 7 - Palermo

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 – 9:00</td>
<td><strong>Registration</strong></td>
</tr>
<tr>
<td></td>
<td>Aula Magna Lobby</td>
</tr>
<tr>
<td>10:00 – 16:00</td>
<td><strong>Book Exhibit &amp; Sponsors Tables</strong></td>
</tr>
<tr>
<td></td>
<td>Classroom Building Hall (1st floor)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 – 9:55</td>
<td><strong>The Samuel J. Messick Memorial Lecture</strong> (Sponsored by ETS)</td>
</tr>
<tr>
<td></td>
<td>A Framework for Defining and Evaluating the Validity of Instructionally Relevant Assessments</td>
</tr>
<tr>
<td></td>
<td><em>Jim Pellegrino</em></td>
</tr>
<tr>
<td></td>
<td>Aula Magna</td>
</tr>
<tr>
<td>10:00 – 10:30</td>
<td><strong>Plenary</strong></td>
</tr>
<tr>
<td></td>
<td><em>Ikkyu Choi and Youngsoon So</em></td>
</tr>
<tr>
<td></td>
<td>Aula Magna</td>
</tr>
<tr>
<td>10:30 – 11:00</td>
<td><strong>Plenary</strong></td>
</tr>
<tr>
<td></td>
<td><em>Ute Knoch and Carol Chapelle</em></td>
</tr>
<tr>
<td></td>
<td>Aula Magna</td>
</tr>
<tr>
<td>11:00 – 11:30</td>
<td><strong>Coffee Break</strong></td>
</tr>
</tbody>
</table>
### 4 Parallel Sessions

<table>
<thead>
<tr>
<th>11:30 – 12:00</th>
<th>Diagnosing EFL Learner’s Writing Difficulties in Chinese Secondary Schools</th>
<th>The Validity of Reader/Text Matching Algorithms</th>
<th>Examining the Writing Construct under Different Contexts Using Automated Features</th>
<th>An Investigation of Young EFL Students’ Performance on a Timed Argumentative Writing Task</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>Cecilia Guanfang Zhao</em></td>
<td><em>Kathleen Sheehan</em></td>
<td><em>Chaitanya Ramineni and Brent Bridgeman</em></td>
<td><em>Mikyung Kim Wolf, Saerhim Oh, Yuan Wang, and Fred Tsutagawa</em></td>
</tr>
<tr>
<td>Room 5</td>
<td>Room 6</td>
<td>Room 7</td>
<td>Room 8</td>
<td></td>
</tr>
</tbody>
</table>

### 12:05 – 12:35

<table>
<thead>
<tr>
<th>4 Parallel Sessions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Construct Validity of a High-Stakes Writing Test: The Effects of Two Types of Expository Essay Writing</td>
</tr>
<tr>
<td>Testing Bilingualism: Incorporating Translanguaging into a Listening Task for University Professors</td>
</tr>
<tr>
<td>Language Assessment Raters in the Call Centre Industry: Interfaces between Language and Subject Matter Experts</td>
</tr>
<tr>
<td>Young EFL Learners’ Strategy Use during L2 Speaking Task Performance</td>
</tr>
<tr>
<td><em>Hyun Jung Kim and Jooseung Lee</em></td>
</tr>
<tr>
<td><em>Beverly Baker and Amelia Hope</em></td>
</tr>
<tr>
<td><em>Jane Lockwood and Michelle Raquel</em></td>
</tr>
<tr>
<td><em>Ching-Ni Hsieh and Lin Gu</em></td>
</tr>
<tr>
<td>Room 5</td>
</tr>
<tr>
<td>Room 6</td>
</tr>
<tr>
<td>Room 7</td>
</tr>
<tr>
<td>Room 8</td>
</tr>
</tbody>
</table>

### 12:35 – 14:00

- **Language Testing EAB Meeting**
  - Room 6

### 14:00 – 14:30

- **Plenary**
  - Face-to-Face and Video-Conferencing Technology Delivered Speaking Tests: Comparing Constructs
  - *Vivien Berry, Fumiyo Nakatsuhara, and Chihiro Inoue*  
  - Aula Magna
## 1 Symposium

**Organizers:** Liying Cheng and Antony Kunnan

How does Context Mediate Construct and Content? A Political, Social and Pedagogical Analysis of English Language Testing in Asia

Paper 1. The contextual mediation of educational and social features influencing test construction (Yan Jin)

Paper 2. The Demise of NEAT: Whose Defeat or Triumph Is It? (Yong-Won Lee)

Paper 3. Understanding “the ability-in-language user-in-context” through the eyes of stakeholders (Jirada Wudthayagorn)

Paper 4. English language demands at universities—what is the construct of “academic English” in the Japanese context? (Yasuyo Sawaki)

**Room 7**

## 4 Demos (Speaking Assessment)

**Chair:** Evelina D. Galaczi

Using Technology to Help Learners and Teachers in a Learning-Oriented Approach to Assessment

Evelina D. Galaczi and Sarah Unsworth

The Use of Live Voice Communication Virtual Environments to Assess L2 Oral Communication

Gary Ockey, Lin Gu, and Madeline Keehner

Say Hello to TELL

Masanori Suzuki, William Bonk, and Alistair Van Moere

Educator Rater Training Materials for the ACCESS for ELLs 2.0 Speaking Test

Mark Chapman and Meg Montee

**Room 8**

## 16:00 – 16:30 Coffee Break

## 16:30 – 17:00 4 Parallel Sessions

<table>
<thead>
<tr>
<th>Paper 1</th>
<th>Paper 2</th>
<th>Paper 3</th>
<th>Paper 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Score Augmentation Approach to the Investigation of the Added Value of Subscores in Reading and Listening Tests</td>
<td>The Effects of Test Task and Pairing Types on Heritage and Non-Heritage Korean Learners’ Speaking Performance</td>
<td>An Eye-Tracking Study of Task Types and Attention to Nonverbal Cues in Video Listening Tests</td>
<td>Towards Developing English Ability Indicators for College Students in Taiwan—Constructing tests of English for specific academic purposes</td>
</tr>
<tr>
<td>Spiros Papageorgiou and Ikkyu Choi</td>
<td>Sun-Young Shin</td>
<td>Aaron Olaf Batty</td>
<td>Jessica Wu and Rachel Y. F. Wu</td>
</tr>
<tr>
<td>Room 5</td>
<td>Room 6</td>
<td>Room 7</td>
<td>Room 8</td>
</tr>
</tbody>
</table>

## 17:05 – 17:50 Works-in-Progress (WIPs)

**Chairs:** Lorenzo Rocca and Barbara Dobson

Classroom Building Hall (upper level)

1) Investigating the Prompt Effect on Test Scores and Score Interpretations of a Group Oral Test

Zhouyang Lu
| 2) | Behavior in Speaking Tests: A Preliminary Model of Interaction | Jayanti Banerjee and India Plough |
| 3) | Justifying the Use of Computer Automated Scoring in Portfolio-Based Writing Assessment in EFL Classroom | Shangchao Min, Lianzhen He, and Dajian Chen |
| 4) | Incentivizing Students to Reach Stated Proficiency Goals | Bill VanPatten, Susan Gass, Paula Winke, and Koen Van Gorp |
| 5) | Mapping Language Use to the CLB and CELPIP-General LS within Workplace Contexts for New Immigrants | Christine Doe, Scott Douglas, and Liying Cheng |
| 6) | Investigating the Validity of the Multiple-Choice Method for Assessing Comprehension of Conversational Implicature | Stephen O’Connell |
| 7) | Investigating the Usefulness of the TOEFL® Primary™ Reading and Listening Tests to Measure Young EFL Students’ English Learning Progress | Ian Blood and Yeonsuk Cho |
| 9) | Investigating Language Assessment Literacy in Brazil | Gladys Quevedo-Camargo and Matilde Scaramucci |
| 10) | The Construct of EFL Academic Writing Ability as Operationalized in the Context of Chinese Colleges | Cecilia Guanfang Zhao and Jinsong Fan |
| 11) | Investigating Teachers’ Attitudes towards Portfolio Based Language Assessment (PBLA) in Language Instruction for Newcomers to Canada (LINC) Program | Fatemeh Mohammadian Haghighi |
| 12) | Investigating Cognitive Validity in a Test of English for Italian Students with SpLDs | Geraldine Ludbrook and Claudia D’Este |
| 13) | A Fresh Look at the Word Associates Format to Measure Depth of Vocabulary knowledge | John Read |
| 14) | Exploring the Relationship between TOEFL Scores and Academic English Performance on Reading, Speaking and Writing Tasks at a Colombian University | Isabel Tejada-Sanchez |
| 15) | Diagnostic Writing Assessment of First-Year Engineering Students: Providing Profile-Based Pedagogical Support to Develop Academic Language Proficiency | Janna Fox and Maryam Wagner |
| 16) | Authorial Voice in Integrated Writing Assessment Tasks | Atta Gebril |
17) ACTFL Proficiency Levels and TOEFL Scores of Students Entering US Universities  
Daniel Reed, Susan Gass, and Erwin Tschirner

18) Investigating the Validity of TOEFL iBT Integrated Writing Assessment: Do the Tasks and the Rating Scheme Used Require Any Revision?  
Maryam Homayounzadeh and Alireza Ahmadi

19) ECD for MSA - Developing a Comprehensive Construct Definition  
Bjorn Norrbom, Yong Lou, and Abdulrahman Alshamrani

17:55  Group Photo (In front of Aula Magna)

---

Friday, June 24: LTRC  
(University of Palermo, Faculty of Engineering classrooms & Aula Magna)  
Scuola Politecnica e Aula Magna Ingegneria - Via delle Scienze 7 - Palermo

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
</table>
| 8:00 – 9:00| Registration
Aula Magna Lobby |
| 10:00 – 16:00| Book Exhibit & Sponsors Tables
Classroom Building Hall (1st floor) |

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
</table>
| 9:00 – 10:30| Symposium
Organizers: Xiaoming Xi and Larry Davis
Evolving Academic English Constructs and New Innovative Tasks |

  - Paper 1: Integrating multiple sources and digital affordances into reading assessment: opportunities and challenges (Tenaha O’Reilly and Mary Schedl)
  - Paper 2: Incorporating audience into source-based writing tasks – design challenges and considerations (Yeonsuk Cho, Ikkyu Choi, Robert Kantor, and Jakub Novák)
  - Paper 3: User perceptions of task design features intended to improve the test-taking experience and validation argument for innovative speaking tasks (Jonathan Schmidgall, Yuan Wang, Larry Davis, Pam Mollaun, and Ching-Ni Hsieh)
  - Paper 4: Development and evaluation of innovative speaking tasks: Increased context to evaluate communicative appropriateness (Larry Davis, Pam Mollaun, Jonathan Schmidgall, Ching-Ni Hsieh, and Yuan Wang)

Aula Magna

10:30 – 11:00  Coffee break
### 11:00 – 11:30  3 Parallel Sessions

<table>
<thead>
<tr>
<th>Nurses’ Perspectives of The Qualities of Referral Letters: Towards Profession-Oriented Assessment Criteria</th>
<th>Advancing Written Assessment Feedback Praxis through a Teacher-Researcher Collaboration in a University-level Spanish program</th>
<th>Demo DELTA Writing Component: An Online Automated Diagnostic Test of Academic Writing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sharon Yahalom</td>
<td>Kathryn Hill and Ana Marie Ducasse</td>
<td>Alan Urmston, Michelle Raquel, and Roxanne Wong</td>
</tr>
</tbody>
</table>

Room 6  Room 7  Room 8

### 11:35 – 12:05  4 Parallel Sessions

<table>
<thead>
<tr>
<th>Exploring Contextualized Tasks Through L2 Students’ Oral Performance: How Discriminated, Constrained and Generalizable are the Observed Task Performance Features?</th>
<th>The Challenge of Capturing the Construct of Written Communication for a Health-Specific Language Test</th>
<th>A Meaning-Based, CEFR-Linked Framework for Assessing Vocabulary Knowledge</th>
<th>Demo An Innovative Online Testing and Rating System for the Integrated Reading-Writing Assessment Research</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naoki Ikeda</td>
<td>Ute Knoch Tim McNamara Catherine Elder Robyn Woodward-Kron Elizabeth Manias Eleanor Flynn Annemiek Huisman</td>
<td>Veronica Benigno and John H. A. L. de Jong</td>
<td>Yu-Chen Lin and Sheng-Bo Chen</td>
</tr>
</tbody>
</table>

Room 5  Room 6  Room 7  Room 8

### 12:10 – 12:40  4 Parallel Sessions

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Soo Jung Youn</td>
<td>John Pill and Sally O’Hagan</td>
<td>Benjamin Kremmel and Norbert Schmitt</td>
<td>David Booth and Paul Seddon</td>
</tr>
</tbody>
</table>

Room 5  Room 6  Room 7  Room 8
12:40 – 13:30  **Lunch Break**

13:30 – 15:30  **ILTA Business Meeting**  
Aula Magna

15:30 – 16:00  **Plenary**  
Is conversation analysis really the answer in pragmatics assessment research?  
*Kirby Grabowski*  
Aula Magna

16:05 – 16:35  **4 Parallel Sessions**

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
</table>
| 16:05 – 16:35 | The Written Discourse Completion Test: Different Versions, Different Constructs  
Afe Labben  
Room 5 |
| 16:05 – 16:35 | Assessing Interactional Competence on Managing Topic Transitions in EFL Group Oral Task  
Patharann Patharakorn  
Room 6 |
| 16:05 – 16:35 | Exploring the Relationship Between Lexical Diversity Measures and L2 Proficiency  
Veronica Benigno, William Bonk, and Masanori Suzuki  
Room 7 |
| 16:05 – 16:35 | Practice Makes Perfect: The Likelihood of Success in Retesting  
Rachel Brooks and Maria Brau  
Room 8 |

16:35 – 17:00  Coffee break

17:00 – 17:30  **Plenary**  
Construct Validation of Multimodal Scenario-Based Language Assessment (SBLA) Tasks for Diagnostic Placement Purposes  
*Eunice Eunhee Jang, Maryam Wagner, and Maggie Dunlop*  
Aula Magna

17:35 – 18:35  **The Cambridge/ILTA Distinguished Achievement Award**  
(Sponsored by Cambridge English)  
Modelling language competence into a global framework: Taking the past into the future  
*John H. A. L. de Jong*  
Aula Magna

18:35 – 18:45  **Closing Session**  
Aula Magna

20:30  **Banquet - Orto Botanico**

---

**Saturday, June 25**

10:00 – 13:30  **Palermo Walking Tour**

9:00 – 21:00  **Excursion to Segesta – Trapani—Mondello**

---

**Sunday, June 26, 2016**

10:00 – 12:30  **Palermo Street Art Tour**
Un populu
diventa poviru e servu
quannu ci arrubbanu a lingua
addutata di patri:
è persu pi sempri.

A people
becomes poor and servile
when the language inherited from their forefathers is stolen from them;
they are lost forever.

Ignazio Buttitta (1899-1997), Sicilian poet
(Translated by Alissandru Caldiero)
In the assessment of spoken language ability, social interaction figures in two ways: as the format of assessment, and as the target of assessment. It is the format in the case of interactive speaking tests (OPI, IELTS, OET), and research has shown that the interactional organization of speaking tests can affect test takers’ performance and impact the validity of decisions based on their speaking scores (Brown, 2005; May, 2009; Nakatsuha, 2011). In addition, interaction itself is increasingly becoming the target of assessment, with scores being interpreted as indicating test takers’ interactional and pragmatic competence (Grabowski, 2013; Roever, 2014; Youn, 2015).

The central role of interaction in tests of spoken language ability raises the question of how interaction can be usefully conceptualized and analyzed for language assessment. This two-day workshop offers an introduction to conversation analysis (CA) (Sacks, 1992; Schegloff, 2007; Sidnell & Stivers, 2013), a transdisciplinary theory and methodology of interaction. CA provides a coherent approach to the analysis of interaction in ordinary conversation and institutional settings, including language assessment contexts. CA has generated a large cumulative literature on the organization of interaction in an increasingly wider range of languages. There is also a fast growing research direction that applies CA to the study of second language and multilingual learning and development inside and outside of classrooms (Kasper & Wagner, 2014).

The first day of this two-day workshop will introduce participants to the fundamental organizations of interaction (turn-taking, sequence organization, preference organization, and repair) and the transcription of talk. We will use examples from casual conversation, institutional interactions, and language assessment. The second day will apply conversation analysis to investigate the organization of interaction in speaking tests and the assessment of L2 speakers’ interactional competence. Specifically, we will discuss the following issues:

1. To what extent is assessment talk representative of “real-world” talk? Can raters and interlocutors affect scores in unpredictable ways? These questions have serious implications for the extrapolation inference in Kane’s (2006, 2013) argument-based approach to validation. Should assessment talk be fundamentally different from “real-world” talk, inferences based on scores from interactive speaking tasks would
be questionable.
2. What are the markers of interactional competence? How can different levels of interactional abilities be determined?

We will investigate indicators of different degrees of interactional competence from a conversation analytic perspective, including (preference) organization of interaction, internal structure of turns, linguistic formatting of social actions, and others. For assessment of interactional competence to move forward, it is important that indicators of interactional competence can be identified in assessment talk and rated. Participants will be enabled to critically examine research on interaction in second language assessment, and gain fundamental skills for carrying out such research in the future.

References
Grabowski, K. (2013). Investigating the construct validity of a role-play test designed to measure grammatical and pragmatic knowledge at multiple proficiency levels. In S. Ross & G. Kasper (Eds.), Assessing second language pragmatics (pp. 149-171). New York: Palgrave.

Carsten Roever (Ph.D., University of Hawai‘i at Mānoa) is Associate Professor in Applied Linguistics at the University of Melbourne. His research interests are second language assessment, especially of pragmatics and interaction, interlanguage pragmatics, quantitative research methods, and conversation analysis. He has co-authored “Testing ESL sociopragmatics” (2014, Peter Lang) with Catherine Elder and Catriona Fraser, and “Language testing: The social dimension” (2006, Blackwell) with Tim McNamara, as well as several journal article and book chapters on second language interaction and assessment of pragmatic competence.

Gabriele Kasper is Professor in the Department of Second Language Studies at the University of Hawai‘i at Mānoa. Her research applies conversation analysis and cognate perspectives to second language talk, learning, and assessment, to cognition and emotion in interaction, and to standard research methods in applied linguistics. Her most recent book is Assessing Second Language Pragmatics, co-edited with Steven Ross (2013, Palgrave Macmillan).
Workshop 2

**Methods and Techniques for Large-Scale Test Construction/**
**Metodi e tecniche per la costruzione di test**

*Roberto Ricci, Istituto nazionale per la valutazione del sistema educativo di istruzione
e di formazione (INVALSI)*

The construction of standardized tests is the result of rigorous technical and scientific processes. INVALSI follows such processes in constructing its own tests by using Rasch models and Item Response Theory.

The aim of this workshop is to illustrate the procedures followed by INVALSI for the construction of its own tests. It is also to highlight the issues that need to be addressed and solutions found to give scientifically appropriate answers to the aforementioned issues.

*This workshop will be given in Italian.

Roberto Ricci is responsible for research and coordinates the national system of construction of INVALSI standardized tests in Italy. He is a member of the Pisa Governing Board and General Assembly of the International Association for the Evaluation of Educational Achievement (IEA). He has over fifteen years of experience constructing standardized tests and using statistical procedures for data analysis in education. Roberto has a PhD in statistical methodology for scientific research.

Workshop 3

**Multi-Level Modeling in SLA and Language Assessment**

*Steven J. Ross, University of Maryland*

Multi-level models (MLM) have been increasingly used in applied linguistics, SLA, and language assessment to examine the influences of context, sampling variation, and bias on outcomes of interest. The research literature in these domains features a growing number of applications of MLM to model diverse phenomena such as
peer influences on motivation, rater training and experience, factors affecting item response patterns, participant and item effects on psycholinguistic experiments, and environmental factors affecting language proficiency growth trajectories.

The primary goal of the current workshop will be to provide participants with hands-on experience in constructing data sets amenable to MLM, and with sample data analyses, using downloadable software, HLM 7 for Students, and R. Data sets will be made available for practice analyses of various phenomena such as immigrants’ community enclosure effects on language proficiency, perceived peer aspiration on class-level language learning growth variation, socio-metric factors affecting the validity of peer assessment, and multi-level Rasch analysis.

Participants should bring their own personal computer with HLM 7 (free student version) and R (lmer4 package) installed.

**Day 1: Morning**
We will begin by focusing on the conceptual rationale for the MLM approach. The difference between aggregate group-level effects versus random effects modeling that identify variance components associated with participants, items, or contexts. This session will also examine how data sets amenable for MLM are constructed in common software such Excel, SPSS, or Systat.

**Day 1: Afternoon**
Using examples featuring nested effects for school composition, the first analysis in this session will model cross-level effects showing how student level individual differences can be conditioned on the dominant L1 of the school. A two-level hierarchical linear model using HLM 7 will be featured in this analysis. A second example will feature contextual factors influencing language learning growth trajectories of learners nested in classes and schools. For this analysis, a three-level growth model will be formulated with student factors at level 2 and class factors at level 3.

**Day 2: Morning**
A quick review of Day 1 will precede an extension of the multi-level framework to linear mixed effects models using R (lmer4). Featured will be an analysis of the validity of peer assessment, focusing on socio-metric indicators such as seating patterns and self-section into group level projects as possible factors injecting bias into peer assessments of student presentations.

**Day 2: Afternoon**
The final session will focus on either multi-level Rasch analyses devised to assess if test-taking variables at level 2 mediate item response probabilities, and/or on analyses of reaction time experiments with items nested under study participants. Workshop participants will practice using R (lmer4) applications for assessing experimental main effects while allowing for variation across random intercepts and slopes at the person and item level.
Steven J. Ross is Professor of Second Language Acquisition at the University of Maryland’s School of Languages Literatures and Cultures, and Senior Research Scientist at the Center for the Advanced Study of Language. He has taught research methods and language assessment in Japan, Mexico, Australia, and the USA, and has served on the editorial boards of Applied Linguistics, TESOL Quarterly, Language Assessment Quarterly, Language Testing, and the Australian Review of Applied Linguistics. He is currently Associate Editor of Language Learning.

Monday, June 20 / Tuesday, June 21
9:00 - 16:30 / 9:00 - 16:00
Room 101
Piazza S. Antonino, 1

Workshop 4

**Scenario-Based Language Assessments in Educational Settings: Theoretical Foundations, Prototype Examples and Future Applications**

*John Sabatini, ETS
Tenaha O’Reilly, ETS
James E. Purpura, Teachers College, Columbia University*

In recent years, advances in technology, theory, and educational practice have opened the door for innovations in language assessment design and delivery. Simultaneously, the ways in which people communicate, collaborate, and apply problem-solving skills in real life contexts has changed significantly. Despite these changes, however, the assessments used to measure these skills have not kept pace with how people apply their language knowledge, skills, and abilities (KSAs) in real world settings. This workshop will explore efforts to leverage the scientific literature, technology, and good teaching practice to rethink the assessment of communicative language ability in general, with a particular focus on reading comprehension. Special attention will be given to describing scenario-based assessment approaches that are designed to not only to measure the examinees’ KSAs in language performance, but also to impact student learning. Finally, this workshop will discuss the challenges of refocusing assessment that has been traditionally centered on measuring achievement, to shifting the focus on supporting learning.

The current workshop will provide participants with a basic understanding of an emerging capability in assessment design, called scenario-based assessment. SBA is a collection of techniques designed to contextualize the assessment experience, leverage technology to expand assessment opportunities, encourage, and support deeper learning and to gather more information about what students know and can do.
Organization of the workshop

Part 1 will begin with a discussion of how current language assessment theory and practice have often not kept pace with addressing the measurement of a broader range of competencies needed to accommodate the increasing demands of people in education and the workforce. In other words, the first part of the workshop will provide a theoretical and practical rationale for the need for “something different” and for the use of scenario-based assessments as one way of addressing some of the current language assessment shortcomings.

Part 2 will then provide an overview of one assessment framework that utilized a scenario-based assessment design. Using a concrete example of an SBA, this review will cover the theoretical and empirical foundations that motivated the design, the constructs, and basic principles underlying scenario-based assessment. It will also discuss the potential tradeoffs incurred when developing SBAs.

Part 3 will provide further examples of scenario-based language assessments from two projects, each intended for summative use. The first project is designed to create a range of assessments intended for use with children ranging in ages 8 to 18. The second focuses on the design of a language placement exam for students entering an adult language program at a university. Although both projects have different origins, they share common design principles related to SBA and to a learning-oriented approach to language assessment.

Part 4 will use the examples from Part 3 to provide a structure for participants to work collaboratively to design their own SBA. We will discuss principles and concepts such as the goal orientations, scenario narratives, component skills, reasoning skills, performance moderators, etc.

Part 5 will provide an overview of some formative assessments, which relax many of the constraints that exist in summative contexts. This part of the workshop will offer a collaborative opportunity to provide feedback on the approach, discuss potential assessment designs, and offer alternate applications of the concepts and capabilities.

After completing this workshop, participants should have a basic understanding of key concepts in scenario-based assessment design and some of the tradeoffs associated with making various design decisions. Participants will also have a basic understanding of the key theoretical and empirical issues in reading comprehension that can potentially improve assessment. Participants will have a familiarity with the key design principles used to create scenario-based assessments and will recognize those concepts in real examples. Participants will have some hands-on experience in designing an SBA. Finally participants should gain an appreciation of the challenges of implementing new approaches in real settings as well as have the opportunity to think critically and creatively for solutions to problems associated with change and innovation.
**Pre-Conference Workshop Abstracts**

**John Sabatini** is a Principal Research Scientist in the Center for Global Assessment at ETS. His research interests include reading literacy development, disabilities, assessment, cognitive psychology, and educational technology, with a primary focus on adults and adolescents. He has been the principal investigator of an IES-funded grant to develop pre-K -12 comprehension assessments as part of the Reading for Understanding initiative. He is co-investigator on projects exploring the reading processes of adolescents and English language learners, and co-lead of the ETS CBAL ELA project. He provides technical consulting to national and international surveys including PIAAC and PISA.

**Tenaha O’Reilly** is a Senior Research Scientist at ETS in the United States. He has been the principal and co-principal investigator on a number of research and development projects geared towards creating innovative assessments of reading. He is interested in designing formative and summative assessments of reading that utilize the cognitive science literature to produce assessments are both meaningful for students useful for instruction. He serves as the co-PI on the assessment strand of IES’s Reading for Understanding initiative and has consulted with testing programs such as NAEP, TOEFL, and PISA.

**James E. Purpura** is an Associate Professor of linguistics and education in the Applied Linguistics and TESOL Program at Teachers College, Columbia University. Jim’s research has focused on grammar assessment and the cognitive underpinnings of assessment. More recently, he is interested in learning-oriented assessment (LOA) as a means of understanding how learning can be incorporated into assessment, and how assessments function when embedded in teaching and learning. He teaches a seminar on designing scenario-based L2 placement exams from an LOA perspective. He is currently Editor of Language Assessment Quarterly, and is series co-editor of New Perspectives on Language Assessment and Innovations in Language and Learning Assessment at ETS. He served as President of ILTA in 2007 and 2008.

---

**Monday, June 20 / Tuesday, June 21**

**Workshop 5**

**Understanding and Using Standard Setting in Language Testing: A Practical Foundation**

*Dorry M. Kenyon, Center for Applied Linguistics*

*David MacGregor, Center for Applied Linguistics*

Standard setting methodologies provide useful tools to link specific examinee performances on language tests to more generalizable verbally-defined descriptors of performance levels. They offer explainable, replicable procedures to supply evidence in support of claims about test performances.
The goal of this two-day workshop is to provide participants with a foundation for understanding, planning, and conducting studies using standard setting methodologies for this purpose. Knowledge of, or even use of, processes suggested to link exams to the CEFR as described in Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment. A Manual (2009) will be helpful, but not necessary. The workshop will seek to contextualize standard setting in language tests into a broader framework that will help participants better understand the issues related to standard setting and how to address them.

**Part 1** defines standard setting and contextualizes it within the larger framework of test validity argumentation, specifically Bachman and Palmer’s Assessment Use Argument. In particular, the various approaches to standard setting will be examined in terms of providing evidences to claims related to interpretations of test performances, decisions about test takers, and consequences of test use.

**Part 2** reviews common features of conducting standard setting, including the selection and training of panelists, the selection of an appropriate standard setting methodology given the context (including decisions about providing feedback to participants), preparation of materials for the standard setting activity, and the collection and documentation of evidences supporting claims about the standard setting process and its outcomes. Above all, strengths and weaknesses of various options in addressing issues related to these features will be explored and discussed.

**Part 3** focuses more concretely on how to analyze and interpret data collected through the various methods of collecting panelists’ judgments. Using several illustrative data sets, approaches to the analysis will be practiced. The importance of matching analytical approaches to the larger interpretive framework of claims and evidences related to the standard setting process will be illustrated. Again, strengths and weaknesses of various options related to data analysis will be explored and discussed.

**Part 4** will provide a practical recap of the workshop material by working, in small groups, through one or more case-studies (depending on the number of participants) related to the presenter’s experience in conducting standard setting studies. Participants will be given a description of the case study, develop claims and approaches to collecting evidences through applying standard setting methodology. They will then need to make and justify their decisions about their selected approaches and describe the expected outcome. As time allows, the presenter will discuss what actually was done in the specific case-study and participants can discuss strength and weaknesses of the actual approach used.
Dorry M. Kenyon has experience in all aspects of designing, developing, validating, and operationalizing language tests through many large projects at the state, national, and international level. He joined the Center for Applied Linguistics (CAL) in Washington, DC in 1987 and currently serves as CAL’s Vice President, Director of Assessment. At CAL, he has directed a variety of projects related to the large-scale, high-stakes assessment of the second and foreign language skills of language learners from pre-school to adult and conducted over 30 standard setting studies using a variety of methodologies for a variety of contexts. With a doctorate in Educational Measurement, he has also served as CAL’s chief psychometrician. Dr. Kenyon is particularly interested in the application of new technology to language assessment problems. Prior to joining CAL, he taught German and English as a Foreign/Second Language for seven years in the United States and abroad.

David MacGregor has experience working on various projects involving all aspects of test development. He has worked at the Center for Applied Linguistics (CAL) since 1998, first as a research assistant for the National Capital Language Resource Center and, since 2000, in language testing. He has served as the director of the Psychometrics/Research team for ACCESS for ELLs®, as well as the lead developer of PODER, a Spanish academic language test for grades K-2. He is currently Senior Research Associate, with a focus on the intersection between psychometric research and test design. Over the years, he has participated in several standard setting studies using a variety of methodologies. Dr. MacGregor holds a B.A. in Linguistics from the University of Chicago, an MAT in TESOL and Spanish from the School for International Training, and a Ph.D. in Linguistics from Georgetown University. He is especially interested in the theory of Cognitive Linguistics and how it applies to the understanding of the construct of academic language. Prior to joining CAL, he taught English as a Foreign Language and Spanish for ten years in Turkey, the United States, and Spain.
Lingua e Dialettu

Un poviru,
c’addatta nte minni strippi
da matri putative,
chi u chiama “figghiu”
pi nciuria

Nuàtri l’avevamu a matri
nni l’arrubbaru;
aveva i minni a
funtani di latti
e ci vippiru tutti,
ora ci sputanu.

Nni ristò a vuci, d’idda,
a cadenza,
a nota vascia
du sonu e du
lament:
Chissi non nni
Ponnu rubari.

Non nni ponnu
rubari
ma ristamu poviri
e orfani u stissu.

Language and Dialect

I’m a poor man,
who suckles the barren breast
of a so-called mother
who calls him “son”
as an insult

Once we had a mother,
they stole her from us;
her breasts were
fountains of milk,
And everyone drank from them,
Now they spit on them.

For us her voice remains,
the cadence
the low note
of it music and
the sorrow:
of these
they couldn’t rob us.

They couldn’t
rob us of these,
but we’re poor
and orphaned just the same.

Ignazio Buttita (1899-1997), Sicilian poet
Last five stanzas of “Lingua e dialettu,” written in 1970. (Translated by Alissandru Caldiero)
(Cited in Seeking Sicily by John Keahey)
The Dynamics of Test Impact in the Context of Australia’s Skilled Migration Policy: Reinterpreting Language Constructs and Scores

Kellie Frost, University of Melbourne

The use of language tests as tools of immigration policy raises critical questions concerning the adequacy of current conceptualisations of test impact and validity in the field of language testing. Within existing theory, these notions rest on the premise that language constructs and associated test score meanings exist as coherent, fixed and stable entities. Test developers and test users are assumed to be in consensus concerning how test scores should be understood and used, and when tests and test scores are used as intended, the premise underlying evaluations of consequences is that positive effects on individuals and societies should outweigh or at least justify any negative effects. Such underlying assumptions, it is argued in this paper, limit our capacity to understand and explain the role of language tests in immigration (and other) policy domains, where, as will be demonstrated here, score meanings, test purposes and test consequences are dynamic, multiple, potentially conflicting and inherently unstable. This paper explores the dynamics of test impact in the context of Australia’s skilled migration policy by examining how four individuals responded to the language test score requirement that exists within the policy as they sought to transition from temporary visas to permanent residency status. Temporary visa holders in Australia are able to become permanent residents via the skilled migration program if they possess certain specified skills and attributes, including English language proficiency as demonstrated on an accredited language test, such as IELTS. In recent years, test scores of IELTS 7 and 8 have been heavily weighted in the overall selection process, and thus represent a vital transition mechanism in the trajectory of those seeking to move from the status of temporary to permanent resident in Australia. A series of in-depth, open-ended interviews were conducted over an 18-month period with four participants during their migration trajectory from temporary to permanent residency, a transition that for each of them involved repeated language test attempts. A grounded theory approach guided analysis of interview data, which focused on identifying how participants interpreted score meanings throughout their trajectory, how they perceived test purposes, if and how their perceptions changed over time as they interacted with the test, how their thoughts and feelings influenced their behaviour, and the ways in which their actions and decisions produced test consequences. Findings show that test impact is co-constructed by test takers in dynamic and unpredictable ways as they act and react according to changing perceptions of score meanings and test purposes in their attempt to practice agency under the constraints imposed by policy. Their
shifting perceptions produce and are a product of their actions and decisions in response to the testing regime, which in turn lead to shifting test consequences. The study evidences the need for a more nuanced understanding of the experiences of individuals as they engage with language testing practices in the context of their migration journeys, in order to develop theoretical frameworks that can account for the role and impact of language tests in these policy domains.

Tuesday, June 21
18:30-19:00
Palazzo Steri
Sala Magna

Plenary
Young learners’ Processes and Rationales for Responding to Different Types of Self-assessments

Yuko Butler, University of Pennsylvania

Self-assessment has gained substantial attention in recent years among educators of young language learners (defined as children up to 11-12 years old). Researchers have developed various types of self-assessments including Can-do Speech Bubble in Language Portfolio in the U.K. (2006) and Lingua Folio Junior (2014) based on the ACTFL proficiency guidelines. In Japan, at the secondary school level, teachers of English are required to develop and implement self-assessments in their classrooms, and this policy is expected to be implemented at the primary school level soon. Despite the popularity of self-assessment at the policy level, however, teachers often express concerns about what exactly self-assessment captures, how accurately it captures learners’ abilities, and how best to capture such abilities. We have limited information about many aspects of self-assessment including: the relationship between a child’s age and the ability to self-assess one’s performance in a foreign language; how the context of self-assessment (e.g., wording, scaling, examples provided, and timing) influence their judgement; and how children make judgments (e.g., comparisons with other children, comparisons to previous performance, etc.).

For example, in responding to the item “I can greet in English,” we do not know what a child considers to be a “greeting,” what counts as evidence of “greeting,” and how children evaluate their performance.

This study investigates young learners’ processes of conducting self-assessments and seeks to understand how their responses and rationales may differ depending on different self-assessment contexts and age. The participants were 31 Japanese primary school children learning English. They were divided into a younger group (ages 8-9, N=17) and an older group (ages 10-12, N=14). The children were asked to complete four types of self-assessment: (1) a popular dichotomous format (selecting items from Can-do Speech Bubble); (2) a five-scale format (the same items as the first type but with responses provided using a five-point scale); (3) a five-point scale
format with concrete examples provided for each item; and (4) a five-point scale format attached to actual tasks. The last type of self-assessment was administered every time a child performed his/her tasks. The children’s processes and rationales of responding to each self-assessment item were investigated through think-aloud or guided interview procedures. The data was collected individually and it took approximately an hour per child. The children’s self-assessment responses and their interpretations and rationales for judgment were compared across four contexts and across children. The results indicate that there were substantial variations among children (regardless of their age) regarding how they interpreted each item and their rationales for judgment. The dichotomous form tended to invite more response bias and to lose much information, while the bases for responding to the 5-point scale items varied greatly among individuals. Providing concrete examples was not necessarily helpful and sometimes confused them. Overall, the children found the self-assessment attached to actual tasks to be the easiest to respond to. The data was interpreted based on self-efficacy theory (Bundura, 1997) and self-regulation (Zimmerman, 1990). The paper concludes with a number of implications for practice.

Parallel sessions

Wednesday, June 22
10:00-10:30
Aula Magna
University of Palermo, Scuola Politecnica - Faculty of Engineering

Examining the Validity of a Performance-Based Oral Test for University Lecture Certification

Slobodanka Dimova, University of Copenhagen

The rapid increase of English-medium instruction (EMI) programs across Europe has raised concerns regarding the oral competencies of non-native English speaking lecturers and the implications for the quality of teaching. Consequently, lecturers’ English proficiency is under scrutiny and universities are developing language policies for quality assurance, which are enforced by implementation of internal assessment procedures. Given the complexity of the local teaching and learning contexts in which these assessments are administered and used, answering the questions about whether and how to address the interface between language, disciplinary content, and pedagogy in the assessment procedure has been a struggle. This presentation will report on validation of an oral English certification test based on simulated lecture. The test is used for assessing university lecturers’ oral English proficiency in relation to the communicative demands of EMI and for providing them with extensive written and oral feedback on their language skills. Holistic scores and formative feedback reports (N=400) from six raters, and interviews with lectures (N=24) were used to analyze questions related to 1) consistency of scale...
use, 2) scalar level distinction, 3) reported lexical and organizational characteristics of disciplinary content in performances across different levels, and 4) lecturers’ opinions about the interplay between language and pedagogy. Rating data were examined using multi-faceted Rasch measurement (Linacre 2010), while formative feedback reports and interview data were analyzed in Nvivo10. Results suggest appropriate and consistent scale use, but differing severity levels among raters. Band distance at the upper scalar end was found to be too wide, yielding much variation within one level. According to the formative feedback reports, lecturers tend to experience difficulties with general, rather than discipline-specific vocabulary. Interview findings suggest that lecturers’ content knowledge and teaching experience facilitate their language performance. This presentation will end with a discussion about whether re-scaling procedures to fine-tune the scale levels are needed, especially for formative feedback purposes, and whether the linguistic, rather than classroom management, aspects of pedagogy are sufficient for inclusion in the underlying construct of the lecturer assessment procedure.

Wednesday, June 22
10:00-10:30
Room 6
University of Palermo, Scuola Politecnica - Faculty of Engineering

Double-rating Methods for Direct Speaking Tests: Construct(s) Measured by Live, Audio and Video Rating Modes

Fumiya Nakatsuhara, University of Bedfordshire
Chihiro Inoue, University of Bedfordshire
Lynda Taylor, University of Bedfordshire

It has long been suggested that double rating of spoken performance is essential to establish scoring validity for a speaking test and to ensure fairness to test-takers (e.g., AERA, APA and NCME, 1999). A review of rating systems currently employed in major international examinations indicated that double ratings of speaking tests for score reporting and quality assurance purposes are often carried out with audio-recordings of test-taker performance. However, rapid advances in computer technology over the past decade have made the gathering and transmission of test-takers’ recorded performances much easier in a video format. This seems a good moment, therefore, to investigate different modes of rating spoken performance, so that test providers have the necessary information for making informed decisions on appropriate rating methods for the future. This study compared IELTS examiners’ scores when they assessed test-takers’ spoken performance under live and two non-live testing conditions using audio and video recordings. It also explored examiners’ perceptions towards test-takers’ performance in the two non-live rating modes. This was a mixed-methods study that involved both existing and newly collected datasets. A total of six trained IELTS examiners assessed 36 test-takers’ performance under the live, audio and video rating conditions, and their scores in the three modes of rating were calibrated using the multifaceted Rasch model analysis. In all modes of
rating, the examiners were asked to make notes on why they awarded the scores that they did on each analytical category. The comments were quantitatively analysed in terms of the volume of positive and negative features of test-takers’ performance that examiners reported noticing when awarding scores under the three rating conditions. Using selected test-takers’ audio and video recordings, examiners’ verbal reports were also collected to gain insights into their perceptions towards test-takers’ performance under the two non-live conditions. The results showed that audio ratings were significantly lower than live and video ratings for all rating categories. Examiners noticed more negative performance features of test-takers under the two non-live rating conditions than the live rating condition. The verbal report data demonstrated how having visual information in the video-rating mode helped examiners to understand test-takers’ utterances, to see what was happening beyond what the test-takers were saying and to understand with more confidence the source of test-takers’ hesitation, pauses and awkwardness in their performance. In the presentation, a recommendation will be made as to enhanced double-rating methods that could be introduced to face-to-face speaking tests such as the IELTS Speaking Test. We will also discuss broader implications for the construct(s) assessed by different speaking formats in relation to the availability of test-takers’ visual information to examiners. This research study will therefore contribute to a better understanding of the extent to which raters, whether or not they also serve as an interlocutor, are co-constructing speaking test outcomes across different modes of rating, thus enabling better test specifications regarding raters’ roles in speaking tests (e.g., McNamara, 1997).

Wednesday, June 22
10:00-10:30
Room 7
University of Palermo, Scuola Politecnica - Faculty of Engineering

**Examining L2 Learning and Learning Outcomes as a Result of Embedded Classroom Assessments**

*James E. Purpura, Teachers College, Columbia University*
*Jorge Beltrán, Teachers College, Columbia University*
*Heidi Han-Ting Liu, Teachers College, Columbia University*
*Michelle Stabler-Havener, Teachers College, Columbia University*

In recent years, many assessment researchers (e.g., Davison & Leung, 2009; Hill & McNamara, 2012) have expressed the need to examine the critical role that teachers play in the use of assessments for learning in L2 classrooms. As a result, considerable research has focused on the teachers and teaching processes with studies investigating teacher assessment literacy and the role of teacher knowledge, experience, and beliefs in using assessments (Brindley, 2001; Cheng, Rogers & Hu, 2004; Gardner & Rea-Dickins, 2001, 2007; Rea-Dickins, 2003; Vogt & Tsagari, 2014), and teacher decision-making processes (Cheng & Sun, 2015). Some work has also
investigated learners and learning processes, specifically with relation to the use of diagnostic or dynamic assessment in promoting teaching and learning (Alderson et al., 2015; Lantoff & Poehner, 2011), and the value of self and peer assessment for fostering self-regulation, autonomy, motivation, and learner outcomes (Patri, 2002; Saito, 2008). However, many important questions remain as to how assessment relates to learning in L2 classrooms—a critical focus of classroom-based assessment. The current study used a learning-oriented approach (Turner & Purpura, 2016) to investigate how assessments, embedded in instruction and mediated through interaction, contributed to the promotion of learning processes and the achievement of learning success in two ESL classrooms. Video-recordings from 8 sessions were uploaded onto Nvivo, transcribed, and coded for several features. Then, assessment episodes were selected and examined through four lenses (i.e., instructional, interactional, proficiency, learning), with special attention paid to learning as it related to evaluative feedback, the quality of the assistance, the role of assistance in promoting learning processes, the cognitive load of the assessment, and the attainment of learning success. Learning patterns were tracked across the lessons with results showing that certain types of assistance seemed to influence the likelihood of uptake and learning success, thereby seeming to contribute to mixed learning trajectories.

Wednesday, June 22
10:00-10:30
Room 8
University of Palermo, Scuola Politecnica - Faculty of Engineering

Validation of an LSP Test Aligned with the CEFR

Joan Jamieson, Northern Arizona University
Daniel Isbell, Michigan State University
L. D. Nicolas May, CaMLA

Language for specific purposes (LSP) testing requires content and task specificity commensurate with a particular test’s use; localization is argued to be effective in accomplishing this requirement (O’Sullivan, 2012). However, LSP test score users often desire scores suited for specific uses to simultaneously be situated in more general frameworks of language competence. This tension, and how it affects arguments for validity, is explored in this presentation. In an argument for test score interpretation and use (IUA; Kane, 2013), a clear description of the interface between context and content provides evidence warranting an extrapolation inference. An LSP test naturally demonstrates this interface, with context-derived tasks and content-based input. But for many workplace assessments, the construct remains elusive. When lacking a strong theory, Kane (pp. 21-22) described the importance of observable attributes (performance tendencies that reflect knowledge, skills, and strategies) in lieu of a specific trait or theory. While not a theory, the CEFR includes a vast typology of language use in terms of competence across tasks, texts, and purposes. By linking examinees’ performance to the CEFR, a test score can have
added meaning in terms of a criterion-referenced interpretation. Additionally, doing so addresses the desire to situate scores globally. Test localization and CEFR alignment principles guided the development and piloting of the Test of Elastomer English (TEE). TEE’s purpose, defined by test users, was to assess trainees’ ability to understand and communicate essential elastomer concepts, processes, and relationships in English at three CEFR levels (A2–B2). To define test content, consultation with elastomer experts yielded four content themes related to basic physical properties of elastomers. Consideration of work context and tasks requiring English led to targeting fourteen of CEFR’s linguistic competencies. Test developers prototyped test tasks that would align both with the training content and CEFR competencies and levels. Review by elastomer experts informed selection of pilot tasks. Detailed item guides and test specifications documented the development process. TEE was piloted with 120 elastomer trainees in July, 2015. For TEE’s IUA, four inferences were examined. The scoring inference was supported by consistently applying the scoring rules, explained and justified in item guides. Generalizability was another inference, supported by estimates of internal consistency (0.91), inter-rater reliability (0.75), the separation index and its reliability (>2, 0.90). The extrapolation inference, or whether the test reflected the goals of the vocational program and future workplace English demands, was also investigated. TEE’s localized development and the careful following of CEFR alignment procedures provided positive evidence. Finally, the criterion-referenced interpretation inference that tests scores reflected levels of selected CEFR competencies was supported through test development, examination of predicted levels and empirical difficulties, and setting cut scores (Kaftandjieva, 2010). Taken together, the evidence supported the claim that the TEE scores represented trainees’ ability to understand and communicate essential elastomer concepts, processes, and relationships in English, as described by the CEFR. The project illuminates the possibility of relating a specific, localized language test to a general framework of language competence.

Parallel sessions

Wednesday, June 22
11:00–11:30
Aula Magna
University of Palermo, Scuola Politecnica – Faculty of Engineering

Assessing Low Educated Learners – Implications for the Test Construct for Large-scale Assessment

Cecilie Hamnes Carlsen, Vox – Norwegian agency for lifelong learning

Adult second language learners with limited literacy skills have long formed part of the immigrant population. As requirements for citizenship and permanent residency are becoming increasingly contingent on language proficiency, low educated L2
learners (LEL2s) are starting to form part of the test population as well (Extramiana, Pulinx & Van Avermaet, 2014). Nevertheless, this most vulnerable group of second language learners have attracted relatively little interest in SLA and language testing research to date (Tarone, 2010, LESLLA homepage). Defining and operationalizing the language test construct so that LEL2s get a fair chance to show their language abilities confront test developers with some new challenges. LEL2s do not only lack in general literacy – they also lack in test literacy (Allemano, 2013). Several neuropsychological tests have shown that LEL2s perform significantly lower on verbal tests than test takers with more schooling (Kim et al., 2014). As Ostrosky-Solis et al. (1998) show, the low test scores may be caused not (only) by lack in the ability being tested, but also by the lack of experience with the testing situation. In this paper, I will first show how we worked qualitatively to reduce construct-irrelevant variance caused by the lack of literacy and testing experience in task development and in scoring of a high stake digital test of Norwegian for adult immigrants. Thereafter, I will show the results of multiple regression analysis, bias analysis, and variance analysis (ANOVA) comparing LEL2s (N=2083) with test takers with higher levels of education to see whether the test does indeed give LEL2s a chance to show their abilities. Finally, I will present the results of a small-scale quantitative study mapping the opinions and attitudes of teachers of LEL2s (N=60) about the suitability, washback effect and consequences of the language test in question on LEL2s. This paper intends to remedy the lack of focus on the LELs in the testing situation and the challenges encountered by the test constructors striving to increase test validity by decreasing the effect of construct irrelevant variance caused by test takers’ lack of general literacy and test literacy.

Wednesday, June 22
11:00-11:30
Room 6
University of Palermo, Scuola Politecnica - Faculty of Engineering

Validation of a Rating Scale for the Assessment of Summary Writing within an Argument-based Framework

Woranon Sitajalabhorn, University of Melbourne

To point out the importance of rating scales in writing assessment, Hamp-Lyons (2011) states that rating scales are “realizations of theoretical constructs, of beliefs about what writing is and what matters about writing” (p. 3). In other words, rating scales embody the skills/abilities that are intended to be measured by the test (McNamara, 2002). Given this critical role of scoring rubrics, scholars take it as axiomatic that rating scales play a central part in determining the validity of writing assessment (McNamara, 1996; Weigle, 2002). Therefore, it should go without saying that the validity of rating scales themselves is not second to that of the test. Despite the valid and carefully-designed writing task, the use of an invalid rating scale in the scoring process can compromise the credibility of test scores and the validity of test score interpretations and uses. Whilst it is acknowledged that the validity of rating
scales is vital and that their empirical validation should be conducted (McNamara, 1996), scale developers pretty much find themselves in a cul-de-sac when they attempt to validate their rubrics. Validation methods proposed thus far (e.g., Bachman and Palmer’s (1996) test usefulness model, Chapelle, Enright, and Jamieson’s (2008) argument-based approach, or Bachman and Palmer’s (2010) assessment use argument) were primarily developed for the validation of language tests and their uses/consequences rather than for the validation of such test instruments as rating scales. With the absence of clear validation guidance, how rating scales should be validated is conveniently left to the researchers’ discretion, making it difficult to compare the results from different studies (see Knoch, 2009; Shaw & Falvey, 2008; Shaw & Weir, 2007 for example). To address the lack of an approach to scale validation, Knoch and Chapelle (in preparation) draw up comprehensive guidelines for the validation of rating processes based upon Chapelle et al.’s (2008) argument-based framework. For each inference in the chain of argument, they provide warrants, assumptions, and sources for backing for researchers striving to validate scales and scoring processes. This paper presents the application of Knoch and Chapelle’s (in preparation) framework to a summary writing rating scale, developed using mixed methods of scale development (writing theories/models, evidence of examinees’ written performance, and experts’ intuitions). An overview of the rating scale and its development will be given. To validate the scale, five raters were asked to use the scale to evaluate 50 summaries, complete a questionnaire, and participate in an interview. Rating results were analysed quantitatively using the many-facet Rasch measurement approach, and the questionnaire and interview responses were analysed qualitatively. Then, how the data were utilised to support the assumptions of each warrant of the seven inferences will be discussed. The results reveal strengths and weaknesses of the scale. Detailed and structured, Knoch and Chapelle’s (in preparation) framework considerably contributes to the scale validation because it not only delineates what evidence needs to be gathered but also offers options for examining such evidence. The framework also aids the scale revision since it clearly indicates areas that still require improvement.

Maximizing Feedback for Learning: Investigating Language Learners’ Differing Cognitive Processing when Receiving Computer-based Feedback

Maggie Dunlop, OISE / University of Toronto

This paper reports a study that investigated the cognitive processes taking place when adult immigrant English as a Second Language (ESL) learners in Canada received cognitively diagnostic computer-based feedback about their English
language skills. In particular, the study examined how these processes differ among students with different learner characteristics (i.e., ESL proficiency level, goal-orientations (Dweck, 1986)). The study contributes to current theories of learner-feedback interaction in second language learning, a central component of the assessment-for-learning practices (Hattie & Timperley, 2007) that are a key feature of effective foreign language instruction. The objectives of this study were to investigate how language learners engage with and process feedback differently, to develop learner feedback processing profiles for use in personalizing feedback for learners. The study adopted a quasi-experimental research design and utilized mixed methods. Data was gathered from 112 adult immigrant English as a Second Language learners in Canada, and used a CELPIP reading test, a Canadian test of English for immigration purposes, adapted to identify current learner abilities in key language skills according to cognitive diagnostic assessment principles (Leighton & Gierl, 2007). Study participants received their feedback report and planned their learning, all online, then completed a survey designed to investigate aspects of attention, processing, interaction, and application. The report included an introduction, skill descriptions, comparisons between assessment and self-assessment results, study suggestions and planning activities. Fifteen participants also received their reports while using an eye tracking device, then completed a stimulated recall interview to elicit information on participants’ attention, processing and interaction. Finally, 15 participants participated in a follow-up interview one month later to elicit information on ongoing processing, interaction and application. The study analyzed and triangulated survey and interview data using iterative coding methods (Creswell and Plano Clark, 2007), latent trait profiling methods (Hagenaars & McCutcheon, 2002), regression analyses and case studies, and identified four feedback profiles based on extent of trust in the report’s diagnosis, amount of reflection taking place, desire to use the report, and confidence in ability to use the report. Profiles supporting effective feedback processing were associated in different ways with low emotional reactivity, higher mastery and performance learning orientations and low performance avoid orientations, and greater and targeted use of metacognition in reflection. Higher and lower English language skills were both factors in at-risk profiles through lack of engagement, but for differing reasons. The study also analyzed eye tracking, interview and survey data using case studies, analyses of variance, and latent trait profiling to assess report feature usage. Discrepancies in assessment and self-assessment results prompted in-depth reflection. Participants with emotional responses emphasized scores. Those with more effective feedback processing strategies used skill descriptions and critically engaged with study suggestions. Plan effectiveness was related to factors such as greater metacognition, mastery orientations and language skills. This study identifies learner profiles representing the cognitive processes by which language learners engage with making meaning of descriptive, process-oriented computer-based feedback. Future applications include developing personalized computer-based feedback systems that provide feedback that learners are likely to use.
Higher education institutions normally associate proficiency in the language of instruction with successful performance in various academic programs. Therefore, language tests used as admission requirements to higher education institutions are high stakes tests whose scores greatly impact the universities’ decision to admit non-native speaking candidates to their various programs. Previous studies on the predictive validity of such tests have, however, mainly focused on standardized English language tests as measures of L2 proficiency (Elder, 1993; Morris & Cobb, 2004). Few studies have examined the predictive validity of the admission tests that measure the candidates’ proficiency in their L1 (Romainville, 1997; Blais, 2001). This study examines the predictive validity of the Test de Français Laval-Montreal (TFLM), a high-stakes French language test administered for admission and placement purposes to French-speaking candidates applying to Teacher-Training Programs (TTPs) in major francophone universities in Quebec, Canada. Adopting a convergent parallel mixed-method design (Creswell & Plano-Clark, 2011), the study is conducted in two phases. Phase I features a comprehensive task analysis detailing the characteristics of TFLM tasks as opposed to those of three major TLU academic and instructional contexts linked to the test. The results of this phase of the study are discussed qualitatively and in terms of the standards of validity (Messick, 1996), Messick’s conception of construct validity (1980), and Bachman & Palmer’s Framework of Task Characteristics (2010). In Phase II, quantitative data —test scores, GPA after year 1, GPA after the first and the second remedial French courses, and the final GPA—is gathered for 145 participants at different intervals during their undergraduate program. In this phase, information is also gathered, through a questionnaire, with respect to the participants’ perceptions of the test. The findings of Phase I reveal that TFLM tasks and constructs do not represent those of the TLU contexts and do not address the language needs of pre-service teachers as identified by the Quebec’s Ministry of Education. The statistical analysis of Phase II data (multiple regression, paired t-test, one-way ANOVA) further shows that the admission test does not predict learners’ success in TTPs. The findings also suggest that the remedial French courses offered to students with lower test scores do not help them catch up with those with higher scores. Data triangulation using the results of Phase I and the questionnaire further confirms that the test tasks and constructs do not represent the real-life language needs of the students in TTPs. Implications of these results for the validation of admission tests measuring test-takers’ L1, as well as the consequential aspect of TFLM validity and the potential negative consequences of using TFLM as an admission test are discussed.
After the Entrance Test. A Longitudinal Study of L2 Students’ Experiences at University

Bart Deygers, University of Leuven

University entrance language tests are often administered under the assumption that even if language is not a guarantee for academic success, a minimum level is required to successfully participate in academia. In Europe this entrance level is most frequently set at B2 of the CEFR and it is commonly understood that this is a minimum requirement, which learners will transcend during their academic career. The longitudinal study presented in this paper questions existing assumptions and has implications for how we conceptualize and operationalize university entrance requirements for L2 students. Previous predictive validity studies do not generally find strong relationships between academic performance and language test scores, and that is also the case in the current study. To date however, little research has focused on how well both successful and unsuccessful testing candidates cope linguistically in the target language use context. Even fewer studies have taken a qualitative, longitudinal perspective on this. This paper describes the linguistic development and experiences of 23 L2 students at three Flemish universities during the academic year 2014 - 2015. All respondents had passed at least one L2 university entrance language test. The data were collected during bi-monthly semi-structured in-depth interviews, resulting in over 130 hours of recorded speech. The interviews were transcribed, and all transcriptions were coded using NVivo for Mac. The results presented in this paper first highlight in what sense the entrance tests accurately operationalize the real-life expectations in terms of proficiency level and task design. They show in what sense the constructs of the tests invite construct underrepresentation or construct irrelevant variance. Secondly, this study shows what limited linguistic progress the respondents made during one year in a Dutch-medium context. Progress was monitored both through the analysis of interview data and by comparing scores in a within-subject test-retest study, which did not find any significant differences between scores on identical tasks, administered at the start and at the end of the academic year (t = -1.093, df = 18, p-value = 0.29). This paper confirms that studying in an L2 context does not automatically result in increased language proficiency in the fields that matter to academic language use. Consequently, it may not be in the best interest of the candidate to administer tests that are below the level that is actually required. Lastly, this study shows that there is much more to academic success than language ability, suggesting that the construct of academic language as it is defined now, might be somewhat narrow.
"I can see that": Developing Shared Rubric Category Interpretations through Score Negotiation

Valerie Meier, University of California Santa Barbara
Jonathan Trace, University of Hawai‘i at Mānoa
Gerriet Janssen, University of Hawai‘i, Mānoa & Universidad de los Andes, Colombia

Performance assessments using human raters will always contain some degree of subjectivity, and disagreement among raters necessitates valid and reliable methods for resolving discrepant scores. Previous research on a high-stakes writing assessment (Trace, Janssen, & Meier, 2015) has shown that when raters disagree, score negotiation is one potentially effective method to guide scoring decisions and reduce raters’ tendencies to be unexpectedly severe or lenient when scoring specific rubric categories or test takers. Beyond its utility for scoring, however, rater negotiation in performance assessments also has important implications for raters’ definitions of the constructs the rubric purports to be measuring. As raters may assign individual meanings or values to even well-described rubric categories, these differences can lead to multiple interpretations about examinee performance that can ultimately influence the reliability and validity of the assigned scores and test uses. In the process of discussing their decision practices, however, raters are provided the affordance to create a shared sense of meaning for the language constructs being measured and co-construct interpretations about examinee performance. In other words, the sum of negotiation becomes greater than its individual parts (van Lier, 2004). The current study examines this process by using quantitative and qualitative methods to trace how negotiation impacts raters’ scoring decisions over time and examine in detail how raters develop a joint interpretation of rubric category criteria. Scores from the writing section of a high stakes English language placement exam (N = 80) rated by two program outsiders were used for the basis of this study. Both negotiated and non-negotiated scores were analyzed using ANOVA and many-faceted Rasch measurement to determine which rubric categories were most frequently assigned discrepant scores and how this changed over time in terms of interrater reliability and agreement, as well as Rasch measures of severity and bias. Discourse analysis was then used to examine eight, one hour-long audiotaped rater negotiation sessions in order to determine how the two raters’ understanding of rubric category criteria converged over time. Our results indicate that through negotiation, raters came to use a shared set of terminology and justifications to clarify ambiguous rubric constructs; at the same time, they also developed new understandings of relevant constructs that went beyond the original rubric category descriptions. These results suggest that score negotiation not only influences scoring inferences from a test validation standpoint but also creates affordances for raters to ground those inferences in shared constructions of
meanings. A major implication of these findings is that score negotiation can be both an important step for ensuring the validity and reliability of performance assessment scores (e.g., reducing bias and limiting severity) and a resource raters can draw on to refine their understanding and application of constructs both in large-scale assessments and classroom teaching contexts.

Wednesday, June 22
14:30-15:00
Room 7
University of Palermo, Scuola Politecnica - Faculty of Engineering

**Applying a Concept Model of Dynamic Assessment to a Large-scale Speaking Test**

*Tziono Levi, Ort Network*

Standardized tests are used extensively to influence high-stakes decisions in the education field and specifically in the L2 context. Hence researchers and practitioners seek ways to align assessment practices with instruction. Many studies (e.g., Leung, 2004; McNamara, 2001; Purpura, 2004; Rea-Dickens, 2008; Shohamy, 1998; Turner, 2012) have indicated the centrality of assessment and the need to couple assessment practices with L2 teaching and learning. Indeed, in recent years various notions of formative assessment have been pursued to support student learning.

The initiation of assessment-for-learning (Black & Wiliam, 1998), teacher-based assessment (Davison, 2004), diagnostic assessment (Alderson 2006), classroom assessment (Rea-Dickins 2008), interactive assessment (Hamp-Lyons & Tavares, 2011) and most recently - learning-oriented-assessment - (Turner & Purpura, 2015), have generated practices that consider how learning and assessment can impact each other. Accordingly, based on Vygotsky’s Sociocultural Theory which accentuates means to promote thinking and learning, Dynamic Assessment (DA) has emerged as a paradigm connecting L2 learning with assessment emphasizing intervention to enhance learner development (e.g., Negueruela, 2003; Poehner, 2005, 2008). In an effort to optimize the learning-assessment relationship of oral language proficiency (OLP), this paper examines the application of DA to a large-scale speaking test attempting to up-scale the use of DA with larger groups of learners. Affixing DA to L2 OLP draws on Vygotskian theories that learning is primarily a socially meaningful cooperative activity (Kozulin, 1998; Minick, 1987) and the dialogic nature of most OLP tests helps to gain an understanding of test takers’ language performance level. In this study, seventy-three (73) Israeli 11th-grade high school students from three (3) public schools in periphery towns preparing for their matriculation EFL OLP test, underwent a single short-term DA mediation session following a pre-test in interview form. Teacher-testers guided mediation to two treatment groups of 25 students each, one relying on group-work and the other on individual self-examination, as they analyzed a video-recording of their pre-tests. Both groups were equipped with a theoretical-scientific-conceptual instrument (adaptation of Gal’perin’s model, 1967; 1992) aiming to materialize what the ‘speaking skill’ entails.
There was also a control group of 23 students. Post- and follow-up tests were given to all participants. Test scores were compared statistically while student verbalizations during mediation were examined qualitatively. Findings confirm that students who underwent the mediation scored better in the post- and follow-up tests than the control group for the measures of Communicative Ability (fluency) and Accuracy of speaking, suggesting that this form of DA mediation may have beneficial short and longer-term impact on OLP performance. Implications show that the controlled interaction or ‘collaborative-dialogue’ (e.g. Swain, 2000; Swain & Lapkin, 1998; 2001) used for stimulating, noticing and reflecting on OLP with a meta-cognitive tool led to improved L2 oral performance in a large-scale test setting. Thus theoretically, DA is proposed as a means to connect OLP learning and assessment while applying conceptual measures, so that academic achievements may improve with guided cognitive mediation. Consequently, DA contributes to our understanding of how assessment is aligned with learning to guide processes of learner L2 development.

---

Parallel sessions

Wednesday, June 22
15:05-15:35
Aula Magna
University of Palermo, Scuola Politecnica - Faculty of Engineering

Is the Large-scale Testing of English for Academic Purposes a Misnomer?

Jo Lewkowicz, University of Warsaw
Constant Leung, King’s College, London
Jennifer Jenkins, University of Southampton

Large scale tests of English for Academic Purposes (EAP) such as IELTS and TOEFL are considered a passport for entry to English-medium universities by students of non-English speaking backgrounds (NESB). However, there is a growing concern among interested stakeholders that these tests are (no longer) ‘fit–for–purpose’. They are proving to be insufficiently sensitive as predictors of academic performance. This paper explores some of the issues underpinning current EAP testing practices. It starts by considering test validity within the currently influential psychometric tradition in which the construct of any test is established a priori. In practice, therefore, the construct forms the basis for test item development and the validity of the construct is established by demonstrating that test scores have been interpreted appropriately (with reference to the defined construct). We argue that this view of construct validity gives rise to two epistemological problems: that of defining the construct in terms of what counts as relevant and secondly viewing validity as referring not to the properties of a test, but to the interpretation of test scores.
Thus, the measurement score reflects the amount of the construct residing in the test-taker. We suggest that this approach is overly performance focussed and is not able to capture the complexities of real life communication at any level, let alone the academic level. We then consider some of the complexities of language and communication which large-scale EAP tests fail to capture. We contend that the template used to inform much of language testing is both static and outmoded. It does not account for the changing linguistic landscape where linguistic diversity is the norm, and/or where English is a lingua franca and where more often than not interaction is among and between NNSs of English, hence judging their performance against an abstracted native speaker model is often no longer the most appropriate. The communicative competence model also fails to capture the nature of authentic lingua franca use which is constantly evolving and being adjusted as interlocutors co-construct meaning and discover ‘online’ the extent of their linguistic repertoires. We conclude the paper by suggesting a way forward. We believe that language modelling for assessment purposes needs to be more nuanced and much more closely related to the context and situation in which it is to be used. Hence rather than being all-encompassing and fit-for-all, assessment of EAP should be more locally focussed and more context and domain specific. This would suggest moving away from large-scale testing to appropriate assessment focussing on those features of communication that would be relevant and hence authentic in the target language situation.

Wednesday, June 22
15:05-15:35
Room 6
University of Palermo, Scuola Politecnica - Faculty of Engineering

Comparability of Students’ Writing Performance on TOEFL-iBT and Required University Writing Courses

Lorena Llosa, New York University
Margaret E. Malone, Center for Applied Linguistics
Anne Donovan, Center for Applied Linguistics
Christopher Van Booven, New York University
Scott Grapin, New York University

Although studies have been conducted to validate the writing tasks on various versions of the TOEFL, none have explored the comparability of students’ performance on TOEFL writing tasks and actual academic writing tasks. Investigating this comparability is essential for providing backing for the extrapolation inference in the TOEFL validity argument (Chapelle, Enright and Jamieson, 2011). Our study begins to fill this gap by comparing 100 international nonnative-English-speaking undergraduate students’ performance on two TOEFL iBT® writing tasks—one Independent and one Integrated—and their performance on two writing tasks assigned in required writing classes in U.S. universities. We examined comparability of student performance across TOEFL and classroom writing tasks in terms of scores on various dimensions
of writing quality as defined and measured by an analytic rubric. During a two-hour data collection session, students completed a background questionnaire, two TOEFL tasks, and a post-questionnaire. Students also submitted first and final drafts for two assignments completed for their required writing class during their first year of college. All writing samples—TOEFL Integrated and Independent, and the first and final drafts of the two class assignments—were scored by a team of five raters using an analytic rubric focusing on the following dimensions: grammatical, cohesive, rhetorical, sociopragmatic, and content control. Interrater reliability was high with only 8.5% of ratings requiring a third rating. Paired sample t-tests showed no statistically significant differences between scores on the TOEFL tasks and scores on the first drafts of the class assignments. Scores on the final drafts of the class assignments were significantly different and higher than scores on the first drafts and scores on the TOEFL tasks, likely because final drafts had been reviewed and revised at least once. Nonetheless, correlations between scores on TOEFL tasks and first and final drafts of class assignments were positive and significant, ranging from .59 to .71. When we examined the relationship between TOEFL scores and the various dimensions of writing quality, we found that the total scores on TOEFL tasks correlated highest with grammatical and cohesive control scores in the classroom assignments and correlated lowest with rhetorical control scores, which also makes sense given the variability in task characteristics of class assignments and TOEFL tasks. Another interesting finding is that the patterns of relationships were the same for the Independent and Integrated tasks. To further explore the extent to which TOEFL tasks measure the same constructs as class assignments, we will use structural equation modeling. Using a confirmatory factor analysis of multitrait-multimethod data, we will examine overall model fit and parameter estimates to investigate the convergence and discrimination of the scores as well as method effects. The evidence so far provides support for the extrapolation inference in the TOEFL validity argument: the high correlations between scores on TOEFL and class assignments indicate that they measure similar constructs and the comparison of scores across tasks indicate that student performance on TOEFL writing tasks most resembles how they perform on the first draft of a class assignment.

Wednesday, June 22
15:05:15:35
Room 7
University of Palermo, Scuola Politecnica - Faculty of Engineering

**Complex Systems Interacting in a Classroom-based Dynamic Assessment**

*Clarissa Lau, University of Toronto*
*Eunice Eunhee Jang, Ontario Institute for Studies in Education, University of Toronto*

Dynamic assessment posits the perspective that students’ learning capability must be understood with two critical pieces of information: how students’ performance is
assisted through a mediator and the extent of benefit a student can acquire from this assistance (Poehner & Lantolf, 2005). Mediation is an essential aspect in dynamic assessment (Feuerstein, Rand, & Rynders, 1988) and while dynamic assessment is frequently discussed in relation to classroom assessments (Deakin Crick, 2007; Lantolf & Poehner, 2011; Poehner & Lantolf, 2005); little research has examined the nature of changing student agency as suggested by complexity theory in the context of dynamic assessment. Complexity theory is a view that presents students as a dynamic, complex system that intersects and updates as change is introduced and this perspective enriches the student profile by accounting for the rapid changes and interaction between student and mediator (Larsen-Freeman, 1997). As such, this study aims to understand the developing student and mediator systems through the lens of complexity theory to better understand the dynamic nature of the classroom-based assessments. This study recruited a total of 14 Grade 6 students from an urban school in Ontario. Each student received one-to-one diagnostic assessment interventions with individualized feedback from a mediator. Interventions lasted over eight sessions and targeted specific reading goals as identified from diagnostic profiles generated from application of cognitive diagnostic modeling to provincial literacy assessment data. Three researchers with varying degrees of teaching experience mediated the students’ reading strategy use relating to their reading goals. Written transcriptions of each session were deductively analyzed with one coding scheme of student responses identifying cognitive, metacognitive, and affective factors and a second coding scheme of mediator responses identifying mediation focus, directionality, and intensity. Given the unique nature of each interaction, this study took a multiple case study approach to examine the dynamic interplay between student and mediator. Findings of this study highlighted the evolving profiles of mediator-student relationship. Results identified variance in impact of mediators regarding their mediation focus and intensity in response to students’ cognitive and metacognitive strategy use. First, mediators who focused on improving the complexity of students’ reading strategy use led to more effective cognitive and metacognitive strategy use. Second, mediators’ effective moderation of mediation intensity for different students identified an increase in cognitive and metacognitive strategy use. Lastly, results identified that when the interaction between mediator and student shifted the mediating direction from mediator-initiated to student-initiated, students were more likely to display positive use of cognitive and metacognitive strategies as well as positive affect. Implications of this study suggest that assessment should not overlook individual differences in both students and mediators (e.g. teachers in classroom). By aggregating classroom assessment results, important differences that may impact the assessment process are often overlooked. In a classroom setting, even though an assumption is that a teacher provides similar instruction to all students, the interaction between a teacher and individual student is dynamically changing; therefore, students’ non-linear growth and response to assessment should be examined as processes of change.
Plenary

**Diagnosing Development of Reading and Writing in a Foreign Language - a Longitudinal Perspective**

*Ari Huhta, University of Jyväskylä*
*Charles Alderson, Lancaster University*
*Lea Nieminen, University of Jyväskylä*
*Riikka Ullakonoja, University of Jyväskylä*
*Eeva-Leena Haapakangas, University of Jyväskylä*

The ability to read and write in another language is of growing importance in a globalised world and therefore diagnostic language testing is potentially important both for classroom and large scale assessments. Relatively little is known, however, about how foreign language (FL) abilities develop, how to identify strengths and weaknesses in them, and which abilities contribute most to the development of, e.g., overall FL reading and writing performance. The paper is based on 5-year research on the construct and development of reading and writing in English as a foreign language that sought to identify those features of language use and learning which predict a learner’s development or lack of it, and to better understand causes of strengths and weaknesses. The project brought together scholars from different theoretical orientations (e.g., language testing, L1 and L2 learning). Given how little is known about FL diagnosis, the project is best seen as an example of basic rather than applied research. Thus, its main contribution is to offer novel, well-grounded theoretical insights and to develop a range of methodologies to study FL development and its diagnosis. The research reported presents the findings of the longitudinal part of a larger project carried out with young learners of both second and foreign language learners. The study reported here focuses on learners of English as a LF and started in 2010 when the learners were in grade 4. In the first phase of the study the learners were administered a wide range of standardized cognitive and linguistic measures of both L1 and FL skills as well as motivational and background questionnaires. The learners were followed and re-tested with linguistic and motivational measures during the main project till 2013 when they were in grade 6. About 90 students could be located and recruited to the final stage of the study two years later in the autumn of 2015 when they took a range of English tests covering reading, writing, vocabulary and structures and replied to a background and motivational questionnaire. Some of the measures (e.g. motivation, part of vocabulary and reading) were exactly the same that the learners had taken in grade 6, while others were more challenging as appropriate to learners’ more advanced English skills. Data analyses included the investigation of learners’ writing performances with tools such as Coh-Metrix developed for detailed, automated analysis of English language texts, and regression analyses with grade 9 FL measures as dependent
variables and the measures in grades 4-6 as independent variables. The study sheds light on how well the cognitive, linguistic, motivational, and background variables measured up to five years earlier could predict later development in FL skills. Besides overall development in these skills, we also report whether the degree of prediction was different for different aspects of reading (e.g., understanding main ideas, making inferences) and writing (e.g., syntactic complexity, lexical diversity). Finally, we discuss how the development of FL skills related to learners’ motivation to study English and their use of the language outside school.

---

Thurday, June 23
10:00-10:30
Aula Magna
University of Palermo. Scuola Politecnica - Faculty of Engineering

Plenary

A Measurement Model for Integrated Language Assessment Tasks

Ikkyu Choi, ETS
Youngsoon So, Seoul National University

Language assessment tasks that require the integration of multiple skills have gained interest in the field of language testing (e.g., Cumming, 2013). A popular form of such integrated tasks elicits written or spoken responses, based on the comprehension of listening and/or reading sources. The potential benefits of integrated tasks include greater authenticity and positive washback (Feak & Dobson, 1996; Lewkowicz, 2002). However, the engagement of both receptive and productive language skills poses a challenge to standard measurement models that attribute test taker responses to a single latent trait. A measurement model designed to account for the interaction between receptive and productive skills has not been investigated in detail in the language testing field. In this study, we investigate the research gap in measurement models for integrated tasks. In particular, we propose a measurement model for integrated tasks and illustrate the model’s application and added-value in terms of the precision of skill level estimates. The proposed model is based on the Bayesian Network approach to building measurement models for educational assessment (Almond, Mislevy, Steinberg, Yan, & Williamson, 2015), and utilizes the inhibitor relationship in which the receptive skill places a threshold for the productive skill to be activated (Levy & Mislevy, 2004; Mislevy et al., 2002). By considering the interaction between receptive and productive skills in responding to integrated tasks, the proposed model estimates the level of the receptive skill required for the comprehension of source information, as well as the productive skill level indicated by observed scores. We illustrate the application of the proposed model using a TOEFL iBT data set. The Speaking section of TOEFL iBT has four integrated tasks that provide listening input for comprehension.
Therefore, we focused on the interaction between listening and speaking skills. We fit the proposed model to the Listening and Speaking item response data of 5,000 randomly selected test takers. The results showed that the threshold listening skill levels required for source comprehension ranged from 5 to 7 on the TOEFL iBT Listening section score scale. We also examined the added-value of the proposed model by comparing the precision of estimated listening skill levels from the proposed model to the corresponding estimates from a unidimensional model (for the Listening section only) and a correlated multidimensional model (one dimension for each of the Listening and Speaking sections). The point estimates from the proposed model were, overall, 10 and 15 percent more precise than the corresponding estimates from the unidimensional and correlated multidimensional models, respectively. The increase in precision was especially pronounced among test takers whose listening skill level was around or lower than the estimated comprehension threshold. We believe that these results indicate the potential of the proposed model as a measurement model for integrated tasks. Although we illustrated the application and the added-value of the proposed model using TOEFL iBT Listen-Speak integrated tasks, the model is applicable to a wide variety of integrated tasks, such as “read-and-write” tasks and mathematics tasks for English learners.

**Thursday, June 23**
10:30-11:00
Aula Magna
University of Palermo, Scuola Politecnica - Faculty of Engineering

**Plenary**

**Language Performance Rating Issues in Argument-Based Validity**

_Ute Knöch, University of Melbourne_
_Carol Chapelle, Iowa State University_

Diverse purposes of language tests prompt a variety of approaches to defining the constructs measured and the contexts for extrapolation of test scores. Accordingly, argument-based validation requires test developers/researchers to specify any claims that they hope to make about constructs and contexts on the basis of test results. Such claims need to be made explicit in order for validation research to address the warrants and assumptions that underlie construct and context-related claims about a specific interpretation and use. The clarity ultimately to be gained by argument-based validity, however, comes at a cost of effort spent learning a different way of thinking about validity of test interpretation and use: each tester needs to create a specific validity argument in part by drawing upon existing knowledge and methods to develop the relevant claims, inferences, warrants and assumptions. Our paper begins to address this challenge by showing the positioning in an overall validity argument of the multiple issues involved in rating language
performance. Previous presentations of argument-based validity (e.g., Chapelle, Enright & Jamieson, 2008) include performance rating issues primarily if not solely in the assumptions underlying the evaluation inference (also referred to as the scoring inference). This elaboration of the evaluation inference agrees with Kane’s (2013) description of it advancing the validity argument “from the observed performances to an observed score. It typically makes assumptions about the appropriateness of the scoring criteria and the rules for combining scores (which are based on judgement and/or statistical criteria)” (p. 10). Performance rating issues clearly fit with evaluation/scoring inferences in a validity argument, but in language testing, rating issues also include assumptions about the correspondence of the rating scale to the test construct and the context of extrapolation. Therefore when test developers and researchers attempt to develop an interpretation/use argument for a test, they are challenged by the fact that the assumptions they want to make about rating do not fit neatly within one inference. In view of the central role of performance rating in language test score interpretation and use in addition to the well-developed methods for investigating aspects of the rating process, our paper outlines the types of warrants underlying score interpretation that are affected by the rating process. We show that such warrants extend across multiple claims and inferences in a validity argument, including claims about constructs (explanation) and contexts (extrapolation). As a basis for organization, we have used inferences roughly aligned to those introduced by Kane (2006) and used by Chapelle, et al. (2008). For each inference, we suggest wording for warrants and their assumptions as well as illustrate the backing that would be used for support of each assumption. Our paper therefore opens discussion of how to elaborate on the validity argument conceptualizations that have begun to provide some common ground for fruitful discussion of slippery conceptual issues in language testing.

---

Parallel sessions

Thursday, June 23
11:30-12:00
Room 5
University of Palermo, Scuola Politecnica - Faculty of Engineering

Diagnosing EFL Learner’s Writing Difficulties in Chinese Secondary Schools

Cecilia Guanfang Zhao, Shanghai International Studies University

The evaluation of students' educational progress in Chinese schools still depends heavily and almost exclusively on the use of timed tests. The extensive use of such tests, however, leads to a general lack of writing ability among Chinese students. In EFL teaching and learning, this problem is especially salient (Yan, 2012) and is seen as “directly affecting [Chinese students’] research and learning ability at the postgraduate level” (Zhang, 2010, p.311). To help students develop their EFL writing
ability, however, we need to first understand the nature and the sources of their writing difficulties. Consequently, it is important to gather evidence and data that help us explain why Chinese students have the problems they have with EFL writing. As part of a larger project, the present study focuses on collecting and analyzing such data from students at the secondary level, especially the high school level. Considering that high school students are less able to verbalize and articulate their writing difficulties, a diagnostic tool that taps into the specific writing difficulties they face seems to be most appropriate. While commercially available diagnostic tools do exist, they are developed for use either with native speakers of English (SWET, TOWE, TOWL-3) or with L2 learners of English at the college or adult level, often also in an English-medium educational setting (DIALANG, DELNA, DIAKUL). None focuses on the EFL student population at the secondary level. As a response, this study developed a diagnostic tool tailored to this student population, considering their learning objectives and overall language proficiency, as well as the assessment practices in the Chinese context. Unlike the existing diagnostic tools that either focus on assessing discrete aspects of writing ability or rely on a detailed analytic rubric to diagnose strengths and weaknesses, this instrument examined both writing processes and written products. It first presented pre-writing questions that both tapped into students’ key cognitive ability in areas such as task interpretation, planning, and organizing. In the meantime, each pre-writing question also tapped into a specific area of language knowledge addressed in their curriculum, textbooks, and learning goals. Then, it solicited students’ actual writing products based on their responses to the pre-writing questions. Finally, a post-writing questionnaire was administered to collect students’ immediate reflections on and self-assessment of the writing performance and specific writing difficulties. This instrument was then administered to 190 10th and 11th grade EFL students at a local high school in a pilot study. Preliminary data analysis showed that the most prevalent and salient writing difficulties among these high-school students are the lack of ability to accurately interpret the writing task, effectively address audience’s needs, and clearly present their ideas. Results from this study can offer a more nuanced and context-specific interpretation of Chinese high school students’ EFL writing difficulties, which in turn could help classroom teachers more accurately determine students’ weaknesses in their writing practice, and hence offer timely and appropriate remedial services to students before they exit high school and enter college.
The Validity of Reader/Text Matching Algorithms

Kathleen Sheehan, ETS

Advances in natural language processing techniques combined with the increasing availability of large collections of electronic text have sparked a renewed interest in algorithms designed to match language learners with texts of appropriate difficulty, that is, texts that are expected to be challenging, yet not so challenging as to cause frustration or reduce motivation. In a symposium presented at LTRC-2015, Stenner and his colleagues argued that a reader/text matching algorithm (RTMA) implemented via the Lexile Framework was capable of explaining more than 90% of the variance in readers’ comprehension levels during encounters with text. Yet data that fail to support this assertion have also been presented (Sheehan, 2014). This study evaluates one possible explanation for these inconsistent findings: the parameter shrinkages that may be introduced into validity analyses when students’ observed scores are reexpressed on a common vertical scale (Mislevy, Sheehan & Wingersky, 1993). Analyses are focused on a set 76,538 reader/text matches generated via the Oasis system, an online reading platform powered by the Lexile RTMA. Each “match” involved one or more of 2,694 students who responded to one or more of 586,300 items after being “matched” to one or more of 372 different texts. Two text complexity estimates are generated for each text: one estimated entirely from students’ observed item responses, and one estimated from the same data after all scores have been reexpressed on the Lexile scale. If the parameter shrinkages predicted by Mislevy and his colleagues are present, then text complexity estimates generated from the rescaled data will be more closely aligned with the Lexile theory compared to those generated entirely from students’ observed responses. The observed rate of shrinkage is estimated as follows: Shrinkage = 100*(N2 - N1)/(372 - N1) where N1 is the number of texts classified as functioning in a manner that is consistent with the Lexile theory before rescaling, and N2 is the number of texts classified as functioning in a manner that is consistent with the Lexile theory after rescaling. Note that the observed rate of shrinkage ranges from 0% (when N2 = N1) to 100% (when N2 = 372). Analyses of the observed data yielded N1 = 141 and N2 = 243, so shrinkage equals 44%. This high degree of shrinkage suggests that current estimates of the validity of the Lexile RTMA may be significantly inflated. Implications with respect to the goal of generating valid evidence about the performance characteristics of proposed RTMAs will be discussed.


Examining the Writing Construct under Different Contexts using Automated Features

Chaitanya Ramineni, Research Consultant
Brent Bridgeman, ETS

Taking a corpus based approach, in this study, we use automated features to analyze two corpora of texts (written by the same group of students). The first corpus is a set of responses written on impromptu topics assigned on a high-stakes admissions test. The topics vary widely and the written texts were produced under timed conditions. The second corpus is a set of course-related writing samples written as part of graduate course work. The topics in this corpus vary widely as well and these written texts were produced over an extended time period. While all the test responses appear in essay format, the course-related writing samples appear primarily in the format of long research papers that make use of source texts and argumentative writing skills with some short technical papers that are more informational in nature. Hence, there is sufficient reason to anticipate that the writing construct is instantiated differently under each condition/context. However, one of the claims of the admissions test is that the writing performance on the timed test is a predictor of writing performance in graduate school, thereby suggesting that we expect some overlap between the two instantiations of the writing construct. To investigate the claim, we scored the course-related writing samples using a modified version of the rubric used for the timed test to assess primarily the writing quality (we excluded samples that were highly technical in nature) and computed correlations between the scores on the texts in the two corpora. We observed moderate correlations which are indicative of the relationship between student writing quality across the two contexts, but it is not a direct indicator of the overlap (and differences) in how the writing construct is defined under the two contexts. Hence, we next used a set of automated features to analyze these texts and gain a better understanding of which dimensions of the writing construct are in play in different contexts, in this case a timed testing context and an academic writing context. For example, we found that language use related features are more relevant in the untimed condition and predictive of writing quality, than features measuring conventions and discourse elements which are more predictive of writing quality in timed samples. The set of automated linguistic features we have used come from one current automated essay scoring engine and is limited. We therefore plan to extend the set by including additional automated features from other capabilities which should help us further differentiate the construct under the two conditions. These additional analyses will be complete in time for the conference presentation.
Thursday, June 23  
11:30-12:00  
Room 8  
University of Palermo, Scuola Politecnica - Faculty of Engineering

**An Investigation of Young EFL Students' Performance on a Timed Argumentative Writing Task**

*Mikyung Kim Wolf, ETS*  
*Saerhim Oh, Teachers College, Columbia University*  
*Yuan Wang, ETS*  
*Fred Tsutagawa, Teachers College, Columbia University*

Academic writing skills are increasingly important for young school-aged EFL students who are planning to study in English-medium instructional environments. For example, the academic standards for Grades K-12 in the United States explicitly call for greater academic writing skills so that students are ready for college and careers. Assessment tasks asking students to construct informational or argumentative writing have also become prevalent in large-scale standards-based assessments in K-12 schools. To support young EFL students’ academic writing development, it is essential to examine their writing performance and identify areas for improvement. In particular, this study aimed to investigate young EFL students’ performance on an argumentative writing task from a standardized English language proficiency assessment. This assessment, including listening, reading, speaking, and writing sections, was designed to assess young EFL students’ social and academic English language proficiency. Using a portion of this assessment dataset, we focused on examining the writing patterns of EFL students at different proficiency levels. Further, we investigated the common types of errors that were present in the students’ writing at the linguistic and organizational levels. We were also interested in finding any writing patterns of students from different countries. The data for this study were comprised of a total of 2,175 EFL students from Korea (n=1,551) and Brazil (n=624) within the age range of 10 to 15. The writing responses were rated on a holistic scale with the scores ranging from 0 to 4. In order to examine the linguistic characteristics of their writing, we utilized an automated tool to count various features at the lexical, syntactic, and discourse dimensions. The outcomes of the linguistic analysis were compared among the students at each score level as well as by each country. In order to conduct more in-depth analysis of argumentative-specific writing features (e.g., inclusion of claims in the introduction, the quality of supportive evidence), 235 argumentative writing samples were randomly selected, maintaining a balanced sample size for each score level from each country. A pair of trained researchers analyzed the student writing based on a coding scheme. The coding scheme focused on types of errors at the organizational, content, and language use levels based on previous research (e.g., Biber & Gray, 2013; Friginal, Li, & Weigle, 2014). The codings were then compared by student score level and by country. Descriptive statistics, multiple ANOVAs, and regression analyses were conducted with both the larger and smaller data sets. Preliminary results indicate that increased composition length, lexical diversity, use of academic words,
nominalization, and passive structure were strongly associated with higher writing scores. Some discernible error patterns also emerged, such as a higher frequency of run-on sentences for the Brazilian students. The results from the human rating analysis suggested that organizational ratings significantly correlated with higher scores. Some notable organizational errors included omitting a conclusion and using inadequate sources as supportive evidence. In this session, we will present detailed findings and discuss implications regarding how the large-scale assessment results could be utilized for instruction to improve young EFL students' academic writing.

Parallel sessions

Thursday, June 23
12:05-12:35
Room 5
University of Palermo, Scuola Politecnica - Faculty of Engineering

**Construct Validity of a High-stakes Writing Test: The Effects of two Types of Expository Essay Writing**

*Hyun Jung Kim, Hankuk University of Foreign Studies*
*Jooseung Lee, Hankuk University of Foreign Studies*

Due to the nature of L2 performance assessment in which examinees' performance is necessarily affected by construct-irrelevant factors (e.g., raters and tasks), previous studies have identified such factors and have proposed a framework of performance assessment for its construct validity (e.g., Fulcher, 2003; McNamara, 1996; Milanovic & Saville, 1996). L2 writing performance assessment has paid special attention to the task variable, among many other factors, because of its importance; examinees normally respond to only one or two tasks due to limited time and the cost of rating (Weigle, 2002). While the different aspects of tasks (e.g., discourse mode, task type, and topic) have been examined, little is known about the effects of genre, or of different types of writing within a genre. In an attempt to address this issue, the present study investigated (1) whether two types of expository essay writing (i.e., compare/contrast writing and problem/solution writing) used for a high-stakes admissions test had different effects on examinee test performance, and (2) to what extent examinees' linguistic features differed between the two types of writing. One hundred forty-three applicants to an M.A. TESOL program at a university in Seoul took an English writing test across five administrations as part of the admissions process. The examinees were presented with the two types of writing prompts (compare/contrast and problem/solution), with different TESOL-related topics for each administration. They were asked to select one of the two types and write a 300-word essay for 30 minutes. The examinee responses were scored by two native English-speaking professors using an analytic scoring rubric. The analytic ratings were analyzed using many-facet Rasch measurement models to compare the
effects of examinee ability, raters, and analytic rating scales between the two writing groups. The examinees' written responses were further analyzed using Coh-Metrix to compare the use of linguistic features of the two writing groups. Moreover, the two raters were interviewed to examine how they perceived the examinee performance in the two types. Overall, the analysis of the examinees' test performance provided evidence of construct validity of the writing admissions test. The two types of writing were found to share many common features from the statistical analysis of examinee, rater, and rating scale facets. The only difference between the two types of writing was found in the relative difficulty of the organization of an essay in comparison with other aspects of writing (content elaboration and grammatical control). It was more difficult to receive a high score in organization in the compare/contrast writing group, while getting a high score in organization was not necessarily more difficult in the other group. Further, a Coh-Metrix analysis of the linguistic features also revealed differences in the cohesive-related linguistic features of the two writing groups. The interview data analysis suggested that the differences regarding the difficulty in organizing an essay were partly attributed to rater behavior and the writing prompts. The findings of the study provide implications for the development/use of different writing types and rater training in the context of a high-stakes admissions process.

Thursday, June 23
12:05-12:35
Room 6
University of Palermo, Scuola Politecnica - Faculty of Engineering

Testing Bilingualism: Incorporating Translanguaging into a Listening Task for University Professors

Beverly Baker, University of Ottawa
Amelia Hope, University of Ottawa

This paper reports on the development of a translanguaged French and English listening test task for professors in a bilingual Canadian university. This university prioritizes French-English bilingualism in its faculty members; indeed, attainment of either “passive” or “active” bilingualism, as determined upon hiring and as measured by tests, is required for tenure. In classes, meetings, and public spaces on campus, both official languages are mixed constantly. However, all tests of “bilingualism” contain a single language only. Indeed, recourse to the other official language on a test is normally considered as evidence of lower proficiency. However, research on translanguaging—a dynamic process where resources from more than one language are employed in meaning making (Garcia, 2009)—suggests that fluid translanguaging could be considered a sign of advanced bilingual ability. Translanguaging is increasingly valued and encouraged in pedagogical contexts (Lewis, Jones, & Baker, 2012), but beliefs about the compartmentalization of languages for assessment are still strongly entrenched (Shohamy, 2011). In order to capture these developments in the understanding of the construct of bilingual competence, as well as to more
authentically represent the target language use domain, a pilot translanguage task was added to the listening test for professors. The task prompt is a departmental meeting, a place of frequent alternation between the two languages in this institution. Passive bilingual ability for this task was operationalized as the ability to follow meetings that alternate between French and English, with production in the L2 not being necessarily required. Task development began by convening and recording simulated meetings with faculty volunteers, to capture typical meeting language as well as realistic examples of translanguage practice. Transcriptions of these meetings guided the creation of draft listening tasks. For professors taking a French examination (i.e., those whose first or strongest language was English), the prompt was a recorded meeting conducted primarily in French with some insertions in English. Professors then read short answer comprehension questions in French and responded in the language(s) of their choice. The draft tasks were piloted with 20 professors, who also took the existing monolingual listening tasks in order to establish the relative difficulty of the new tasks. Task responses were also analyzed qualitatively, to examine the use of translanguage by the professors during the assessment. Because of the innovation of this task, we suspected that there may be resistance to the use of the other language. Therefore, we also conducted a survey of test takers and other test stakeholders, with specific questions related to their perceptions of construct coverage as well as the suitability of using translanguage. Our presentation will briefly describe the development process, as well as results of the pilot phase and implications for future task revision. This study aligns closely with the conference theme, in that we have attempted to expand the conceptualization of the problematic construct of “passive bilingualism” with a task that makes explicit the value of dynamic bilingual practices.

Thursday, June 23
12:05-12:35
Room 7
University of Palermo, Scuola Politecnica - Faculty of Engineering

Language Assessment Raters in the Call Centre Industry: Interfaces between Language and Subject Matter Experts

Jane Lockwood, City University of Hong Kong
Michelle Raquel, Lingnan University

Increasingly workplaces and professional bodies want a voice in the development of performance assessments measuring communicative competence at work and such assessments are bound to be high stakes for candidates determining employment, certification, appraisal and promotion. Exploring the implications of this shift, some recent studies have investigated how language assessment can better account for subject matter expert judgments about what comprises good and poor communication exchange in their fields and how these judgments, translated into ‘indigenous criteria’ can be captured in a language assessment rubric. This study
investigates call centre how well SMEs, with no language teaching or language assessment background, use a language informed assessment tool for quality communications management. Specifically we explore how reliably the SMEs rate English spoken communication skills of customer services representatives (CSRs). Fifteen Indian non-native speaker quality assurance (QA) specialists (the SMEs) employed by an Indian-based call centre of a large multinational bank (Bank A) were trained to use a new linguistically grounded QA tool called the ‘Business Performance Language Assessment System Call Assessment’ (BUPLAS CA) for quality control purposes. BUPLAS CA is specifically designed to measure call centre communication effectiveness by assessing authentic calls. QA specialists received rater training from language experts to understand how to rate using a performance rubric. After the training, they were asked to score 19 calls and also participate in interviews about the rating process and procedures. Quantitative analysis of their ratings showed that QA specialists had high agreement of BUPLAS CA scores and that they were able to distinguish five domains of the scale and separate the levels. Interviews conducted after the rating exercise revealed a very positive rating experience using BUPLAS CA as well as some of the difficulties they had in interpreting some domains. This study therefore builds on other studies investigating the efficacy and procedures involving SMEs in not only the construction, but also the implementation, of workplace communications assessment.

Thursday, June 23
12:05-12:35
Room 8
University of Palermo, Scuola Politecnica - Faculty of Engineering

Young EFL Learners’ Strategy Use during L2 Speaking Task Performance
Ching-Ni Hsieh, ETS
Lin Gu, ETS

This study examines the relationship of age (as a proxy for cognitive maturity) to strategy use while young language learners are engaged in L2 speaking tasks. Cognitive maturity is operationalized as age at testing, following Miralpeix (2007). In line with the conference theme that focuses on the influence of content and topical knowledge on language performance, we hypothesize that young learners of different ages would have differential cognitive abilities and processing efficiency and these differences would impact how young children execute cognitive strategies during L2 speaking task performance (Schwenck, Bjorklund, & Schneider, 2009). The impact of cognitive maturity on task performance unavoidably leads to the reflection on the cognitive demands of tasks imposing on young learners and underscores the importance of taking into account the cognitive differences when designing language assessment tasks for young language learners (Muñoz, 2008). The study investigated (through retrospective cognitive interviews) the cognitive strategies used by young EFL learners while they were engaged in four computer-delivered speaking tasks.
The four tasks, with increasing task complexity, included a four-picture narration, a story retelling based on a short animation, a six-picture narration, and an integrated listen-and-speak task. Thirty-one school children, ages between eight and 11 (M = 10.3), participated in the study. The students were from a primary school in Hong Kong where English is used as the medium of instruction. The students were elementary to upper intermediate English speakers as measured by a large-scale young learner language assessment, TOEFL Primary. The cognitive interviews were conducted one-on-one between a researcher and the participating students at a quiet school conference room. After responding to a task, the student answered a set of cognitive interview questions related to the task. Each interview session lasted between 30 and 35 minutes. A scheme was created on the basis of the strategies emerging from the interview transcripts and informed by previous research on children's L2 development and cognitive strategy use in order to code the interview data (Harnishfeger & Bjorklund, 1990). Results revealed that the participating students employed four major cognitive strategies: organization, elaboration, retrieval, and pre-task planning. All students, even with low to intermediate level of English proficiency, were able to reflect on their strategy use and express how the use of different strategies had helped them process information and organize their thoughts when responding to the four speaking tasks. The data corroborate previous findings (e.g., Coyle & Bjorklund, 1997; Harnishfeger & Bjorklund, 1990), suggesting that older children are more capable of employing a wider range of strategies while completing cognitively demanding performance tasks. The analysis also showed that cognitive strategy use is influenced by the individual student's language proficiency, learning styles, and prior language learning experiences. The study results extend our knowledge of how age impacts young EFL learners' speaking task performance. Implications for task design include the consideration of different strategies that may be called upon by young language learners while responding to L2 oral proficiency tasks.
inevitable narrowing of the test construct, since computer-based speaking tests are constrained by the available technology and include tasks which lack interactional components, a distinct feature of the face-to-face delivery mode. However, face-to-face speaking test administration is usually logistically complex, resource-intensive and can be difficult to conduct in geographically remote or politically sensitive areas. Recent advances in video-conferencing technology now make it possible to engage in online face-to-face interaction more successfully than was previously the case, reducing dependency upon physical proximity. This paper reports on Phase 2 of a major study into the impact on test construct of the mode of delivery of an existing speaking test, online and face-to-face. Phase 1 had presented results and recommendations of a small-scale initial investigation designed to explore how new technologies could be harnessed to deliver the face-to-face version of a standardized speaking test. It investigated what similarities and differences, in scores, linguistic output and test-taker and examiner behavior, could be discerned between the two formats. The results of the analyses suggested that the speaking construct remains essentially the same across both delivery modes. Phase 2 of the project was a larger-scale follow-up investigation designed to: a) analyse test scores using Many-Faceted Rasch Model (MFRM) analysis b) investigate the effectiveness of examiner training for the video-conferencing delivered test (developed following Phase 1 research findings) c) confirm Phase 1 results and gain further insights into the two delivery modes. 99 test-takers each took two speaking tests under face-to-face and computer-delivered conditions. Performances were rated by 10 trained examiners. A convergent parallel mixed-methods design was used to allow for collection of an in-depth, comprehensive set of findings derived from multiple sources. The data collected included an analysis of feedback interviews with test-takers as well as their linguistic output during the tests (especially types of language functions) and score ratings awarded under the two conditions. Examiners responded to two feedback questionnaires and participated in focus group discussions relating to their behavior as interlocutors and raters and to the effectiveness of the examiner training. Eight observers also took field notes from the test sessions. The MFRM analysis of test scores indicated that although the video-conferencing mode was slightly more difficult than the face-to-face mode, when the results of all analytic categories were combined, the actual score difference was negligibly small. While 71.7% of test-takers preferred the face-to-face test, 39.4% reported that there was no difference in the difficulty of the two modes. All examiners found the training useful and effective, the majority of them reporting that under the video-conferencing condition test-takers had more opportunity to demonstrate their level of English proficiency and that it was easier for them to rate test-taker performance. The presentation will conclude with a discussion of the comparability of the construct(s) measured by the two delivery modes, which may have important implications for the future assessment of oral language proficiency.
A Score Augmentation Approach to the Investigation of the Added Value of Subscores in Reading and Listening Tests

Spiros Papageorgiou, ETS
Ikkyu Choi, ETS

Developers of language assessments typically report section scores, corresponding to separate language skills (listening, reading, speaking and writing). In order to provide more fine-tuned information about test performance, assessment providers might want to also report “subscores” for groups of items in the different parts of a test section. For example, subscores for reading and listening in a second/foreign language (ESL/EFL) can be reported in relation to various subskills, such as the ability to understand the main idea, understand specific details, make inferences, and so on (Alderson, 2000; Buck, 2001). Although teachers, students and test users might find such subscores appealing because they offer more information about test performance, reporting them needs to be defensible from a psychometric perspective. Reporting a subscore is justified if it can provide additional useful information about language ability that the (reading or listening) section score cannot provide. In other words, a subscore has “added value” (Sinharay, Haberman & Puhan, 2007) if it is a better predictor of the intended construct (e.g., ability to understand the main idea in a passage) than the section (reading) score is.

Past research in the field of language testing has focused primarily on the reporting of scores corresponding to each of the four language skills and whether they provide added value over the total test score (Sawaki & Sinharay, 2013). Our study aims to contribute to the literature by exploring the added value of individual subscores for groups of items within a test section assessing a specific language skill (reading and listening). Our study also demonstrates the application of score augmentation techniques (Haberman, 2008, Wainer et al., 2001), in order to “strengthen” the reliability of subscores within a reading or listening test section.

We addressed the following research questions in our study:
• Do subscores for groups of items in reading and listening test sections offer added value over the reading and listening section scores respectively?
• Does score augmentation improve the added value of these subscores enough to justify reporting them?

We analyzed the responses of 116,489 test takers to reading and listening items from operational administrations of two large-scale EFL tests. These tests varied in terms...
of their design and psychometric characteristics so that we could explore the added value of subscores and the application of score augmentation techniques in more detail.

Our analysis indicated that, in general, there was lack of support for reporting subscores, at least from a psychometric perspective, and score augmentation only marginally improved the added value of the subscores. We discuss several implications of our findings for test developers wishing to report more fine-grained information about test performance. We conclude by arguing that research on how to best report such refined feedback should remain the focus of future efforts related to English-language tests.

Thursday, June 23
16:30-17:00
Room 6
University of Palermo, Scuola Politecnica - Faculty of Engineering

**The Effects of Test Task and Pairing Types on Heritage and Non-heritage Korean Learners' Speaking Performance**

Sun-Young Shin, Indiana University

Pair or group oral activities have been increasingly accepted as popular L2 speaking assessments since they allow students to interact as active participants in assessment, thus enhancing test authenticity and learner autonomy (Katz, 2013; Ockey, Koyam, & Setoguchi, 2013). However, due to the nature of co-constructed interaction in a paired speaking test context (May, 2009), students are likely to be awarded scores differently depending on who they are paired or grouped with (Taylor & Wigglesworth, 2009). From an assessment perspective, this makes it difficult for test users to interpret and use test scores obtained from a paired or group speaking assessment because raters usually assign individual oral ability scores to each participant (Bonk & Ockey, 2003; Ockey, 2009; Van Moere, 2006). The degree of separability of each student’s contribution and performance in a paired speaking test is, thus, a matter of great concern to many language teachers and testers particularly when they have a mix of students with different language learning abilities and linguistic profiles as in many postsecondary foreign language programs in the U.S. (Brinton, Kagan, & Bauckus, 2008; Carreira & Kagan, 2011). Such differences between two groups pose pedagogical challenges for many foreign language teachers when they have both heritage language (HL) and non-heritage language (NHL) learners in the same classroom because they tend to have different linguistic profiles and instructional needs (Montrul, 2010; Montrul & Perpiñán, 2011). In particular, many HL learners have better pronunciation and higher fluency than NHL learners (Kagan & Friedman, 2003) despite their relatively weak literacy skills in the target language. These notable differences in spoken language abilities between two groups of learners (Kondo-Brown, 2003; Montrul, 2010; Sohn & Shin, 2007) could be a concern for foreign language teachers and testers especially when they
try to design and implement pair- or group-oriented classroom assessment. Language proficiency differences among interlocutors in a pair or group work might unfairly affect a learner’s speaking performance leading to biased scores (Davis, 2009). The present study thus undertakes the issue of whether university-level heritage language (HL) and non-heritage language (NHL) learners (N=78) of Korean perform differently as reflected in their overall scores and subscores (Pronunciation, Grammar, Vocabulary, and Fluency) on monologic and dialogic speaking tasks when they are paired with partners of the same or different language backgrounds in Korean using a univariate and a multivariate G-study. The results of this study show that HL students performed similarly across different test tasks regardless of heritage backgrounds of their partners. Likewise, NHL students did not perform differently between two test tasks when they were paired with the NHL students. On the contrary, when NHL students were paired with HL students whose level of Korean oral proficiency was higher, they tended to gain higher scores in the paired speaking test task than in the monologic one, particularly in their fluency subscores. This study will provide important insights into our understanding of the relationships between test constructs and contextual features in paired and group oral assessment.

Thursday, June 23
16:30-17:00
Room 7
University of Palermo, Scuola Politecnica - Faculty of Engineering

An Eye-Tracking Study of Task Types and Attention to Nonverbal Cues in Video Listening Tests

Aaron Olaf Batty, Keio University

Multiple-choice listening items can be broadly categorized into two general task types: explicit and implicit. Explicit items require examinees to comprehend discrete linguistic features in order to answer correctly, whereas implicit items require a degree of inference. Much research has been conducted into these task types on audio-mediated listening tests, but video-mediated tests may present a special case in this regard. Research into nonverbal communication has shown that listeners tend to use the visual channel as support for the verbal channel, contributing emotional or social context in order to aid interpretation. As such, examinee attention to these nonverbal cues may differ by task type on listening tests. Although examinee attention to visual cues on video listening tests is a frequent topic of research, no studies have sought to objectively and quantitatively track attention to specific nonverbal cues or to investigate its interaction with task type. The present research addresses this gap via a mixed-methods study employing eye-tracking methodology for quantitative measurement of visual attention and stimulated recall for qualitative interpretations of the behaviors observed. Twelve Japanese university students sat a video-mediated English listening test of six multiple-choice items while wearing a Pupil DEV eye-tracking headset. Item stems were presented before the videos.
to provide a purpose for listening, and answer options presented after. All videos featured one man and one woman in conversation, and ranged from 38 to 99 seconds in length. After each test, the researcher conducted a stimulated recall interview of the participant while watching scanpath-overlaid video of his/her viewing behavior. A list of twelve visual cues of interest was developed during analysis of the eye-tracker output, and five main reasons for behavior were identified in interviews. Although most visual attention was focused on faces, regardless of task type, the facial cues attended to in explicit and implicit items were found to differ. Scans of the speaker’s face, wherein the participant’s eyes moved rapidly around the face rather than settling on one feature, accounted for an average of 15.10% of the viewing behavior in explicit tasks, as opposed to 24.12% in implicit tasks. Furthermore, scans of the listener’s face increased from 1.50% in explicit tasks to 3.97% in implicit tasks. Finally, quick alternating between the faces of the speaker and the listener accounted for an average of only 7.12% of the viewing behavior in explicit tasks, as opposed to 13.52% in implicit tasks. Paired t-tests on these differences were significant at below the .01 level, with large effect sizes. Respondents explained these behaviors as attempts to glean more information about the characters’ motives and feelings through facial expressions, which they believed would aid in answering the implicit questions posed. The findings demonstrate that listening test items relying on inference, especially that of a social nature, are likely to be more affected by the inclusion of video than items relying strictly on linguistic comprehension. The question of whether this difference in attention results in a difference in score, and whether such a difference— if identified—would be construct-relevant, remains open.

Thursday, June 23
16:30-17:00
Room 8
University of Palermo, Scuola Politecnica - Faculty of Engineering

Towards Developing English Ability Indicators for College Students in Taiwan – Constructing tests of English for specific academic purposes

Jessica Wu, Language Training and Testing Center
Rachel Y. F. Wu, Language Training and Testing Center

The focus of English curricula in Taiwanese universities has shifted from developing students’ general English proficiency to advancing their competencies in professional contexts to meet both academic and career requirements. The trend is evident from the growing number of English-medium instruction (EMI) courses and English for specific academic purposes (ESAP) programs offered in higher education. However, during the policy-making and program design stages, little attention has been given to the issues of defining appropriate professional English ability indicators for evaluating the students’ learning outcomes. To step forward towards the goal of identifying English ability indicators, this study explored the direction of developing assessment tools in three broad academic disciplines, including science...
and engineering (SE), health and medicine (HM), and business and management (BM). Unlike major academic-oriented language proficiency tests, e.g., TOEFL and IELTS, which adopt a one-size-fits-all approach by including a wide range of topics in the test questions in order to accommodate students of all majors, an ESAP test defined in this study is one that contains tasks or items which are topically related to a specific academic domain. Therefore, this study was particularly interested in exploring whether and how topical knowledge or discipline domain specific (DDS) background knowledge impacts students’ test performance (Bachman & Palmer, 1996; Alderson, 2000) and examining the extent to which language proficiency interacts with DDS background knowledge (Clapham, 1996). Data were collected in January 2015 from a total of 1,704 students who were taking ESAP courses at a university in southern Taiwan. The participants consisted of 756 students in SE courses, 306 in HM, and 642 in BM. All of the students took both an EGP (English for general purposes) test and an ESAP test corresponding to their ESAP course and filled out a post-test questionnaire about their test-taking experience. The order of the use of the EGP test and the ESAP test assigned to each student was counter-balanced. Overall, no significant difference was found between students’ performance on the ESAP and EGP tests. However, an interaction between proficiency level and test performance was observed: across disciplines, students whose language proficiency was below CEFR B2 performed better on the ESAP tests than the EGP test. That is to say, the ESAP tests may help elicit lower proficiency students’ best performance. In addition, survey results showed that students considered the ESAP test more relevant to what they learned in ESAP courses. The results of the study provide justification for adopting ESAP tests to assess students’ English language competencies in higher education. The findings also provide insights into the feasibility of developing future ESAP tests for other academic areas. Implications for developing English ability indicators for Taiwanese college students are also discussed in this paper.

References
Nurses’ Perspectives of the Qualities of Referral Letters: towards Profession-Oriented Assessment Criteria

Sharon Yahalom, The University of Melbourne

Good-quality written communication among overseas-trained health professionals is critical for accurate diagnosis, safe patient care and effective delivery of multi-disciplinary interventions (Vermeir et al., 2015). In Australia, the written communication of overseas-trained health professionals, including nurses, is assessed through use of the Occupational English Test (OET), a specific-purpose language (LSP) test. Nursing candidates are required to write a letter, usually one of referral, to another health professional (OET, 2014). In the field of LSP testing, test designers and researchers are becoming increasingly aware of the need to develop criteria that more accurately reflect work-related performance (Härmälä, 2010). Assessment criteria are more likely to be professionally relevant if specialists are involved in the decision-making process and their perspectives on what is required of candidates are considered (Pill, 2013). However, the current OET writing test assessment criteria were developed by language professionals without direct input from health professionals. The aim of this study is to identify what constitutes an effective referral letter from the perspectives of nurses. The findings from this qualitative study will be subsequently used as a basis for reviewing the assessment criteria of the OET writing test. Data include thirty interviews and six workshops conducted with nurses from two Australian hospitals – one rural and the other metropolitan. For the interviews, participants were asked about their perspectives on the qualities of effective referral letters and the advice they would give to newly registered health professionals to ensure that they write referral letters effectively. Thirty-four nurses took part in six workshops in which they commented on referral letters extracted from real hospital medical records from the same two hospitals. The aim of the workshops was to establish the valued features of referral letters. The findings show that referral letters are formally written and generally follow a prescribed structure; however, the appropriateness and comprehensiveness of the letters vary. Nurses commented that referral letters must be legible, clear, concise and accurate; although these qualities are not always evident in the letters. The ability to effectively communicate about patients through the inclusion of relevant information in referral letters is considered to be more important than linguistic features such as grammar and spelling. Lack of standardisation of referral letters means that vital information necessary for high-quality patient care is sometimes omitted. The findings of the study also show that the current assessment criteria for the OET writing test do not accurately reflect nurses’ criteria for effective referral letters, which has implications for the validity of the test. The findings contribute...
to the emerging field of research (e.g. Elder et al., 2012) which seeks to include specialists’ perspectives about what is considered important in the review and development of assessment criteria for LSP tests.

Friday, June 24
11:00–11:30
Room 7
University of Palermo, Scuola Politecnica - Faculty of Engineering

**Advancing Written Assessment Feedback Praxis through a Teacher-Researcher Collaboration in a University-level Spanish program**

*Kathryn Hill, La Trobe University & University of Melbourne
Ana Marie Ducasse, RMIT University*

A number of authors have highlighted the need for the research community to take greater account of teachers’ language assessment literacy (LAL), e.g. professional experience, perspectives and knowledge (e.g., Freeman & Johnson, 1998; Leung, 2005, 2014) in context with regard to assessment in the service of learning (Carless 2011; Mcmillan 2010). This paper reports on a collaborative dialogue (Scarino, 2016) between a teacher and researcher regarding written feedback practices in a Spanish as a foreign language program at an Australian university. Starting with Hill’s (forthcoming) framework, designed as a heuristic for promoting teachers’ assessment literacy, the study sought to explore the teacher’s ‘tacit’ expertise with the aim of uncovering how the researcher’s theoretical knowledge and the teacher’s practice might inform each other. Following Hill’s (forthcoming) framework, the study started with the following questions: 1. What does the teacher do (types of feedback)? 2. What does she look for (features of writing, quality & standard)? 3. What theories, understandings or beliefs inform her feedback practices? 4. What are learners’ understandings of feedback? 5. How does the teaching context (including task & learner characteristics) influence the nature and content of feedback? Participants comprised a language assessment researcher, an ‘expert’ Spanish as a foreign language lecturer and 15 students from beginner (CEFR A1), intermediate (CEFR B1), and advanced (CEFR C) levels in a university-level Spanish program. Data comprised written feedback on writing tasks for each of the three levels collected over a 12 week semester, unstructured think-aloud protocols (Ericsson & Simon, 1993) produced by the teacher during the act of providing written feedback (e.g., Leung, 2014), semi-structured interviews with the teacher and students respectively and reflective journals (teacher and researcher). Verbal report and interview data were recorded and transcribed. Data were analysed using thematic content analysis. Following Turner and Purpura (2016), relevant contextual variables will include student attributes (specifically CEFR level and prior language learning experience) in addition to task and topic. Discussion will focus on the ways in which the collaboration served to advance both theory and practice in teacher’s LAL as well as highlight issues warranting further research.
Exploring Contextualized Tasks Through L2 Students' Oral Performance: How Discriminated, Constrained and Generalizable are the Observed Task Performance Features?

Naoki Ikeda, University of Melbourne

With L2 speaking assessment being widely used for English proficiency tests for university admission, several discourse studies have empirically revealed L2 speakers' distinguishing features in tasks of TOEFL iBT (Iwashita et al., 2009) and boundaries of IELTS assessment levels (Brown, 2002). Performances on the tasks employed for the examinees in these studies have provided a refined picture of L2 learners' discourse that characterizes general oral proficiency for test users and test developers. The present study investigates performance in more contextualized oral tasks requiring examinees to utilize contextual elements and their linguistic resources to achieve communicative goals. It ultimately explores capabilities of contextualized tasks, which are desirable but remain vulnerable in generalizability (Xi, 2015). The study used oral data from L2 students (n=67) in English-mediated education including university students with IELTS 6.0/TOEFL iBT 60 to near-native level, and language school students with IELTS 6.0-6.5 preparing for university admission. They completed three dialogue and three monologue oral tasks simulating activities that involve individuals in using language to achieve particular goals in particular situations (Bachman & Palmer, 1996) in real life at university. Communicative goals, medium of communication, and intended imaginary listener were specified to characterize the language use context (Xi, 2015) that account for features of spoken product in quantity and quality. Further, imaginary listeners' social background (Brown & Levinson, 1987) and pragmatic demands (Grabowski, 2009; Youn, 2013) were operationalized. Task interlocutors (for dialogue tasks) were trained to interact in ways that require examinees to interpret the interlocutor's implied meaning (Purpura & Grabowski, 2012). The task performance was analyzed in features characterizing oral proficiency (e.g., fluency) in relation to the literature, pragmatic device (Bardovi-Harlig, 2013) and interactional management (Roever, 2011). The performances were examined by discourse-based analyses quantitatively and qualitatively and also were converted into scores by trained raters using rating criteria for Rasch analyses. Results combined with participants' interview data suggest the following: (1) task performance discriminated a group of university students (IELTS 7.0 level or higher) from other examinees significantly/substantially in many aspects as reported in the literature, (2) the task variables did not affect performances of highly proficient students (IELTS 8.0 level or higher), who consistently showed positive evidence in all of the observed features across the tasks, whereas other examinees' performances showed inconsistencies.
depending on task situation or/and modality and made the proficiency boundaries unclear, (3) interactional behaviors of some language students in dialogue tasks confused interlocutors in turn-taking although they were not significantly different from university students in many aspects at the same proficiency level, (4) relative difficulty of the tasks identified by Rasch analyses can reflect imaginary listeners’ social background in each task, but possible method effects (Iwashita et al., 2001) were also implied in some examinees’ interview comments. By examining L2 students’ performance in contextualized conditions combined with the participants’ views, this study will discuss potential utilization and methodological considerations of contextualized dialogue and monologue tasks, as an assessment instrument to elicit L2 learners’ distinctive features of spoken production.

Friday, June 24
11:35-12:05
Room 6
University of Palermo, Scuola Politecnica - Faculty of Engineering

The Challenge of Capturing the Construct of Written Communication for a Health-Specific Language Test

Ute Knoch, University of Melbourne
Tim McNamara, University of Melbourne
Catherine Elder, University of Melbourne
Robyn Woodward-Kron, University of Melbourne
Elizabeth Manias, Deakin University
Eleanor Flynn, University of Melbourne
Annemiek Huisman, University of Melbourne

Central to construct definition in the assessment of languages for specific purposes are the “indigenous criteria” or values that underlie domain experts’ judgments of communication in the relevant occupational setting. It is now widely accepted that such values should inform judgments of performance on a specific purpose language test if the resultant test scores are to be considered valid indicators of the capacity to communicate in the target language use domain (Jacoby and McNamara 1999; Douglas 2000). Identifying and interpreting what matters to domain experts is not straightforward however. Methodological and theoretical challenges relating to the manner and context in which data are elicited may influence informants’ responses and hence the validity of research findings. While these challenges have been explored in relation to spoken interaction (e.g., Elder & McNamara, in press), scant attention has been paid to written communication. A project involving collaboration between language testers, applied linguists and health professionals in medicine provided an opportunity to explore these challenges in relation to writing assessment. The project focused on inter-professional written communication in health settings, which forms the basis of a writing task on the Occupational English Test (OET), an English for specific purposes (ESP) test for assessing the language
proficiency of health professionals applying to practise their profession in Australia, New Zealand and Singapore. The project aimed to identify criteria indigenous to the work setting, which would inform the criteria used by language professionals in judging written performance on the OET. Two methodological approaches were adopted: 1) 25 doctors were interviewed individually by researchers to identify what factors they saw as facilitating or obstructing effective written communication in the healthcare context and 2) 5 workshops were conducted in which groups of 6-8 doctors (clinicians and educators) were asked to judge the adequacy of examples of written communication (such as discharge summaries and referral letters) in the health setting, and to verbalize the reasons for their judgements. Both sets of data were transcribed and an inductive thematic analysis of the transcripts was undertaken (Galaczi, 2014). The paper describes the challenges associated with each methodological approach (including issues of participant representativeness, selection of stimulus tasks and potential observer effects) and considers the validity of insights emerging from the interview and workshop data respectively. It concludes by discussing the potential impact of using one or other elicitation technique on the way the construct of written communication in healthcare is defined and operationalized for OET testing purposes, and reflects on the implications for construct definition in specific-purpose language testing more generally.

References

Friday, June 24
11:35–12:05
Room 7
University of Palermo, Scuola Politecnica - Faculty of Engineering

A Meaning-Based, CEFR-Linked Framework for Assessing Vocabulary Knowledge

Veronica Benigno, Pearson
John H.A.L. de Jong, Pearson

An overview of studies on vocabulary acquisition, teaching and assessment (Bogaards and Laufer, 2004; Meara, 2009; Nation, 2001; Read, 2000; Schmitt, 2000) shows that this field of investigation has evolved quite rapidly over the last few decades, yet there is still little agreement on which and how many words are needed to communicate efficiently at increasing proficiency levels. Two crucial questions with regard to assessment of vocabulary remain partially unanswered: how to assess vocabulary holistically and how to relate vocabulary knowledge to proficiency. Research has shown that vocabulary knowledge is related to both
the size of vocabulary and its depth. Assessment of vocabulary mostly relies on quantitative measures (such as frequency and diversity) whereas scarce attention has been paid to functional aspects such as communicative usefulness. Moreover, some concerns have been raised about using lemmas or word-families as the unit of counting in measuring individuals’ vocabulary size, as well as about focusing on single words, with no attention paid to lexico-grammar (Purpura, 1997; Wray, 2002) and pragmatics. Qualitative aspects of vocabulary knowledge are generally assessed using human rater judgements but no framework seems to have been put into place for a more systematic investigation of features such as vocabulary range or appropriateness to the context. Attempts to relate vocabulary knowledge to proficiency levels have so far been irresolute and mostly related on learner data, therefore showing what learners do know instead of what learners should know to communicate efficiently in an L1 context. In this paper we report on a corpus-based study to develop a CEFR-aligned graded lexical inventory of words and phrases in general English, with the purpose of providing the lexical exponents for English and complementing the functional guidance found in the CEFR descriptors. We followed four methodological steps: corpus analysis, semantic annotation, human ratings, and vocabulary scaling. The first step required corpus analysis and computational processing of the data. We compiled an L1 corpus of British and American English of about 2.5 billion words to inventory real-life usage of English in written and spoken text. In a second phase, words and phrases extracted from the corpus using measures such as frequency and dispersion were semantically annotated using the Council of Europe Vantage Specifications’ distinction in topics, general notions, and functions. Thereby the vocabulary was identified needed to express concrete and abstract concepts and to fulfil specific communicative purposes, e.g., to apologize, to make a request. A third step was to ask 15 teachers to rate vocabulary on a pre-defined 5-point scale built on the principle of communicative usefulness. In the final step frequency data and teacher judgements were combined in a weighted algorithm to grade vocabulary on the CEFR, taking into account research evidence on vocabulary size requirements (e.g., Hazenberg & Hulstijn, 1996). Our study proposes an integrated model of vocabulary assessment which combines quantitative observation of L1 usage of vocabulary in English with qualitative information about usefulness of vocabulary to talk about a range of different topics and carry out different communicative purposes.
Language test design appropriate to a target construct fundamentally determines obtaining key characteristics of observed language performances, which refers to the evaluation inference in an argument-based approach to validity (e.g., Chapelle, 2012; Kane, 1992, 2006). When it comes to assessing a target construct that involves context-specific rich language features, the evaluation inference becomes more critical with additional steps of collecting valid test scores, such as developing task-specific rating criteria and training raters. For this reason, ensuring validity evidence for the evaluation inference in assessing pragmatic competence in interaction, namely learner’s ability to accomplish various pragmatic actions in interaction, has been particularly challenging in the field of language assessment. Attempting to address the validity challenge, this study examines how a task-based assessment framework (Long & Norris, 2000; Norris, 2009) enables eliciting validity evidence of interaction-involved pragmatic performance employing sequential mixed methods (Green, 2007; Tashakkori & Teddlie, 2003). Two research questions guided the study: 1. In what ways do role-play tasks elicit task-specific features of interaction-involved pragmatic performance?; 2. In what ways are the features of observed task-based performances reflective of a varying level of pragmatic competence in interaction? Based on a large-scale needs analysis of stakeholders’ pragmatic needs in an English for Academic purpose context, five role-play tasks that involve different interlocutors (professor, classmate) and speech acts (request, refusal) were developed and completed by 102 ESL learners. 12 trained raters scored examinees’ role-play task performances using task-dependent analytic rating criteria. Multi-faceted Rasch measurement was employed to analyze the task difficulty and rater performance. For the first research question, select learners’ role-play performances were qualitatively analyzed employing a conversation-analytic (CA) method to examine how the examinees employed linguistic and interactional resources in achieving role-play task-specific pragmatic actions. The CA findings were compared across the role-play task situations. For the second research question, select features of interaction-involved pragmatic performances (e.g., sequence organization, discourse markers) from the CA findings were quantitatively analyzed to examine how such features are related to 12 raters’ scores. Findings include qualitative and quantitative validity evidence for pragmatic competence in interaction, which warrant that observed task-based pragmatic performances provide valid scores reflective of the target construct. Based on the findings, this paper further argues how the task-based language assessment framework and key concepts in pragmatics can be combined to offer
a useful guideline for designing pragmatic assessment tasks that tap into a target construct, that ultimately determine obtaining key validity evidence of assessing pragmatic competence.

Friday, June 24
12:10-12:40
Room 6
University of Palermo, Scuola Politecnica - Faculty of Engineering

**Language Testing in National Medical Licensure Procedures: Perceptions of Language and Communication Skills**

*John Pill, American University of Beirut*
*Sally O’Hagan, University of Melbourne*

Migrating doctors’ training and work experience can differ from what is expected in the destination country. The language of their previous study and practice may not match the new context either. Regulatory authorities in the new jurisdiction seek to ensure that quality of care and patient safety are maintained through the implementation of pathways to licensure that involve assessment of language and communication skills as well as of professional knowledge and clinical skills. Different practices have developed in different countries. In the United States, assessment of spoken English proficiency is integrated into a performance test for clinical skills taken by all medical graduates regardless of their place of training, while Australian law requires language skills – of migrating doctors only – to be assessed separately from professional competence. Policy has recently been revised in the United Kingdom in an attempt to manage tension between European Union freedom of movement legislation and professional accountability, following high-profile cases of malpractice involving doctors trained in other jurisdictions and media coverage about poor language skills among such doctors. A UK licensing assessment for doctors, similar to the US model, is currently being considered. As members of a research team on a large project in Australia, we investigated what doctors value in doctor–patient consultations and subsequently elicited doctors’ opinions to determine passing standards for the speaking component of a specific-purpose English test for health professionals. In research workshops held during this study, doctors expressed a range of views on the definitions of language, communication and professional skills and how these elements interacted in practice. Common themes included the difficulty in considering any one dimension separately from the others, and the effects of inauthenticity, perceived as inherent in all simulation for assessment purposes. Our paper seeks to relate these indigenous perspectives to the assessment practices implemented for migrating doctors in three jurisdictions – Australia, the UK and the USA – and to concerns expressed about these practices, for instance, in the media. In this context, we revisit fundamental issues in language testing regarding, for example, the definition of the construct being measured and test fairness. The assessment
regimes represent attempts to operationalize an often disputed boundary between language performance and content knowledge, taking into account at the same time particular legal frameworks and community expectations. We find that the constructs involved remain contested. The complex nature of language testing in practice is therefore highlighted.

Friday, June 24
12:10-12:40
Room 7
University of Palermo, Scuola Politecnica - Faculty of Engineering

**What's Vocabulary? Challenging Conventions to Re-define the Construct of Vocabulary Size Tests**

*Benjamin Kremmel, University of Nottingham*
*Norbert Schmitt, University of Nottingham*

What are vocabulary size tests assessing and what do their scores mean? The construct of vocabulary tests is often perceived to be relatively straightforward. However, it is based on a number of assumptions that do not seem empirically motivated, but nevertheless have become unquestioned conventions. This talk will problematize and challenge two key assumptions, and suggest a re-definition of these aspects of the construct to make vocabulary tests more diagnostically useful. The first convention addressed is the counting unit. Most vocabulary size tests have used word families, but there are reasons to think that this may not be the best unit. Even if learners know one or more members of a word family, they do not necessarily know all of its members. This assumption, however, is the basis of score interpretation in all major vocabulary tests. The first study presented will question this by comparing 99 German ESL learners’ knowledge of root forms with their knowledge of the entire word family. Results showed that learners only managed to make a connection between base word and other word family members in about 73% of the cases. Given the inappropriateness of word families as a counting unit, we give reasons why the lemma would be a more suitable counting unit, including enhanced interpretability of test scores and the potential to integrate multi-word expressions into lemmatized item sampling. The second assumption is that the vocabulary continuum should be divided into bands or levels of exactly 1,000. We will discuss current research which indicates that a one-size-fits-all approach to frequency division may not be the best solution. Based on coverage figures of frequency lists from 4 different corpora, we will suggest that more narrowly defined bands (500 units per band) might be more useful for higher frequencies, while wider bands (2,000) might be perfectly workable for lower frequencies, particularly for tests aimed at beginner to intermediate language learners, who are most often the target audience of vocabulary tests. Scores from vocabulary size tests only make sense if they are used for particular purposes, e.g. to check whether learners have enough vocabulary in order to read authentic texts. That is, vocabulary tests can no longer be seen as suitable for just any
purpose. The above changes to vocabulary test constructs and procedures would allow vocabulary tests to be better matched to specific purposes, and the scores interpreted in more informative ways.

Friday, June 24
15:30-16:00
Aula Magna
University of Palermo, Scuola Politecnica - Faculty of Engineering

Plenary

Is Conversation Analysis Really the Answer in L2 Pragmatics Assessment Research?

Kirby Grabowski, Teachers College, Columbia University

Although there is a long-standing tradition of qualitative methods in the second language (L2) testing literature (Lazaraton, 2008), pragmatics assessment researchers have only recently begun to provide evidence of test-score interpretation (primarily) through a discourse analytic lens. Since pragmatics is principally concerned with talk-in-interaction, it is not surprising, then, that conversation analysis (CA) has emerged as the method of choice for researchers interested in describing the interactional practices of learner-test takers. Uniquely, CA provides a way for researchers to highlight what is elicited in test data, specifically in terms of turn-taking, preference organization, and structuring practices (see Ross & Kasper, 2013). Furthermore, since many pragmatics assessments are designed for classroom use rather than for large-scale contexts, the link from descriptions of talk to pedagogy is a benefit that can and should be mined through these methods (Wong & Waring, 2010). However, in this presentation I will outline how the explanatory power of CA, first and foremost, is in relation to interactional practices in particular, and does not necessarily give primacy to pragmatic meanings in context (Crystal, 1997; Purpura, 2004), nor learners’ knowledge, skills, and/or abilities (KSAs) (i.e., the student model) (Mislevy, Almond, & Lukas, 2004; Mislevy, Steinberg, & Almond, 1999, 2002), which are all but ignored. This incongruity is also evident in many language assessment studies where pragmatic competence is conflated with interactional competence. In my view, it is the use of CA that has contributed to this phenomenon. Is interactional competence really what we are interested in measuring? If so, then we need to understand the resulting limitations on our test-score interpretations and pedagogical applications. If not, then a more precise definition of pragmatic competence for assessment is needed—one that not only holds the conveyance and interpretation of meaning-in-context as the fundamental unit of analysis, but one that also honors the ‘ability – in language user – in context’ view (Chalhoub-Deville, 2003). Ultimately, I will argue for an alternative discourse analytic framework—one that accounts for the inseparability of context (i.e., task descriptions/prompts) and the learners’ KSAs from the interpretation of the talk elicited (Chapelle, 1998; Mislevy, 2007)—as a more tenable approach than traditional CA.
The Written Discourse Completion Test: Different Versions, Different Constructs

Afef Labben, University of Tunis

One of the basic criticisms leveled against pragmatics elicitation tasks such as role plays and Discourse Completion Tests (DCTs) is that they fail to reproduce the authentic features of contexts in which real-life communication takes place. Such failure no doubt affects the quality of data gathered by these instruments. Yet, DCTs in particular have been the most popular methods used to collect pragmatics data due to their practicality and to the very limited access to authentic English data in EFL settings. Interlanguage Pragmatics (ILP), in particular, still relies on the use of DCTs to make inferences about learners’ pragmatic ability. A survey of a number of ILP studies shows that different versions of this instrument exist and that researchers often opt for comparing data collected via different DCT types. Billmyer and Varghese (2000) designed an unstructured DCT with enhanced situational prompts to investigate the effects of prompt enhancement on subjects’ production of requests and concluded that enhanced DCT items “elicit more robust external modification and elaboration than do the archetypal content-poor prompts which most DCT studies to date have used” (p.543). Johnston et al. (1998) investigated the effects of different types of rejoinders on the production of complaints, requests, and apologies and concluded that speech act strategy selection “is differentially affected by rejoinder type” (p.157). Such findings raise doubts as to the construct validity of different DCT versions and call into question the comparability of the data they yield. This paper reports on a study undertaken to validate different DCT versions in an EFL context. Six different DCT versions varying according to absence or inclusion of different types of rejoinders (positive and negative), and absence or inclusion of contextual details in situational prompts were administered to six equivalent and homogeneous groups of respondents (total number = 240). Each version included twelve apology and request situations set in different contexts. After completing the DCT, a questionnaire was administered to investigate informants’ evaluation of the DCT items to which they responded. Statistical analysis of the data used Logistic Regression analysis and the Mann-Whitney U test to investigate whether the respondents’ choices of apology and request strategies differed based on the DCT characteristics. Results show that the different DCT versions generated different types of data. Specifically, the context enriched DCT prompts produced longer utterances and affected respondents’ strategy selection and modification differentially while different rejoinder types interacted differently with content poor and content enriched situational prompts. Based on such results, the study raises the question of whether these different versions measure the same construct and calls for reconsidering the development of the written DCT.
Assessing Interactional Competence on Managing Topic Transitions in EFL Group Oral Task

Patharaorn Patharakorn, University of Hawaii at Manoa

Nearly two decades ago, He and Young (1998) made a very influential argument that interactional competence is fundamentally co-constructed, local, and situated; and therefore, interactional competence is not an attribute, nor a trait, of an individual. Instead, interactional competence, should be treated as jointly achieved (Jacoby & Ochs, 1984) only within that situated time and contexts. This study takes a slightly different viewpoint as it is seeking to provide evidence to support that, in a given specific set of actions and situations, individuals can show themselves to be more or less interactionally competent, and this could be viewed as the person’s interactional competence within that co-construction (Kasper & Ross, 2013). Talk is comprised of interactive practices and moments that individuals come together to articulate, and manage their collective histories via their resources (Hall, 1995). A skilled participant with strong interactional competence is someone who in that situation can make good use of the resources that are available. These resources include linguistic and pragmatic resources to the practices – knowledge of rhetorical scripts, knowledge of lexis and syntactic patterns, knowledge of turn-taking management, knowledge of topical organization and the means to transition from one topic to another (He & Young 1998). For this presentation we will focus on learner’s ability of management of topics demonstrated within an assessment group task. Thirty two undergraduate Engineering students from a university in Thailand participated in the study. They were grouped into 5-6 students per group. The data collection for this presentation consisted of video recordings of 6 small groups performing on a task eliciting group talk of 10 – 15 minutes during which each student played a role of a company representative attending an international trade show where English is the lingua franca. The task required students to engage in a casual conversation about their character’s general information, and then move the discussion to talk about their work. The transitions between social-talk and work-talk in engineering context lends itself very well to an examination of interactional competence construct for a specific institutional purpose. Conversation Analysis (CA) has been used as the methodological lens for this study provided its attention to detail in capturing the linguistic and semiotic resources, turn-taking designs, and topic management. The findings show interactional patterns how participants employ embodied actions and pre-sequences to create an open slot for them to subsequently carry out the topic transition. In all of the group interactions, students who took this as their responsibility demonstrated considerable amount of planning, and an ability to recycle old information as resource for later turn constructions. In all these cases, the scores assigned by raters who have no prior training in CA also corroborate with
this study’s result. It is evident, therefore, that ability to manage topic transitions is salient for participants and raters within this context. Further inspection into other aspects of interactional competence is still needed to uncover the construct of interactional competence for language assessment purposes in this context.

Friday, June 24
16:05-16:35
Room 7
University of Palermo, Scuola Politecnica - Faculty of Engineering

Exploring the Relationship Between Lexical Measures and L2 Proficiency

Veronica Benigno, Pearson
William Bank, Pearson
Masanori Suzuki, Pearson

Vocabulary is a crucial component of language use which has been shown to contribute to variance in proficiency, particularly in reading ability (Alderson, 2000; Qian, 2002). As much research has pointed out in the past few decades, vocabulary knowledge is a multidimensional construct involving both quantitative and qualitative aspects, namely vocabulary size and depth such as meaning, form, and use (Meara, 2005; Nation, 2001; Schmitt, 2010). The depth dimension in particular has been neglected in assessment research, perhaps because assessing vocabulary choices in a particular domain or situation presents some practical challenges. The quantitative dimension of vocabulary is generally assessed by measures such as frequency profiling, but assessment of qualitative aspects (e.g., use of collocations) generally requires expert rater judgements of, for example, whether particular lexical choices were appropriate. This research will bring both approaches to bear on a large dataset, to help understand vocabulary’s relationship with L2 proficiency. Over 5,000 adult learners of English participated in field testing for a four-skills English proficiency test. Four item types out of 20 specifically elicited responses in speaking and writing that contributed to the vocabulary score on the test. For spoken vocabulary, each test-taker responded to two item types, Describe Pictures and Story Retells. In Describe Pictures, test-takers see an image and describe it orally. In Story Retells, test-takers listen to a story in English and retell it orally. Each of these item types elicited about 30 seconds of extended spoken responses. Additionally, Short Writing and Passage Recalls collected productive vocabulary in writing. Short Writing required a small amount of reading (a sentence or two) to understand the brief writing prompt, and test-takers had time to write their opinion about a given topic. Passage Recalls was an integrated-skills task combining reading and writing skills. Test-takers read a passage for 30 seconds, and reproduced its content in writing in 90 seconds. The responses from these four item types were subsequently presented to human raters trained in using CEFR-based rubrics for collecting independent judgments on test-takers’ vocabulary levels based on spoken and written responses, as well as their overall spoken or written proficiency levels. Over 150,000 ratings
were available for analysis. With this dataset, two research questions are currently being investigated: To what extent do lexical measures such as lexical diversity and correct use of collocations in terms of type, token, and corpus frequency predict human raters’ CEFR estimates of test-takers’ overall proficiency and vocabulary levels? Among the four item types that elicit productive vocabulary skills, which task exhibits the most discriminatory power in measuring test-taker’s overall proficiency level? Data collection on lexical diversity measures is still in progress and the analysis of the data is upcoming. This study has implications both for theoretical approaches to proficiency testing as well as for the practical aspects of vocabulary assessment. Our study supports the need for an integrated model for assessing vocabulary use that combines quantitative and qualitative measures and shows how effective different assessment tasks are in assessing vocabulary knowledge and overall proficiency.

Friday, June 24
16:05-16:35
Room 8
University of Palermo, Scuola Politecnica - Faculty of Engineering

**Practice Makes Perfect: the Likelihood of Success in Retesting**

*Rachel Brooks, FBI*
*Maria Brau, FBI*

Testing research indicates that familiarity of test format, as is the case in instances of retesting, retesting increases the chance of test success (Greenfield, 1997), or test score increase. However, practice and retesting does not always guarantee success as may be expected by the examinees (Bloomfield et al, 2012). In the case of the US Federal Government, linguists are often hired with minimum test scores and can only provide limited services. They hope that practice over time will allow them to improve their test scores and be promoted to a higher status where they can provide more services and earn additional pay. Retesting after practice, however, does not always lead to consistent success. Decreased scores are not predicted, particularly with the expectation that high-proficiency language users are resistant to language loss (Bardovi-Harlig & Stringer, 2010). This research project investigates whether or not retesting after a period of practice does typically increase test scores. The data come from over 1,000 instances of retesting in both translation performance tests and speaking proficiency tests. Data are categorized for analysis by test type, original test score, and language, with more than 10 languages represented. Factors such as time between the original test and retest are considered in the analysis as intervening variables. Data are analyzed using ANOVAs and other inferential statistics. Interviews and surveys about test perceptions and retest expectations were conducted with a subset of participants to triangulate the test results. Preliminary results suggest that practice far from guarantees success in retesting across languages. In fact, retesting can sometimes lead to lower test scores, particularly when the original test
score was quite high. There are some differences in how much practicing helps in a performance test (translation test) over a proficiency test (speaking test). The results have broad implications for the examinee, as they give an indication of when and how much practice is useful to give the best performance possible on a test. Additionally, the results reflect the impact of test type and gives insight to whether the test is measuring what it should.

Friday, June 24
17:00-17:30
Aula Magna
University of Palermo, Scuola Politecnica - Faculty of Engineering

Plenary
Construct Validation of Multimodal Scenario-Based Language Assessment (SBLA) Tasks for Diagnostic Placement Purposes

Eunice Eunhee Jang, Ontario Institute for Studies in Education, University of Toronto
Maryam Wagner, Carleton University
Maggie Dunlop, Ontario Institute for Studies in Education, University of Toronto

Recently, there has been increasing interest in scenario-based language assessment (SBLA) tasks in order to broaden the construct of assessing language ability in order to engage test takers in activities designed to elicit processing underlying communicative activities in non-testing contexts in an integrative way (O’Reilly & Sabatini, 2013; Purpura, 2015). Although promising, innovative features of SBLA tasks can pose challenges in calibrating test takers’ performance data using traditional scoring models partly due to their nested structures violating the assumption of local independence. Few studies have examined the principles of SBLA task design and validity evidence for psychometric rigor in scoring and score interpretations.

The present paper discusses the principles of SBLA task design, their psychometric properties, and implications for guiding program placement and informing instruction in a bilingual college context, based on a multi-round SBLA development and validation project over three years. The resulting on-line SBLA system replaced the college’s old placement test and became operational in 2014. The project took place over three phases: (1) context and needs analyses based on curriculum materials, interviews with instructors, and classroom observations; (2) SBLA task design and pilot-testing based on 153 college students’ task performance and think aloud verbal protocols; and (3) SBLA field testing involving 435 students. The target construct was operationalized into six knowledge and skills based on the thematic analyses of qualitative data from classroom observations, interviews, and curriculum materials, as well as analyses of the definition and operationalization of the construct in the literature and in major language tests. Four multimodal integrative academic learning scenarios were created to elicit students’ deep processing. They included a total of 7 tasks: students’ online application forms (9 items), class lectures (5), assignment instructions (6), a group project assignment (5), class readings (8),
summary of readings (4), and synthesis of multiple text genres (10). The SBLA tasks were multi-staged, integrating four modalities, and were designed to elicit authentic communication gaps. The specifications of the SBLA tasks were guided by Skehan’s model of task difficulty (1998). Students’ performance data were calibrated using confirmatory multidimensional latent trait analysis techniques. The study results showed strong empirical evidence supporting the psychometric properties of the SBLA tasks. Mean item-level discrimination values for each of the tested skills ranged from .49 to .59. In particular, over 78% of items assessing grammar knowledge showed the highest diagnostic discrimination power. Overall skill mastery estimates ranged from 39 to 50%, suggesting that the SBLA tasks were appropriate for placement purposes prior to instruction. Constructed-response tasks contributed more significantly to discriminating mastery levels among the skills. We conclude that SBLA development and validation approaches can strengthen the triangular relationships among theoretical models of language constructs, observations from tasks eliciting the operational components and skills underlying the constructs, and test-score interpretation and use (Mislevy, Steinberg, & Almond, 2003; NRC, 2001). We further discuss implications for enhancing SBLA practice through the design and validation principles.
Assessment of Indigenous Languages: Examining the Cases of Australia and Canada

Beverly Baker, University of Ottawa
Gillian Wigglesworth, University of Melbourne

Australia and Canada share a great deal in terms of their colonial history and their treatment of Indigenous minorities. Both countries are home to a vast diversity of Indigenous peoples who once spoke hundreds of languages. Both countries have a history of active repression of these languages in favour of the colonial languages (English, and also French in the case of Canada). Now, most Indigenous language varieties have disappeared in both contexts, and the remainder are classified as endangered by UNESCO (2011). However, with increasing recognition of the importance of the vitality of Indigenous languages for the well being of these populations, both of these countries are experiencing renewed and increasing interest in Indigenous language revitalisation.

This symposium brings together scholars working in the area of indigenous language assessment in Australia and Canada. Their work includes an examination of macro-level concerns regarding educational and aboriginal policy, as well as micro-level inquiry into Indigenous language assessment practices. Symposia such as this are vitally important for knowledge translation. As Rau (2014) maintains, “Continuing the interaction and collaboration between indigenous groups (and others) means cross-fertilisation of ideas and responses [...] and provides a rich source of mutual support and direction for ongoing and future development” (p. 328).

This symposium will cover considerations of appropriateness of construct definition for these contexts, and cultural relevance of assessment tools and procedures, as well as alignment with locally-based language learning initiatives that seek to work within indigenous worldviews. We end by proposing characteristics of an indigenous approach to assessment, which include primacy of oral communication; the importance of collaboration rather than competition, for example through peer assessment and collaboratively-produced assessments; and the importance of involving Elders and other community members in the assessment process.

The 90-minute symposium will follow the following schedule:
- Introduction: Challenges in the assessment of indigenous languages (15 min)
- Effects of standardized testing on language endangerment (15 min)
- Canadian narrative inquiry (15 min)
- Development of an indigenously-informed tool in Western Canada (15 min)
- Discussant: Elana Shohamy (10-15 minutes)
- Questions and discussion (15 minutes)
**Paper 1**  
**Challenges in the Assessment of Indigenous Languages**  
*Gillian Wigglesworth, University of Melbourne, Australia*

Indigenous languages in both Australia and Canada have been severely depleted since colonization. This loss of languages represents not only a loss in understanding our cognitive processes, but for their speakers also means an incalculable loss of culture and knowledge and often impacts severely on their self esteem and feeling of self worth and belonging.

Indigenous languages can be assessed in two ways. The first is through measures of endangerment, which aim to identify the potential for the language to be maintained, or alternatively, its potential for complete loss. In Australia, of the original 250 to 300 languages spoken when Europeans first arrived on its shores, only around 20 now continue to be learned by children. For the remainder, they may be spoken only by the older generation in particular communities, but many are now only spoken by a very few of the oldest generation, whose life expectancy tends to be relatively short. A good example of this is the two last speakers of Traditional Tiwi, a language spoken in the Tiwi Islands, who both died with a few months of each other in 2010.

The second way in which we assess Indigenous languages are through the assessment of the linguistic knowledge of individuals. These kinds of assessments may be undertaken in schools, in language revitalization programs, or in programs which teach indigenous languages to non indigenous people. This type of assessment presents multiple challenges in these different contexts. Assessment tools are often limited, and are rarely standardized due to small numbers; it is crucial for assessments to be culturally relevant and appropriate, and they must be validated, whenever possible, by Indigenous people themselves.

This paper focuses on the complexities of the assessment of indigenous languages in Australia, including the challenges which have emerged and the ways in which these challenges can be addressed.

**Paper 2**  
**The Impact of Standardised Testing on Children in Indigenous Communities**  
*Ute Knoch, University of Melbourne, Australia*

This paper reports on the National Assessment Program – Literacy and Numeracy (NAPLAN), a standardised assessment administered to all Australian school children at key year levels. The results of NAPLAN are intended to guide government educational policy, and understanding the extent and nature of the effects precipitated by a test is a key aspect of test validation, but it is often neglected. This study aimed to investigate the impact of NAPLAN testing on children in remote Indigenous communities from a range of perspectives, with a particular focus on their first languages.
Children in remote indigenous communities in Australia do not learn English as a first language. Rather they come to school speaking either an Indigenous language, or a creole, depending on what is spoken in the community. The range of creoles spoken across the north of Australia reflects the very significant loss of many indigenous languages in Australia.

A series of interviews with key stakeholders (principals, teachers, and education support staff) were conducted to establish informants’ first-hand experiences of the use of NAPLAN at their school and its impact. The focus of this paper is on the impact of the NAPLAN test on the children’s first languages, both in schools, which teach bilingual programs, and English only schools.

The findings showed that all informants agreed that the test in its current form is not an appropriate tool to assess children in remote indigenous communities, mostly due to the different language and cultural experiences of indigenous children who often only start learning English when they start school. The findings also showed that the test had an emotional impact on students on the student-teacher relationship. In addition, poor NAPLAN results by indigenous students have been directly linked to the closure of bilingual programs in remote indigenous communities. While the impact of the test differed across contexts, there was across the board concern about the children’s first languages, particularly in the context of literacy development.

This study responds to the need identified in the introduction to this symposium to investigate the cultural relevancy of assessment and to include indigenous stakeholders in assessment validation activities.

**Paper 3**

**A Narrative Inquiry into the Formative Assessment Practices of an Indigenous Language Teacher**

*Joyce Germain, Listuguj First Nation*

*Beverly Baker, University of Ottawa, Canada*

This presentation addresses the assessment of the linguistic knowledge of young children within a school-based Mi’gmaq immersion context in Listuguj, eastern Canada. In Listuguj, the proportion of fluent speakers is less than 20%, most aged over 65. In this presentation, I will share the stories of one educator’s formative assessment practices in the classroom and on the land. This work is part of a nearly decade-long university-community partnership dedicated to Mi’gmaq revitalization across the lifespan.

Through a series of conversations and anecdotes, this educator reveals how her observations of student language use inform her decision-making and the evaluation of her explicit and implicit learning objectives. Critical moments of assessment are identified which reveal the importance placed on the use of language for reinforcing specific community values, such as skills on the land, teamwork, and demonstration of respect for Elders.
This study was conducted using narrative inquiry (Chase, 2005; Creswell, 2008; Riessman, 2008), a research methodology uniquely well-suited to research within indigenous contexts. Narrative inquiry involves the analysis of data from a variety of field-based sources (including stories, photographs, and interviews) in creating meaning within a narrative or story framework. Storytelling is an important part of indigenous culture, and is found front and centre in curricula for Canadian indigenous language and culture education programs (Author, in press). This work responds to the call to decolonize the research process within indigenous communities (Smith 1999; Steinhauer, 2002), attempting to give voice to a group that has been previously marginalized in the research enterprise.

**Paper 4**

**NETOLNEW 'one mind, one people': Developing a Context-Relevant Assessment Tool for Adult Indigenous Language Learners in Canada**

Onowa McIvor, Peter Jacobs, & Barbara Jenni, University of Victoria, Canada

This final presentation explores the creation and validation of an assessment tool to bridge the gap between Western notions of what constitutes progress and success, and the more holistic understandings found in Indigenous worldviews.

The NETOLNEW ‘one mind, one people’ project investigates adult Indigenous language learning and the contribution the increasingly popular Mentor-Apprentice Program (MAP) method can make to the language revitalization movement. The prompt creation of adult speakers of Indigenous languages is essential for communities seeking to revitalize their languages, in part to meet the growing demand for language teachers for children, be it in immersion pre-schools or grade school.

Embedded within the overarching research study is the following question: How successful are these adult learners in MAP? However, there is no comprehensive, contextually relevant assessment tool available to measure this. Therefore, a sub-project within the larger study was the collaborative creation and piloting of such a tool.

The assessment tool is currently being validated with learners and teachers across multiple languages in British Columbia, focusing on how the tool:

- relates to self-directed learning goals defined in one-on-one immersion settings;
- responds to various levels of proficiency, including beginners with no prior language exposure;
- considers the lack of conventional opportunities to use language, such as interacting with other speakers in diverse settings, literature, media, etc.;
- provides a set of standards or criteria applicable across different languages, cultural contexts, but without reference to a curriculum or classroom setting; and
Using a Comprehensive Framework to Integrate Constructs, Contexts, and Content in a Large-Scale Language Assessment

Dorry M. Kenyon, Center for Applied Linguistics
David MacGregor, Center for Applied Linguistics

This 90-minute symposium illustrates, through four related papers, the need for and practical use of a comprehensive framework for the development and validation of language tests. The first paper presents a seven-layered framework joining Bachman and Palmer’s Assessment Use Argument with Mislevy et al.’s Evidence-Centered Design. The following papers illustrate its use for understanding, planning, developing, implementing, communicating and researching complex issues in language assessment. The symposium will be of interest to a wide variety of language testers who work on practical language assessment projects integrating theory and practice, and who need to situate their own work and expertise into larger multidisciplinary contexts and changing societal issues.

Language testing is a multidisciplinary endeavor. Foundational theory and practice in the fields of applied linguistics and educational measurement are challenged and expanded by insights from fields as diverse as cultural anthropology, cognitive psychology, and critical policy and ideology studies. As the scale of the language testing endeavor increases, challenges in bringing multiple insights together also increase. In this symposium, illustrations of the use of the comprehensive framework come from a large-scale English language assessment of Speaking, Listening, Reading and Writing for English language learners (ELLs) in grades Kindergarten to 12 in the United States. Results on the assessment are for two main accountability purposes: (a) to ensure that ELLs are acquiring the type of English required for academic success in English-medium classrooms, and (b) to identify ELLs who may be re-classified out of that legally-protected status. This assessment program, used by a consortium of states, began over 10 years ago. As of the current academic year, it has grown to be used by 36 states and assesses about two million ELLs a year.

The framework was developed to provide guidance to work across disciplines and organizations on a four-year federally-funded project to respond to societal
and technological changes. The societal changes included the adoption of more rigorous academic standards. The technological changes included the government’s encouragement of the use of technology in large-scale standardized testing.

Paper 1 (10 minutes) provides background on the need for and the development of the framework to guide complex language assessment research and development work. Paper 2 (20 minutes) focuses on context and illustrates how the framework was used to understand and articulate the impact of the societal changes in terms of changes to interpretations, decisions and consequences in the use of test scores.

Paper 3 (20 minutes) focuses on content and construct, and illustrates how the framework was used, in light of these societal changes, to guide the re-investigation of the assessment domain.

Paper 4 (20 minutes) focuses on assessment implementation in the light of the affordances of new technology, particularly for reaching measurement goals. This paper focuses on the part of the framework where Evidence-Centered Design joins the Assessment Use Argument.

Our discussant (10 minutes) will address some of the strengths and weaknesses of the framework.

The symposium will end with 10 minutes for audience questions and discussion.

**Paper 1**

**The Need for a Comprehensive Assessment Framework**

*Dorry M. Kenyon, Center for Applied Linguistics*

This short paper (10 minutes) provides an introduction to the comprehensive framework illustrated in this symposium, as well as background to the need for, development of, and purposes for the framework. In 2011, a large consortium of over 30 states and its partners were awarded a four-year grant from the U.S. government to transform an extant large-scale annual assessment of academic English language development for grades Kindergarten to 12, used in a national accountability program, in light of two principle developments: the introduction of more vigorous academic achievement standards and the push for the use of technology in large-scale educational assessments. A tool was needed to describe all the moving parts to internal and external stakeholders, including national review committees. To facilitate communication, a seven-layered framework was built through the integration of Evidence-Centered Design (Mislevy et al.) and the Assessment Use Argument (Bachman and Palmer), which were already familiar to some stakeholders. The framework was needed to help all parties involved (a) understand the complex issues to be addressed, (b) situate their own work and the work of other researchers and experts within the larger endeavor, (c) allow them to value their own work and the work of others, (d) prioritize the use of limited resources, and (e) more clearly communicate between internal teams within the complex project and also with external stakeholders.
**Paper 2**

**Using the Framework to Address Issues Related to Test Consequences, Decisions, and Interpretations**

*Ahyoung (Alicia) Kim & Mark Chapman, WIDA at The University of Wisconsin-Madison*

This paper concerns an annual summative large-scale assessment of academic English language development for K-12 English language learners (ELLs) within the United States. Approximately two million students take the exam each year, creating a large impact on various stake-holders (e.g., states, districts, teachers, parents, students). After the exam, students' test performance is reported to various stake-holders using four different types of score reports (i.e., Individual Student Report; Student Roster Report; School Frequency Report; District Frequency Report). Using the comprehensive assessment framework presented in this symposium, this paper discusses the use of the test scores. It describes the interpretations that are made by stake-holders based on the scores, and the decisions and consequences that follow. The interpretations of the scores are meaningful with respect to its foundational standards; impartial to all groups of test takers; generalizable to English language use tasks in academic settings; and relevant to and sufficient for the various decisions to be made in K-12 educational settings. The decisions that are made based on the scores by the stake-holders include the type and amount of ELL service provided to students, and reclassification of the ELLs. The decisions that are made on the basis of the interpretations take into consideration existing K-12 ELL educational values and relevant legal requirements (e.g., No Child Left Behind Act of 2001), and are equitable for ELL students who are affected by the decision. These decisions further impact the consequences of using the scores.

**Paper 3**

**Using the Framework to Address Issues Related to Constructs and Content**

*Jennifer Norton, Center for Applied Linguistics*

Assessing the construct of academic English development via a large-scale English language proficiency test in the U.S. K-12 context presents the unique challenge of drawing on academic grade-level content in test items, while avoiding background knowledge as a requirement to respond to a task or item. In developing the test, the domain analysis and modeling must account for a variety of perspectives, including foundational English language development (ELD) standards and college and career readiness standards of academic achievement. In the case of the consortium’s test, the consortium’s foundational ELD standards define six levels of language proficiency in developing academic English language. In constructing test items based on these standards, it is imperative to create items that test language at the target proficiency level, especially at the higher end of the scale at which high-stakes decisions are being made about whether students have enough English language proficiency to be re-classified out of their status as English language learners. In this paper, we use
the comprehensive assessment framework to report on work intended to give item writers guidance on how to create selected response items at the desired level of difficulty, based on construct-relevant factors, with a focus on more linguistically demanding items. We first identify operational items that are testing at the upper levels of proficiency. We then examine those items to identify construct-relevant characteristics that seem to contribute to their difficulty.

**Paper 4**

**Using the Framework to Address Measurement Issues in Assessment Delivery**

*David MacGregor & Xin Yu, Center for Applied Linguistics*

In creating an assessment of developing academic English language proficiency which assesses multiple domains across six proficiency levels, test designers need to balance competing factors. Items must span a wide range of difficulty; however, for both psychometric and affective reasons, the difficulty of the items that individual students see should closely mirror their ability levels. Using the comprehensive assessment framework illustrated in this symposium, this paper explores issues at the juncture of Evidence-Centered Design (test assembly and delivery) and the Assessment Use Argument (assessment performances). Affordances of the new on-line approach to assessment allow for the use of multi-stage adaptive testing for English language learners in the domains of listening and reading. This approach replaces the tiered approach used in the earlier paper-based test, whereby students were placed into one of three overlapping tiers by local educators based on the educators’ assessment of the students’ abilities. This earlier approach to tailored testing was limited by the accuracy of the educators’ assessment of student proficiency and by the necessity of students being placed in the same tier for both domains, despite potentially having different abilities in those domains. In this paper we illustrate how the multi-stage adaptive approach, arrived at through using the comprehensive assessment framework, allows these issues to be minimized. Multi-stage adaptive testing ensures students are more appropriately challenged, allowing for a more precise measure, while maintaining full coverage of the academic English language of the content areas.

Discussant: Lyle Bachman, University of California at Los Angeles
How does Context Mediate Construct and Content? A political, social and pedagogical analysis of English language testing in Asia

Liying Cheng, Queen’s University
Antony Kunnan, University of Macao Guangdong University of Foreign Studies

Language testing is a social practice, and to place language tests beyond the micro level of test construct and content into their macro social context, language testers need to fill in gaps in the theorizing of social and educational contexts, so that the values implicit in test construct and content can be investigated and articulated explicitly (McNamara & Roever, 2006). Recent research has highlighted the paramount importance of context(s) in understanding the concept of “ability-in-language user-in-context” (Chalhoub-Deville, 2003), that is, the social and educational contexts where test-takers are taking tests, and such understanding is critical for test development and for test score validation (Cheng, et. al., 2014, Kunnan, 2008, Fox, 2004). However, context in language testing is understood differently: Does context mean the level of language learned (school, university, or a EAP program), the type of language used (business, health care, or tourism), the geographical location where the test is designed, administered, and used (China, Japan, or the Philippines), the location (a classroom, a test preparation center, or a test centre) where a test takes place, or various test-taker characteristics (Kunnan, 1995)?

This present symposium wrestles with the meaning of context and explores how a given context or context(s) mediate(s) construct and content using four distinct case studies as the focal points for discussion. The goal of the symposium is to highlight the importance of context in test development and test validation. The geographical context of Asia is focused on in order to raise convergent and divergent issues across other geographical contexts. The first paper discusses issues of concern to the College English Test (CET) developer in China to exemplify how the social and educational context influences the construction and revision of the CET. The second paper presents a macro-level analysis of the competing social, political, and cultural values linked with the government’s decision to first develop and then discontinue a new large-scale English test series in Korea: NEAT (National English Ability Test). The third paper discusses test fairness from the points of views of the stakeholders involved with test design, test administration, teaching and testing on how they understand and interpret the Thailand testing context. The fourth paper presents various ways in which academic English is conceptualized across disciplines and programs and its relationship to knowledge and skills in L1 unique to the Japanese EFL context. Taken together, the symposium highlights the mediating role of
context(s) and raises the issue of how language tests serve as instruments of public policy from the political, social and pedagogical perspectives (Pellegrino, 2004).

In this 90-minute symposium, the organizer (Liying Cheng) will first introduce the nature of language testing in Asia in relation to the political, social, and pedagogical contexts (5 minutes). This introduction will be followed by four case studies, each of 15 minutes. At the end, the discussant (Antony Kunnan) will highlight recurring themes, and indicate future directions for studies on the role of context in mediating construct and content (5 minutes). The audience will have 20 minutes for questions and discussions.

**Paper 1**  
**The Contextual Mediation of Educational and Social Features Influencing Test Construction**  
*Yan Jin, Shanghai Jiao Tong University*

China is known for its long history of testing and its heavy reliance on testing for making high-stakes decisions. Tests therefore have a powerful impact on China’s educational and social systems (Cheng et al., 2010). Our experience with the development of the College English Test (CET), a high-stakes test in China, shows that the educational and social context in which a test is developed and used could be important contextual features mediating the test’s construct and content. In this presentation, I will discuss issues of concern to the CET developer to exemplify how the educational and social context influences the construction and revision of the CET. The pair discussion task in the computer-based CET Spoken English Test (CET-SET), for example, was implemented, at the considerable expense of test efficiency, in order to draw attention from learners and teachers to interactional competence (He & Young, 1998; Young, 2003, 2010, 2011). The translation task was designed with the intention of answering the government’s call for promoting Chinese culture. The item-writing guidelines for translation specify that source texts must be on topics of Chinese custom, culture, history, social and economic development. Impact by design, therefore, should be an important guiding principle for conceptualizing and defining test construct in the Chinese context. Future research needs to further expand the context part of the interactionalist construct definition proposed and advocated by Chapelle (1994; 1998; 1999) through a macro-societal analysis of the features that influence a test’s construct definition and content specifications.

**Paper 2**  
**The Demise of NEAT: Whose Defeat or Triumph Is It?**  
*Yong-Won Lee, Seoul National University*

In recent years, there has been a heightened awareness among language testers that language assessment is a social practice (Cheng et al., 2004; McNamara & Roever, 2006). Due to such awareness, it is now standard practice to consider the social context of testing in designing and developing new assessments and evaluating the
quality of these assessments (Bachman & Palmer, 2010). One way to incorporate these social values into the test development and validation process is to examine the intended purposes and uses of the test as well as the aspirations of major stakeholder groups in a given testing context. Especially, when the government initiates a new test development project as a policy tool to push for educational change, it is critically important not only to evaluate the feasibility and adequacy of the proposed project but also to seek harmonious coordination and resolution of potentially conflicting values of the stakeholder groups, including the government. A major focus of the present investigation is a macro-level analysis of the competing social, political, and cultural values linked with the government’s decision to develop and scrap a new large-scale English test series in Korea: NEAT (National English Ability Test). More specifically, the presentation discusses the results of analyses on social, political, and ideological factors that might have prompted the initial decision to develop NEAT and ultimately led to the doomed fate of the test. The results are also analyzed in terms of their ethical and practical implications for language testing.

Paper 3
Understanding “the Ability-in-Language User-in-Context” through the Eyes of Stakeholders
Jirada Wudthayagorn, Chulalongkorn University

In Thailand, the O-NET (Ordinary National Educational Test), a high-stake and mandatory test for core subjects including English, is administered annually by the National Institute of Educational Testing Service to millions of students covering grade 6, grade 9, and grade 12 students in public and private schools. O-NET, a paper-pencil test, is aligned with the Basic Education Core Curriculum (2008). The final score, which determines grade promotion, is based on the O-NET score (20%) and school-based assessment score (80%). Over the past years, there have been paramount criticisms of the quality of the English O-NET and school-based assessments. Saiphet (2014) investigated item bias of English O-NET multiple-choice questions for grade 12 students. The difficulty index showed that some test items are easier for female students such as items about “flight attendant” and “restaurant.” Another bitter criticism of this test is the construct underrepresentation where listening, speaking, and writing have never been tested. School-based assessments are also controversial. Many students from remote areas usually attend schools with less resources and facilities. Fairness of school-based assessments across the nation is doubtful. This presentation focuses on test fairness at both micro and macro levels where the English O-NET test is designed and administered. “The ability-in-language user-in-context” (Chalhoub-Devile, 2003) is unfolded by interviewing the stakeholders, e.g., policy makers, item writers, school teachers, and students who are involved with test design, test administration, teaching and testing on how they understand and interpret the testing context.
Paper 4

English Language Demands at Universities - What is the Construct of “Academic English” in the Japanese Context?

Yasuyo Sawaki, Waseda University

In the context of English-as-a-foreign-language (EFL) education in Japan, there is growing interest in adopting four-skills English language assessments for university admissions reflecting the Ministry of Education, Culture, Sports, Science and Technology’s (MEXT) movement toward enhancing Japanese students’ communicative English language ability and promoting English medium instruction at universities. While this movement has drawn the attention of English language researchers and practitioners to the feasibility of implementing such tests in a large scale, relatively little has been discussed so far regarding what such tests should assess in light of the nature of English language demands at universities in Japan. A thorough examination of this issue is essential because English is used in the Japanese EFL context as the medium of instruction at various degrees across disciplines and programs with unique conceptualizations of academic English and its relationship to knowledge and skills in L1. This qualitative case study examines the nature of academic English language ability required in six disciplines in natural and social sciences as well as humanities at a large private university in Tokyo. Data from semi-structured interviews of faculty members from those disciplines will be analyzed in terms of (1) frequent and important language use tasks involved in content courses at undergraduate and graduate levels taught in various degree programs involved and (2) the levels and profiles of English language ability the faculty members perceive to be required for successful performance in their programs. The degree of congruence between MEXT’s expectations and value implications of the current practice of English language use in instruction that emerge from the study results will be discussed.

Friday, June 24
9:00 to 10:30
Aula Magna
University of Palermo, Scuola Politecnica - Faculty of Engineering

Evolving Academic English Constructs and New Innovative Tasks

Xiaoming Xi, ETS
Larry Davis, ETS

Evolutions of the constructs of academic English proficiency have primarily been motivated by two major factors -- advances in language testing theories and practices and the changing nature of communication in real-world academic contexts. In particular, recent conceptualizations of language constructs have
converged on the interactionalist view, which treats context as an integral part of a construct definition. This conceptual approach has prompted an increasing attention to context-oriented constructs in practice. Furthermore, the continuing emphasis on the use of sources in communication and growing prevalence of computers and multi-media technologies on university campuses have profoundly impacted how people communicate and thus how the constructs of academic English proficiency should be defined and operationalized.

This symposium focuses on the design and evaluation of innovative reading, speaking and writing tasks that are aligned with the interactionalist approach and responsive to the evolving academic language use domain. In alignment with the interactionalist view, the speaking and writing tasks are situated in authentic academic contexts to elicit expanded evidence of pragmatic competence, i.e., the appropriateness of candidates’ language for a particular audience and situation. The reading, speaking and writing tasks are also designed to reflect real-life communication in the academic domain, which engages digital literacy and critical thinking skills and involves extensive use of source materials. A theoretical orientation to defining test constructs using the interactionalist approach will be provided and the roles of language ability models and domain (contextual factor) models in designing contextualized tasks will be explicated. Additionally, the complex trade-offs involved in task design will be discussed, including striking the right balance between assessing linguistic skills vs. higher-order thinking skills, balancing the need to represent key language competencies and contextual factors in test design against practical testing constraints, determining the degree of task contextualization while being cognizant of fairness and accessibility issues, and determining an optimal way to represent a particular context in task design clearly, efficiently and effectively.

Four empirical studies on the design and validation of innovative reading, speaking and writing tasks will be presented, focusing on the theoretical underpinnings for these tasks, design challenges and proposed solutions based on multiple lines of evidence. Such evidence will include analysis of candidates’ cognitive interview data, response data, reactions to overall and specific design features and quantitative analyses of candidates’ performances. These studies illustrate a theory-based, principled way of task design that involves multiple rounds of iterations and a task evaluation approach that carefully triangulates multiple lines of validity evidence.

**Paper 1**

**Integrating Multiple Sources and Digital Affordances into Reading Assessment: Opportunities and Challenges**

Tenaha O’Reilly, ETS  
Mary Schedl, ETS

Advances in technology have reshaped how people read and communicate in the 21st century. For instance, modern reading environments require citizens to deploy...
a host of digital literacy skills (Leu et al., 2013), to seek information, solve problems and make decisions. These skills often require people to evaluate and synthesize information from multiple sources, authors, and perspectives (Rouet & Britt, 2011). Despite these advances, many assessments have not kept pace with these changes and how people read in authentic contexts (Gordon Commission, 2013).

The purpose of this paper is to describe two studies that evaluate prototype reading tasks requiring students to integrate information over multiple sources within a digital environment. The first study investigated whether language abilities and cognitive processing strategies beyond those involved in the comprehension of single texts are used in responding to questions that require processing information from multiple expository texts. Think-aloud data and comprehension questions probed single-text understanding and the ability to integrate information from multiple sources written by different authors.

In the second study, over 100 students took a computer-based reading assessment requiring them to understand thematic content of three passages on two different topics. The multiple sources presented different perspectives or conflicting theories on a particular phenomenon. Digital affordances such as supporting graphics and hyperlinks of technical vocabulary were provided to aid understanding. Test takers’ attitudes towards the prototype and their use of digital affordances were also collected. The results indicated the assessment was feasible and demonstrated adequate properties.

**Paper 2**

**Incorporating Audience into Source-Based Writing Tasks – Design Challenges and Considerations**

Yeonsuk Cho, Ilkyu Choi, Robert Kantor, & Jakub Novák, ETS

We will report on an effort to design innovative academic writing tasks in order to show how the design of such complex tasks evolves iteratively, based on a series of evidence gathering and revision stages, to achieve the test designers’ expectations regarding how the tasks should function. In this effort we articulated a construct definition of writing that includes the ability to write effectively for a specific audience and purpose, and that puts emphasis on writing from sources. Based on this construct, we designed various assessment tasks to represent the complexity of authentic writing in English-medium, college-level academic contexts by requiring test takers to consider a specific audience and purpose while incorporating information from sources in their responses. The new writing tasks were trialed in multiple stages with groups of adult ESL learners. We will present various types of empirical evidence including writing task responses, and data from cognitive interviews, surveys, and keystroke logs that supported our decisions to make revisions to task design. We hope that our session will demonstrate that test designers’ intentions for how newly designed tasks will function do not necessarily match how the tasks are perceived and experienced by test takers, particularly when the task types in question are new and unfamiliar. Moreover, we hope to build a convincing argument that successful task design depends on an iterative evaluation process including multiple stages of user feedback gathering and task revision.
**Paper 3**  
**User Perceptions of Task Design Features Intended to Improve the Test-Taking Experience and Validation Argument for Innovative Speaking Tasks**  
*Jonathan Schmidgall, Yuan Wang, Larry Davis, Pam Mollaun, & Ching-Ni Hsieh, ETS*

Technology has the potential to support innovation in language assessment design in a variety of ways. Utilizing technology and design features to expand construct representation would be critical for true innovation (Chapelle, 2010), but there are other potential uses of design features that may benefit test-takers and enhance the quality of an argument for test use. In this presentation, we describe design features that were incorporated into a prototype computer-based assessment of academic speaking proficiency with the intention of improving the test-taking experience and TLU domain representation. A set of features related to the visual presentation of the task was introduced (e.g., videos for language input, visual support to complement the presentation of new academic content) in order to enhance representation of the multimodal TLU domain within assessment tasks. Another set of features enabled test-takers to have some control over the pacing of the assessment procedure and choice over the topic of discussion, in order to increase the test-taker’s sense of agency during the assessment.

To investigate the impact of these design features on test-takers’ experience of the assessment and perceptions about the authenticity and validity of the assessment, we administered a prototype version of this assessment alongside an assessment of academic speaking proficiency that lacked these features. After completing both assessments, test-takers completed a survey about their perceptions of the tasks and their features. Survey results are analyzed using mixed methods in order to explore test-taker perceptions of how task design features may positively (or negatively) impact their testing experience, and the potential implications for an argument for test use are discussed.

**Paper 4**  
**Development and Evaluation of Innovative Speaking Tasks: Increased Context to Evaluate Communicative Appropriateness**  
*Larry Davis, Pam Mollaun, Jonathan Schmidgall, Ching-Ni Hsieh, and Yuan Wang, ETS*

Large scale speaking assessments often feature tasks where the context and purpose of communication is underspecified, limiting the inferences that can be made regarding test takers’ ability to achieve a specific communicative goal in an appropriate way. Providing additional context can be challenging in large scale tests, however. On a conceptual level, aspects of appropriateness in communication must be identified that will support desired inferences yet not create unintended fairness issues for test takers from varied cultures; on a
practical level, materials and testing time are needed to establish the context for the task.

This presentation describes the development and evaluation of prototype academic speaking tasks for large scale assessment where the goal was to elicit evidence of the ability to communicate appropriately in a given situation. It will demonstrate a principled design process that includes conceptualizing the aspects of appropriateness intended to be measured, defining contextual and task design features to elicit evidence of the intended ability, identifying the evidence obtained in test taker responses, and iteratively testing assumptions and improving design features through multi-stage prototyping. Differences in appropriateness seen in test takers of differing proficiency levels will also be presented, along with test taker reactions to design features. As expected, what it meant to 'be appropriate' varied with task and context; how this variation was dealt with in scoring criteria will also be described. The presentation will close with a discussion of the prospects for use of more fully contextualized tasks in large scale speaking assessment.
1) Language Assessment Expert, Yes, But Are You a Police Interrogation Expert?

Margaret van Naerssen, Immaculata University

In legal cases involving non-native speakers (NNSs), attorneys might ask language assessment experts two questions. First, what is X's English language proficiency? Second, how might this proficiency have affected, for example, X's ability to participate effectively in a police interrogation? To connect assessment levels to the second question, an assessment expert should also at least have expertise in another relevant linguistics discipline, e.g., language acquisition research or sociolinguistics. (FL, "forensic linguist", refers to this expert.) This poster shows the challenge facing a dually-qualified FL: language assessment and sociolinguistics. The FL completed the language assessment and viewed police interrogation videos. Then the FL explained to the hiring attorney how sociolinguistics could help connect language proficiency to the interactions in the police interrogation. However, the attorney continued to think of the expert only as his language assessment expert. How were these conflicting perspectives resolved? As an expert consultant/witness, one must maintain objectivity as much as possible. Also, the hiring attorney expects the expert to adhere to specific constraints. First, stay within the scope of the assignment. Second, do not go outside the boundaries of one's areas of expertise. Such constraints may seem obvious, but once an expert becomes familiar with the language evidence, this can be challenging. If the expert goes outside the boundaries, even on a small point, the expert's credibility might be seriously questioned. The expert might be stopped from testifying and/or the expert's report might be thrown out. The worst case scenario is: The case might fall apart. However, what if the sociolinguistically qualified FL, sees patterns in the language evidence that raise issues beyond the attorney's view of the expert's assignment? Perhaps there are patterns reflecting aggressive interrogation strategies that might have been difficult for the NNS to handle. Should the FL offer to examine relevant interactions? In this case the attorney challenged the FL: "Are you a police interrogation expert?" (The FL responded, "Uh, no.") "You're our language assessment expert! If you're challenged, we could lose you as our expert! Besides, we have a police interrogation expert!" The FL counters "OK, I understand, but can THAT expert tie strategies to NNS language?" How were these different perspectives resolved? The FL, as sociolinguist, saw the proficiency interviews and police interrogation as interview contexts. The FL, as assessment expert, examined test administration instructions for the proficiency interviews used. The FL then identified six interview factors for organizing assessment practices and general interrogation strategies: (1) Goals, (2) Effect of Stress, (3) Objectivity, (4) Type of Encouragement, (5) Language Supplied to Interviewee, and (6) Topic Content Continuity. The FL concluded: Interacting under
such interrogation conditions would require the linguistic resources to handle cognitively demanding tasks when functioning in a second language. Using general context descriptions with supporting details, framed in language assessment terms, the FL stayed within the expert’s dual areas of expertise. This also allowed the FL to remain as objective as possible, not attacking a particular police officer and not claiming expertise in police interrogations.

Wednesday, June 22  
16:30-17:45  
Classroom Building Hall - upper level  
University of Palermo, Scuola Politecnica - Faculty of Engineering

## 2) Oral Composition and Mini-Debate: Do They Measure the Same Construct?

**Liu Xu, Guangdong University of Foreign Studies**

Many test validation efforts in language performance assessment have focused on the rater’s response to a single speaking or writing task (e.g., May, 2011; Kuiken & Vedder, 2014). As a result, the delineation between task effects and trait effects on the score is not always clear (McNamara, 1996). On the other hand, studies that involve task comparison have focused on differences in scores, performance discourse, and test process, but have largely neglected raters’ use of criteria categories (e.g., Bae & Bachman, 2010). The missing link here is a comparative view on the rater cognition when rating different tasks, which may shed new light on validity. In this test validation study, we compared the rater cognition involved in two different tasks in an EFL speaking test, the Test for English Majors, Band Four, Oral Test (TEM-4) in China. These tasks are a 3-minute narrative oral composition and a 4-minute peer-to-peer mini-debate, both computer-delivered and with 3-minute planning time. Nine TEM-4 oral test raters were invited to give scores on task completion for each task to two/three TEM-4 candidates. Each of the twelve candidates was scored by two raters. The raters gave reasons for their scores immediately after they scored one task. The oral reports were transcribed and coded to identify patterns in the raters’ references of the criteria categories prescribed in the rating scale and criteria not in the rating scale. These patterns were interpreted as the ways in which raters adapted to task performances and operationalized the construct to be measured. The major findings of this study are:

1) Attention of the raters was distributed in different patterns peculiar to each of the two tasks. On one hand, task requirements on topic relevance for oral composition were more frequently heeded than task requirements on stance relevance for mini-debate. And on the other, interaction commanded a large proportion of the raters’ attention when the mini-debate task was rated. 2) Although both content and language were paid major heed to, language was relatively a more prominent criterion for the rating of oral composition than mini-debate. Even by the same criterion, the raters meant different subcategories when rating different tasks. Most notably,
the content criterion for oral composition was operationalized as detailedness, specificity, logic, and relevance, whereas diversity, legitimacy, and elaboration of arguments, persuasiveness, and topic-development, were mentioned as the content subcategories for mini-debate. 3) The two tasks involved different types of construct-irrelevant rating processes. Time management was frequently mentioned by the raters in rating oral composition, while impact of interlocutor’s performance and assigned stances were involved in the rating of the test-takers’ performance in mini-debate. In sum, the comparison has found both overlaps and discrepancies in the constructs measured in the two different tasks. These similarities and differences will be interpreted in terms of construct representation and relevance (Messick, 1989). The implication is that the comparative view is conducive to a clearer distinction between the common construct measured by different tasks and the constructs specific to each task during the rating process.

Wednesday, June 22
16:30-17:45
Classroom Building Hall - upper level
University of Palermo, Scuola Politecnica - Faculty of Engineering

3) Assessment of the Speaking Proficiency and the Civic Competence within the Migration Context: a Case Study on a Large Scale Certification

Lorenzo Rocca, CVCL - University for Foreigners of Perugia

This contribution reports the preliminary results of a case study aimed to investigate how a large scale certificate of Italian L2 administered within the migration context assesses both speaking and civic competence (McNamara: 2009) according to the current Italian law. In 2011 the Italian Government introduced the so called “Accordo di Integrazione”: as a consequence, the law nowadays requires newcomers from non EU countries to satisfy the following two requisites. These requisites have to be assessed at the end of two years, after the migrants’ arrival in Italy. 1. Assessment of the speaking proficiency (A2 CEFR level) 2. Assessment of the KoS (Knowledge of Society), according to a list of compulsory contents included in the law (DPR 179/2011), that cover different areas related both to rights and duties within the Italian Constitution. According to these requirements the Ministry of Internal Affairs asked CVCL (Centre for Assessment and Language Certifications - University for Foreigners of Perugia) to standardize the speaking component of the A2 certificate for adult migrants (CELI 1 i) to address these legal requirements, introducing the contents of the list provided by Accordo. Therefore, the process of production of the speaking component acted on this request, in the following phases: 1) Preliminary considerations: how the KoS could be accessible for A2 users within the migration context 2) Definition of the new CELI 1 i speaking component: structure, tasks, assessment criteria scale 3) Trialling of the new format 4) Revision, according to the
results of the trialling 5) Administration of the live test (around 8,000 candidates involved) 6) Case study: collection of feedback and impact The contribution includes: • Detailed description of the examination structure: • Tasks description • Assessment scale • Videos of the exam • Methods used to collect feedback • Data analysis: first results • Conclusions Two possible implications: 1. The European scenario (Council of Europe Survey: 2014): from a corpus of 36 countries, 50% demand compulsory tests of language and civic competences for migrants (Van Avermaet: 2009): Italy is the only country that introduced the assessment of the KoS within the speaking component. 2. Accordingly, the hypothesis of revising the construct, as in the ALTE LAMI Booklet (2015), of the current tests for migration purposes. Many countries in Europe require the assessment of both language competence and knowledge of the civic norms and cultural routines of the host society, asking migrants to cope with situations where the language is necessary to comply with civic expectations. Thus, the purpose of the test is assess both these competences, taking into account the migrants’ repertoire and emphasizing their experiences through tasks focused on the public domain. If the tasks, in terms of contents and contexts proposed by the test, reflect the real-life ones (Bachman: 1990, 2007, 2010, Weir: 2005) and if the test is able to combine civic expectations with migrants’ everyday life, probably it will have more chance to be perceived as useful.

Wednesday, June 22
16:30-17:45
Classroom Building Hall - upper level
University of Palermo, Scuola Politecnica - Faculty of Engineering

4) The Interface of Construct, Contexts and Content in the Revision of a Reading Test

God Lim, Cambridge English Language Assessment
Mark Elliot, Cambridge English Language Assessment

This poster details the process by which the reading section of a large-scale test of English language proficiency was revised and validated. Changes to the content of the test were made following a review of the construct so as to enhance the test’s suitability for use in the context of entry into higher education. The construct given the target language use context was defined as ‘language proficiency for the purpose of entry into an English-medium university environment’. The definition makes clear that inferences are being made about language proficiency rather than academic preparedness, about one entering higher education as a novice rather than one who knows the specialized conventions of academic discourse, and about ability to function within the university environment rather than just the strictly academic elements of it. This definition implies a different set of abilities than the oft-used but imprecise notion of ‘academic English’, and therefore a different and better specified test construct. Guided by a socio-cognitive framework for test validation, a multi-phase exploratory sequential mixed methods design with nested parallel
convergent elements was used. In the first phase, it was identified that existing test tasks failed to cover one important construct-relevant level of Khalifa and Weir’s (2009) cognitive processing model of reading, namely creating intertextual representations, and draft specifications and prototypes for two task types targeting this were developed in response. In the trialling phase, expert judgment based on the literature and the socio-cognitive framework was combined with classical and Rasch statistics to arrive at an interpretation of trial test outcomes, and one task type selected for inclusion in the test and specifications finalized. In the validation phase, multiple examples of the selected task were produced and pretested to ensure reproducibility, the reading texts within the tasks were subjected to textual analysis using Coh-Metrix and Lextutor to determine comparability with texts encountered in the university environment, and the evidence from all phases combined to produce a validation argument ahead of the live use of the test. The poster highlights how, in the various phases of the revision process, the intersection of construct, context and content created affordances and limitations, determining the final design of the test. Consideration is also given to how changes in one of these parameters would change the resulting test, illuminating the role of each element. This poster is therefore of relevance and implications not just to the conference theme, but for those involved in the practice of designing and delivering language ability examinations.

---

**5) Using Proficiency Test Data to Set Program Goals: Interpretation, Speculation, and Collaboration**

*Daniel Reed, Michigan State University*

*Susan Gass, Michigan State University*

*Paula Winke, Michigan State University*

*Shinhye Lee, Michigan State University*

At our university, in spring of 2015, we administered 796 standardized, computer-based reading and listening proficiency tests to 1st, 2nd, 3rd, and 4th year students taking French (n=234), German (n=100), Italian (n=23), Portuguese (n=25), Russian (n=25), and Spanish (n=390). The goals of the test administration, which is part of a larger, grant-funded initiative to investigate proficiency-based learning at the university level in the United States, were (a) to better understand where our students are in terms of proficiency and (b) to help us set listening and reading goals across these language programs. The students in the programs took these exams in March and early April of 2015, approximately one month before the end of the semester. Standardized testing could not take place any later given that classroom obligations and achievement-exam schedules were already in place.
Thus, when trying to use the test outcomes to understand where students are in listening and reading at the end of the academic year (which is in the beginning of May), concerns arose as whether the scores and designated proficiency levels (i.e., Novice high, Intermediate mid) accurately indicated end-of-term outcomes. Thus, in collaborative sessions aimed at establishing realistic goals for each level in each language, the researchers and program directors had to interpret the existing data and speculate “what might have been.” To compensate for the continued month of learning, we wanted to identify students on the cusp of achieving the next level of proficiency, and tentatively label them as having achieved that level so that we could consider a likely more realistic scenario that gave those students the benefit of the doubt when describing end-of-year learning expectations. We wanted to define “on the cusp” as someone whose listening- or reading-test score was within one Standard Error of Measurement (SEM) below the cut-off for identification into the next-up level of proficiency. Unfortunately, there was no reported reliability for these commercially available, standardized tests. To obtain a perspective on the probable reliability estimates for each of the 8 tests, we first recorded the reliabilities of other standardized listening and reading tests similar to the ones we gave but which have reported reliability data. We also calculated the reliability of the tests we gave using KR-21, which only requires the number of items on the test, the mean score, and the variance of the whole test (standard deviation squared), knowing that KR-21 also assumes the items on the test are of approximately equal difficulties (we violated this assumption) but can provide a conservative estimate. We considered all reliabilities we had for each test format (listening and reading), and used those in the calculation for the SEM for each listening and reading test. In our poster we present this model of estimating reliability and setting goals, and we also present charts for Spanish showing (a) the proficiency levels obtained without SEM taken into consideration, and (b) the proficiency levels obtained (or expected to be obtained) taking SEM into consideration.

Wednesday, June 22
16:30-17:45
Classroom Building Hall - upper level
University of Palermo, Scuola Politecnica - Faculty of Engineering

6) Monitoring Assessment of Writing in Large Scale Examinations Context: a Tool to Improve Raters’ Self-Awareness

Danilo Rini, CVCL - Università per Stranieri Perugia

This contribution addresses the issue of monitoring rater’s behavior in assessment of the writing component of large-scale exams in Italian as L2. In 2009, systematic monitoring of writing raters was introduced by an Italian test provider to its examinations system. As a result of the application of the Association of Language Testers in Europe Quality Management System (Saville 2014, Grego Bolli 2014),
assessment of writing has been monitored through systematic data collection and statistically analysed since 2012. A more structured training system was introduced at the same time. Rater performance is analysed using Multi-Facet Rasch Measurement in order to provide evidence of raters’ behaviour across different sessions. This contribution presents the results of the statistical analyses over time. Excerpts from interviews with the raters about the usefulness of the training and the feedback they received, in terms of increased awareness and professional development, are also presented as a supplement to the statistical investigation into their performance. Writing assessment is carried out by a group of experienced teachers of Italian as L2, with a team leader being a permanent member of the staff. Over the years, the number of raters varied from 4 (in 2009) to 9 today. Before every session of exams, “older” raters undergo refreshment training, while “new” ones have a more structured training session led by the team leader. Exams are clerically marked against a four category rating scale. Raters work in close contact with each other and with the team leader, and weekly meetings are used to compare the assessment of ‘borderline’ sample papers. Their behaviour has been systematically monitored once a year since 2009, with the raters remaining the same for the last 3 years. Data reported in this paper suggests that two of the “older” raters, monitored since 2009, though appearing consistently lenient in their assessment (Knoch, 2011; Weigle, 1998) when compared to the standard, over the years tend to realign; on the other hand “new” raters, who joined the organisation in 2012, have been more aligned to the standard since the beginning. This contribution presents the findings of the statistical analyses over time and the interview data. In terms of the qualitative data, bearing in mind that raters are also teachers of Italian as L2, the interview data suggests that training in the assessment of subjective tasks can contribute to increased awareness of what contributes to candidates’ performance at different levels, and better use and understanding of tasks in the classroom (Parr & Timperley, 2010). Secondly, it discusses the analyses into rater’s consistency, coherence and severity or leniency, and the related issues to be faced in terms of management, training and recruiting of raters (Ackermann & Kennedy, 2010). It demonstrates how quantitative data analysis could be useful to inform decision making for positive impact and how this was managed by a test provider with limited resources. This contribution is of interest to those who need an overview of how simple statistic procedures and practical data collection can be easily applied to their own testing contexts.
Wednesday, June 22  
16:30-17:45  
Classroom Building Hall - upper level  
University of Palermo, Scuola Politecnica - Faculty of Engineering

7) Improving Quality in Classroom-Based Language Assessment: The Case of a Toolkit Alternative Assessment

Koen Van Gorp, Centre for Language and Education, University of Leuven  
Piet Van Avermaet, Centre for Diversity and Learning, UGent  
Fauve De Backer, Centre for Diversity and Learning, UGent

In 2010 the Flemish Minister of Education asked a committee of experts to conduct a feasibility study on the introduction of a standardized language test at the beginning of secondary education in order to (1) optimize school language assessment policies, (2) raise the language proficiency level of every student and especially, (3) identify students at risk of failing academically. The expert committee advised the Minister that a single standardized language test would not be able to fulfill these goals and that in fact a single test would send the wrong message to schools. It would focus the teachers’ attention on students’ language deficits and strengthen teachers’ belief in a remedial approach rather than have an impact on the overall classroom instruction (Struyf, 2004). The committee recommended instead to develop a toolkit of alternative assessment instruments that would allow school teams to provide a more complete picture of students’ Dutch language proficiency and to monitor the process of language learning from K-12 (Van Avermaet et al., 2011). Underlying this advice was the claim that developing alternative forms of assessment at school and classroom level would not only ensure a fairer form of assessment (Van Petegem & Vanhoof, 2002) but also would allow schools to develop a language assessment policy that takes into account the dynamic and complex language development of each individual learner (i.e., learning-oriented assessment – Purpura & Turner, 2013).

Commissioned by the Flemish Department of Education, a team of the University of Leuven and the University of Ghent developed two toolkits in 2012-2013: one for primary and one for secondary education. The starting point was to create a coherent framework bringing together all the language and language-supported competences (e.g., social, self-regulation) that students should develop throughout primary and secondary education and that are necessary to function academically at K-12. Secondly, all language assessment instruments (observation tools, tests, portfolios) available in Flemish education were inventoried and placed within the framework of language competences. Each instrument was evaluated and annotated following the VRIP parameters: validity, reliability, impact and practicality (Bachman & Palmer, 2010). Thirdly, scenarios were developed to help teachers implement these assessment instruments in a coherent language assessment policy. The scenarios provide teachers with information concerning all aspects related to assessing language learners in a broad and alternative way. Videos and training materials were added to support teacher trainers and pedagogic advisors in conveying this information to teachers and school teams, and training them in the use of these instruments. Finally, both Toolkits were made available free of
Poster Abstracts

Wednesday, June 22
16:30-17:45
Classroom Building Hall - upper level
University of Palermo, Scuola Politecnica - Faculty of Engineering

8) Testing for Differential Item Functioning with No Internal Matching Variable and Continuous Item Ratings

Michelle Y. Chen, The University of British Columbia / Paragon Testing Enterprises
Wendy Lam, Paragon Testing Enterprises
Bruno D. Zumbo, The University of British Columbia

Writing ability is usually measured through performance assessments in which test takers write an essay or other forms of written expression, in response to a prompt. Writing assessments like this typically only have two or at most three writing prompts (and hence writing samples). Test takers’ performance, however, may be affected by factors other than their writing ability; for example, types of task, and wording or topic of the prompt. All of these can be potential sources of differential item functioning (DIF) in a writing test. DIF occurs when test takers from different groups of the same ability levels have different chances of achieving the same score levels on a task. Many techniques and procedures have been developed to test for DIF (e.g., Rogers & Swaminathan, 1993; Zumbo, 1999); however, performance assessments, such as writing tests, often pose two unique challenges in DIF investigations: lack of an internal matching variable and a continuous rating score. It is difficult to find an appropriate internal matching variable for writing tests because of the small number of tasks within writing tests. Therefore, typical DIF methods that match on an internal total test score are not feasible. Another challenge is that ratings on writing assessments may be on a continuous metric whereas DIF methods are designed for binary or polytomous scores. Building on work by Zumbo (2008), a method to test DIF for a continuously scored writing test with only two prompts on each test form is proposed and demonstrated with real test data. Data from a high stakes English proficiency test is used for this study. The overall test has four components: listening, reading, writing and speaking. Our focus is on gender DIF for the writing prompts. All test takers respond to two different writing prompts. Prompt-level scores for each test taker are on a continuous metric...
ranging from 0 to 12.3, with over 260 possible scores. Forty-two writing test forms, which consist of 81 unique writing prompts were used in this study. Each prompt was answered by at least 120 test takers from each gender group. A total of 56 writing raters were involved in rating these writing samples, with each sample rated by two to three raters. A multiple matching strategy was applied. Given that the correlations among the four different components of the tests are fairly high (above 0.77), listening and reading scores were used as multiple matching variables to investigate writing prompt DIF. In light of the continuous prompt scores, linear regression models are used to test DIF. The full linear model is: Writing score = b0+b1(reading)+b2(listening)+b3(gender)+{interactions of group by matching}. Linear regression provides both a significant test and several effect size measures, which offer useful descriptions of uniform and non-uniform DIF effects. Three writing prompts out of 81 were flagged as DIF items with moderate magnitude. Details of the analysis procedures, decision rules to flag DIF items, and results from this DIF investigation will be presented in the final paper.

Wednesday, June 22
16:30-17:45
Classroom Building Hall - upper level
University of Palermo, Scuola Politecnica - Faculty of Engineering

9) **Effects of Changing TOEFL Cut-off Scores: The Impact of Raising the Bar**

Laura Decker, Brigham Young University
Troy Cox, Brigham Young University
Jesse Egbert, Brigham Young University

Universities feel that diversity plays an integral part of improving the quality of education and seek to have a balance between diversity of both students and ideas. When admitting students whose native language is not English, they need some assurance that the students will have the language to participate in university life as well as succeed academically. Many U.S. universities use test scores from the TOEFL to screen prospective students' language ability, and as it is a norm-referenced test, each university has the task determine a cut score that ensures that students who are admitted will have the highest probability of success. This study will evaluate the effect that changing the TOEFL admittance cut score has had on a large, private university. In 2010, the university changed raised its cut score based on feedback that the nonnative English speakers admitted to the university did not have the language skills to succeed. This study examines the effects of that change by answering the following questions. • What predictive value did the TOEFL (both overall and subcomponents) have on first year GPA? How did the change affect that?

• What effect did the change have on admission acceptance rates?
• What effect did the change have on first GPA's across different majors?
• What effect did the change have on ESL service courses?

Anonymized data of 9,837 students was collected from the admissions office (including TOEFL Scores, acceptance decision, native language, age, gender, first year GPA, etc.) over the course of ten years (2005-2015). This data will be analyzed to see the quantitative effect the 2010 change had on university life. Furthermore, interviews with faculty who regularly interact with non-native English speaking students will provide qualitative data and shed light on the potential effects from raising the bar.

Wednesday, June 22
16:30-17:45
Classroom Building Hall - upper level
University of Palermo, Scuola Politecnica - Faculty of Engineering

10) Validating a Scale Revision Project: Results and Implications for the Underlying Model of Validation

Jamie Dunlea, Assessment Research Group, British Council
Judith Fairbairn, Assessment Research Group, British Council
Barry O’Sullivan, Assessment Research Group, British Council

This study reports on a project to revise the rating scales used in an operational, computer-based EFL proficiency test. The project integrated the collection of qualitative and quantitative data in a series of iterative steps that overlapped with expanding concentric circles encompassing gradually wider stakeholder groups. The process began with a core group of internal test development team members. Anecdotal evidence had indicated some potential problematic areas for raters in applying the scales. A retrospective review of both operational and research project data was carried out to establish any empirical support for these claims. Draft revised descriptors were then developed and trialled with a small group of experienced raters. Trialling included both the collection of questionnaire data and verbal feedback in a face-to-face group discussion. After a further round of revisions, a large-scale field trial was carried out in which previously scored performances were re-rated by all accredited raters using the new scales. Multi-facet Rasch measurement analysis of the rating data using FACETS indicated a high degree of consistency in the rank ordering of performances by raters, with little misfit and a low range of severity. The results, in conjunction with questionnaire data on the interpretability of the new scales, provided support for their introduction in operational use. The study highlights the importance of integrating multiple sources of evidence in the design and validation of rating scales. The model employed here focused on more qualitative feedback in the early stages. As the circles expanded to take in wider numbers of participants, the focus changed to a more quantitative data collection and analysis procedure. From a different perspective, the project also provides some critical insight into the application of the socio-cognitive model of
language testing development and validation, which underpins the testing program. The study provides support for O’Sullivan’s (2015) recommendation to reduce the focus on the temporal distinctions made in the SC model as it was originally conceived in Weir (2005) and O’Sullivan and Weir (2011). The temporal distinction associated some kinds of validity evidence (e.g. contextual and cognitive parameters) as a priori evidence and other kinds (e.g. scoring, criterion-referenced, consequential aspects) as a posteriori evidence. However, all forms of evidence may be relevant throughout the a priori development stage and will continue to be of use in the a posteriori, post-operational phase. Indeed, in this case, as the scale revision was being carried out on an already operational assessment, operational (i.e. a posteriori) scoring data was reviewed at the very beginning of the project. Throughout the project, both quantitative and qualitative data was collected. If a distinction were to be made, the distinction proposed by Kane of viewing evidence, and validity arguments based on such evidence, in terms of developmental and appraisal stages may be more useful. All forms of evidence associated with the categories used in the validation model may be applicable in both stages, but identifying whether a validity argument, and the evidence supporting that argument, is essentially a developmental- or an appraisal-stage argument may be useful.

Wednesday, June 22
16:30-17:45
Classroom Building Hall - upper level
University of Palermo, Scuola Politecnica - Faculty of Engineering

11) The Effects of Test Re-Dos on Learner Performance and Development

Kristen Sullivan, Shimonoseki City University

A test re-do is a test in which the learner is given the opportunity to literally “re-do” the test after conducting reflection and, if possible, receiving feedback. It is theoretically based upon ideas from self-regulated learning and learning-oriented assessment, and the primary aim is to support and encourage learner development in terms of language proficiency development and the development of learners’ ability to take control of their learning. Teachers may be concerned about the time involved in conducting test re-dos, or with issues of validity. In order to encourage the wider use of test re-dos, it is thus important to show how they can contribute to improved outcomes. This poster will investigate this through an examination of test re-do results for an Independent Speaking test task conducted within an optional TOEFL preparation course at a Japanese university. Results from 22 non-English majors of slightly varying proficiency levels and learning backgrounds will inform the analysis. Three aspects of learner performance will be considered: improvements in learners’ test performance (comparing the original performance with performance on the test re-do), connections between learner reflections and their performance on the test re-do, and the content of learner reflections. Analysis suggests that 16 of the
22 students showed improvement of varying degrees in their test re-do performance. These students could be further broken down into two groups: students whose performances were closely connected to their post in-class test reflections, and students whose performances were only weakly or not at all connected to the content of their reflections. The significance of this for the stated purposes of test re-dos, potential reasons for why this occurs, and possible consequences for the grading of test re-dos will be discussed. The remaining six students did change some aspect of their performance, but this change did not lead to an improved test performance; indeed in some cases the quality of the performance dropped. For both groups of students it is clear that identifying and articulating problematic areas of performance is a skill that needs to be developed. Particularly for the latter group, knowledge of strategies for improving performance is an area, which also needs to be addressed. The results strongly suggest that there is a limitation to what learners can achieve by themselves when re-doing their test performance, and that the role of teacher feedback here cannot be ignored. How and when this feedback should be provided needs to be further considered. This finding could be used to imply that test re-dos do not achieve their stated aim. However, the presenter will argue that the notion of improved performance needs to be considered more critically in the case of alternative assessment tasks, such as the one in question, where long-term learner development is the underlying aim. This raises various issues concerning the grading of student performance as well as the evaluation of the utility of the test task in general, and it is hoped that visitors to the poster will be able to engage in a discussion on this matter.

### Comparability of the GEPT – Advanced and the iBT: Content and Constructs

Antony Kunnan, Guangdong University of Foreign Studies  
Nathan Carr, California State University, Fullerton

The study presented in this poster investigated the comparability of two large-scale English language proficiency tests: the General English Proficiency Test - Advanced (GEPT-A) and the Internet-Based Test of English as a Foreign Language (iBT). These two tests ostensibly assess similar constructs, but are used in different contexts and for different purposes. Data was collected from test takers in both Taiwan and the United States. The instruments used were item-level participant test performance response data from the GEPT-A reading section; scores on GEPT-A Writing Task 1; iBT reading, writing, speaking, and listening scaled scores; participant responses to a background information survey and five questions involving their perceptions...
of the GEPT-A; the items and passages in the GEPT-A form; and sample iBT reading passages published by Educational Testing Service, Princeton, for test preparation purposes. Two specific analyses were conducted: First, a content and task analysis was performed on the GEPT-A and iBT reading passages. This analysis included the cohesion, syntax, and vocabulary used in passages, as well as a comparison of the construct coverage, scope, and task formats used in the reading comprehension questions on the two tests. Second, an analysis of the participant responses on the tests from the two tests was also conducted. The results of the content analysis showed the reading passages on the two tests are comparable in many ways but differ in several key regards. The task analysis revealed that the construct coverage, item scope, and task formats of the two tests are clearly distinct. In particular, there were construct coverage differences in the areas of assessing vocabulary knowledge, vocabulary in context, reading for the main idea, reading for major points, inferencing, skimming, and scanning. The reading sections of the two tests also differed markedly in terms of item scope and task format. Analysis of participant test responses indicated that the GEPT-A has good reliability, and that reading comprehension items tend to function quite well. Scores on the GEPT-A and iBT are highly inter-correlated with each other: More specifically, scores on the two reading tests had a medium-to-large correlation ($r = .467$), and scores on the GEPT-A Writing Task 1 and iBT writing had a medium-sized correlation ($r = .385$). The GEPT-A may be somewhat more difficult than the iBT, given that the mean GEPT-A reading score was 57.9% of the total points possible, vs. 82.9% for the iBT reading section, and 51.1% for GEPT-A Writing Task 1 vs. 80.1% for the iBT writing section; however, these comparisons across differently scaled scores must be interpreted with caution. Exploratory and confirmatory factor analyses of the test score data indicated that the two tests appeared to both be measuring reading and writing ability, but emphasize different aspects of the reading construct—that is, the different construct definitions for the two tests are reflected in the results of the factor analyses. Results on the two tests are therefore not entirely comparable.

Wednesday, June 22
16:30-17:45
Classroom Building Hall - upper level
University of Palermo, Scuola Politecnica - Faculty of Engineering

13) Developing an Electronic Pre-Test for the EPPLE Examination: Test Design, Technical Challenges and Mobile Technology

Douglas Altamiro Consolo, UNESP - State University of São Paulo
Debora Mieko, UNESP - State University of São Paulo

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

Wednesday, June 22
16:30-17:45
Classroom Building Hall - upper level
University of Palermo, Scuola Politecnica - Faculty of Engineering

14) Measuring Progress in English for Tourism in a Hungarian University Context

Zoltán Lukácsi, Euro Examinations

English language proficiency and communication skills may be important graduate attributes for all undergraduates (Humphreys et al, 2012), yet students are not tested in foreign languages on entry at the Budapest Business School. However, their language proficiency is assessed on completion of the 168-contact-lesson course, for local academic requirements dictate that they must obtain a language certificate in ESP at level B2 if they are to graduate successfully as economists. Standardized instruction also prescribes the course material and the pace of progress regardless of student achievement, rate of learning, or language needs. As language instructors are held responsible for student failure, the need for a standardized and reliable but freely available measure became increasingly evident. Working in collaboration with a state-accredited language examination office, the departmental team designed a CEFR-related exam of objectively scored components. The detailed specifications were constructed following a needs analysis and the evaluation of local resources. By monitoring progress in ESP, stakeholders could gain insight into whether the curriculum effectively guaranteed the realization of targeted learning outcomes. The poster presents the results from a longitudinal study of 319 university undergraduates. They were tested at the beginning and end of each of the three semesters of language instruction in Listening, Reading, and Use of English. The study implemented a common-item nonequivalent groups design,
where the six administrations were linked through the item bank of the state-accredited examination. The item responses were processed in OPLM (Verhelst, Glas, & Verstralen, 1995). As a consequence, the person parameter estimates were directly comparable among students and through time, as well. The results revealed that (a) instruction served the function of level maintenance for high achievers, (b) low achievers failed to meet academic requirements, and (c) strict adherence to the rigid educational framework proved counter-productive and posed a threat to the quality of language education.

References

Wednesday, June 22
16:30-17:45
Classroom Building Hall - upper level
University of Palermo, Scuola Politecnica - Faculty of Engineering

15) Creating a Computer Delivered, Multimedia Listening Test: Challenges, Choices and the AUA

May Tan, Canadian Defense Academy
Nancy Powers, Canadian Defense Academy
Roderick Broeker, Canadian Defense Academy
Ruth Lucas, Canadian Defense Academy

This poster focuses on the development of a new, computer delivered listening proficiency test for NATO military personnel. This new multilevel test has items that meet the specification of the NATO Standardization Agreement 6001 (STANAG 6001) language proficiency descriptors levels 1 to 3. This new test departs from the previous format used to present listening items in that it uses a multimedia format that allows audio and video input to be presented to candidates. As part of the development process, teachers and students were presented with a demo test that allowed them to interact with various types of items on this multimedia platform. These participants provided valuable feedback, which was incorporated into the design of the test, via a verbal feedback protocol. Their impressions of how they perceived the older version of the test versus the new computer-delivered multimedia version is presented. Their reactions to audio-only versus video listening tasks within the new test are also presented. Also discussed are the challenges that the developers faced when creating items in order to meet the criteria of authenticity and interactivity since these are important relative to construct validity (Bachman & Palmer, 1996; Ockey, 2007). It also presents the choices that were made concerning the issue of user control as well as the selection of the types of audio and video
listening tasks that are representative of the target language use domain (Messick, 1996; Wagner, 2014). The poster also explains the validation procedures selected for this test. This posed a challenge because the Benchmark Advisory Test (BAT) is the recommended test against which other listening tests created using the STANAG 6001 should be validated. However, due to the fact that it differed greatly from this benchmark test (the BAT uses audio only in its test tasks and constrains the number of times candidates get to listen to a text: the higher the level of difficulty, the fewer the number of replays allowed), the developers have also had to come up with a novel way of validating this new listening test that includes building an Assessment Use Argument and triangulating different sources of evidence in order to build a validity argument.

Wednesday, June 22
16:30-17:45
Classroom Building Hall - upper level
University of Palermo, Scuola Politecnica - Faculty of Engineering

16) OLA (Offerta Linguistica di Ateneo) at UniCal: Designing and Implementing a New Test Format

Carmen Argondizzo, Università della Calabria
Jean Jimenez, Università della Calabria
Ian Robinson, Università della Calabria

Most undergraduate programs at Italian universities include EFL courses, the contents and levels of which are often decided on by each individual department. This has meant that there is not only a great variety in the syllabi of these courses and the number of credits awarded, but also a difference in how language competence is assessed. In fact, test developing and test writing at universities often occur on a small scale, with teachers in different departments working autonomously to either select or develop their end-of-course achievement tests. This creates a non homogeneous language course offer within a single institution and makes it difficult to determine what exactly is meant by, for example, an ‘English One’ course. With recent university reforms and calls for greater transparency as well as the need to facilitate student mobility, this situation is starting to change. In light of this, as part of the Offerta Linguistica di Ateneo* (OLA), the Language Center at the University of Calabria (UniCal) set out to organize an English for Basic Academic Skills Course which would be suitable for students across all departments at the university. The aim was that of standardizing the ‘English One’ courses and methods of evaluation, while preparing students for the more specific language courses that they are then required to attend as part of their degree course. It was thus necessary to develop a new test format to be used as an achievement test, the level of which was set at a B1 lower CEFR level to realistically reflect the context in which the university is situated. This poster illustrates the phases involved in the development of this new
test, highlighting the rationale behind it, the practices that were implemented to meet the needs of the various stakeholders involved, and the work carried out by the test development team to guarantee validity and reliability. We will focus more specifically on how the team developed its own practice in order to design and then write the tests – with checks and balances to assure quality. The test construct, content and format will be outlined along with the design of the software, which is used to administer the test. It is hoped that this poster will encourage discussion on the merits of this style of university wide testing and the manner in which the test has been constructed and implemented. *University Language Course Offer

Wednesday, June 22
16:30-17:45
Classroom Building Hall - upper level
University of Palermo, Scuola Politecnica - Faculty of Engineering

17) Teacher Assessment Literacy Enhancement: Identifying Contextual Factors

Dina Tsagari, University of Cyprus
Tony Green, University of Bedfordshire
Karin Vogt, University of Heidelberg
Ildiko Csepes, University of Debrecen
Nicos Sifakis, Hellenic Open University

Research has shown that in many educational systems across Europe, English Language Teachers (ELTs) are not well prepared to create quality assessment materials and procedures (Inbar-Lourie, 2013; Green, 2013; Csepes, 2013; Vogt & Tsagari, 2014; Tsagari, 2013). This is at least partly because they are not sufficiently trained in the area of Language Testing and Assessment (LTA). There is therefore an urgent need to develop an efficient, relevant and sustainable LTA training infrastructure for ELTs to help them develop relevant assessment literacy, which can eventually be of benefit to other language teachers as well. This poster will present the aims of a three-year long project, entitled ‘Teachers’ Assessment Literacy Enhancement (TALE). This involves a network of five experts from different European countries and aims to contribute towards: - the development of innovative training materials and services that will be delivered through traditional face-to-face and online learning systems and offered through synchronous and asynchronous modes. The proposed training ecosystem is expected to offer continuous support and mentoring to teachers in the respective countries. - the expansion and the exchange of LTA expertise between European educational contexts in a creative and innovative way – the cooperation between and within different disciplines and various sectors of training in order to foster efficient and meaningful assessments suitable for language learners in primary and secondary education. The first phase of the project involves a needs analysis, which will be achieved through extensive consultation with ELTs and their students to ensure that the training course will meet...
their needs. This study adopts a sequential explanatory mixed-method design with a quantitative and a qualitative phase. In the quantitative stage, a questionnaire will be designed to measure how teachers and their students perceive assessment, the types of assessment used, the impact of these assessments and suggestions for improvement. The survey will be conducted among approximately 250 EFL teachers and students in the various European contexts identified in this project. Descriptive statistics, correlation, and structural equation modeling will be used to analyze the collected data to examine the trends identified per country and also interrelationships between the needs of ELTs and their students in the various contexts involved in this project. In the qualitative stage, individual interviews and focus groups with students sampled from the quantitative phase will be conducted to complement the relationships detected in the quantitative stage. A standard thematic coding process will be used and the co-occurrence of the codes will be identified. The findings may provide nuanced in-depth understanding of the assessment needs of ELTs and their students which may contribute to the identification of assessment priorities and the development of assessment training strategies that are contextually situated. The results of this phase of the project will be used as input in the next phase. With this presentation we aim at sharing the results of the first phase and hope to receive constructive feedback and advice from the conference audience especially those with an interest in the field teacher training and development.

Wednesday, June 22
16:30-17:45
Classroom Building Hall - upper level
University of Palermo, Scuola Politecnica - Faculty of Engineering

18) **Bridging the Gap between Research and Practice: Reflections from Constructing a Cognitive Diagnostic Reading Assessment**

*Junli Wei, University of Illinois at Urbana-Champaign*

Assessment should be made more useful for promoting student learning success (Gordon commission, 2013). Although recent advances in cognitive diagnostic assessment (CDA) research move in this direction, most CDA studies of second or foreign language (SFL) reading tend to identify examinees’ mastery of reading subskills by using existing large-scale English proficiency tests (e.g., Buck, Tatsuoka, & Kostin, 1997; Kasai, 1997). This might help researchers to gain knowledge and experience about CDA as a technique, but it has led to inaccurate and/or unsatisfactory diagnostic inferences, as many reading subskills have been rarely or never measured in the tests that were not originally designed for CDA purposes (e.g., Alderson, 2015; Jang, 2009). By constructing a CDA, this study aims to diagnose the strengths and weaknesses of students’ reading comprehension, and thus contributes to enhance instruction and learning. Over 1,300 senior high school
students were recruited from a non-native English speaking country. Using Evidence-centered Design (Mislevy & Yin, 2012), this study integrated a cognitive assessment framework from the outset into the whole test development primarily concerning three key issues: what to diagnose, how to diagnose, and how to use the diagnostic information. An integrated mixed methods research design (Greene, 2007) was developed over the following three phases to address the issues. The first and most important phase in this study was to identify subskills and specify their relationships. To clearly understand the cognitive processes measured within the specific group of examinees, a thorough analysis was conducted iteratively through literature review, think-aloud, content experts’ suggestions, and assisted by eye-tracking technology with a subsequent stimulated recall interview. Eight subskills with hierarchical relationships were finally identified. Second, test specifications including Q matrix were developed, providing a generative explanation for the creation of test tasks and the specific requirements of test construction for CDA. Thirty multiple-choice test items were developed and refined iteratively through pilot tests. After comparing the model-data fits (the hierarchy consistency index, HCI) and classification accuracy of different cognitive diagnostic models, the R-RUM model was chosen (HCI=.7) and programming codes were written to analyze the attribute hierarchy. Finally, only 25 items with good diagnostic power were maintained. Third, quantitative and qualitative evidence was provided to support test inferences. Quantitatively, statistical evidence (e.g., item difficulty ranging from .4 to .7, item discrimination >.7, subskill-specific item discrimination index >.6) from classical test theory (CTT), item response theories (IRT), and CDA respectively demonstrated test validity of the proposed CDA. Qualitatively, test-users’ evaluations of the diagnostic profiles, and their satisfaction with the CDA test design were analyzed. The integrated multiple sources showed an overall validity argument. Situated at the intersection of theories of L2 reading, cognition, and measurement, this study narrates a validation process for developing a diagnostic reading test. Thus, it strengthens validity arguments and enhances an understanding of the complexity of CDA test construction. As one of the first to develop a diagnostic reading test from the outset, this study is unique in incorporating the hierarchical structures of the reading subskills in test design and validation.

Wednesday, June 22
16:30-17:45
Classroom Building Hall - upper level
University of Palermo, Scuola Politecnica - Faculty of Engineering

19) Adding an Additional Oral English Language Assessment after University Admission

Erik Voss, Northeastern University
Annie Hsu, Northeastern University

International students take many English language exams throughout their academic career. Scores from these exams are used for many purposes; one of which is
admission to an institution of higher education. Part of the admission process for incoming direct-admit international undergraduate students to Northeastern University is an acceptable score on a standardized English language assessment. Any student who submits TOEFL or IELTS scores below the admissions criteria are required to complete an NEU English Language Assessment (NEU ELA). The NEU ELA is an additional oral language screen to evaluate speaking ability of incoming international students. This oral language assessment has been designed to elicit speech samples from students in a more relaxed setting than a high-stakes test. Furthermore, the prompts are localized focusing on topics related to academic life relevant to Northeastern that are not included in high-stakes test tasks. The test construct includes the ability speak in response to a prompt about academic situations, email etiquette and a co-op simulation situation. The NEU ELA has been delivered in 2013, 2014, and 2015. A total of 153 students took the assessment in 2015. The rating scale for the assessment is based on level descriptions for speaking ability from the Common European Framework of Reference. Correspondence tables were used to equate CEF levels with iBT scores. Incoming students usually place between a low B2 to a C1 level. The test is delivered through a third party interview website. Students record responses to the prompts using their webcam, which are scored by two independent (blind) raters with a third rating in the event of a disagreement in the score. These final scores are used to place a student into one of three levels: Level 1 Course: Recommendation to place student in English for Academic Communication course; Level 2 Resource: Exempt from course with recommendation for additional language support throughout the term; Level 3 Exempt: Exempt from course. The English course was designed to engage students in language used in the classroom, with advisors and instructors, and interviewing for co-op positions. At the end of the course, students are rated again by two raters to determine gains in language ability. Our poster will describe and evaluate the usefulness of the scores from the oral English language assessment to determine if students should receive supplemental English language instruction during their first semester of study. Data were examined from students placed into the course showing language gains, grades, and student course evaluations. Results will show the placement of students who completed the exam and try to draw conclusions about necessity of an additional speaking assessment for incoming international first-year students. Results are relevant to the usefulness of an additional oral language assessment in addition to the scores from standardized tests that students submit to enter the university.
Test Development at the Intersection of Context, Construct, and Content

Yasuko Okabe, CEES
Catherine Pulupa, University of Maryland / Second Language Testing, Inc.
Yuko Kashimada, Benesse Corporation

The subject of this poster is the development of the passages and test items for the listening and reading subtests of a high-stakes, four-skills, multiple-choice test, used to qualify students for admission to university in an EFL setting. The test is based on the principles of task-based language teaching (TBLT) and is aligned to the Common European Framework of Reference (CEFR). This poster will show how the test is designed at the point of interaction among construct, content, and context, as defined for this test. The overarching context for this test is the setting of university life: all passages and items must be plausibly situated in a university setting, whether social or academic (or on the continuum between those ends). During the initial writing of the test specifications, all item types were linked to particular can-do statements from the CEFR to create an unbroken chain between the construct being assessed and the assessment tool. For lower-level items (A2-B1), the first stage in the item development process is the "proposal stage," which ensures that item content is appropriately aligned to test specifications. At the proposal stage, item writers submit their idea for a passage-item set, and the proposal is reviewed for its appropriateness to the target context. Item proposals that do not meet this criterion are rejected. The test items must also be tied to the CEFR: writers specify not only the targeted item type but also the can-do statement related to the task that the item requires the test-taker to do. The proposal will also be rejected if it does not reflect a CEFR-aligned task appropriate to the context, thus ensuring that items are developed with respect to both the specified context and construct, as defined by the CEFR.

The development process for higher-level test content (B2-C1) is distinct and further illustrates the test's alignment of content to the target context and construct. While language learners at the A2 and B1 levels are able to interact in a variety of social situations, their capabilities in more "purely" academic settings are more limited. At the B2 and C1 levels, however, language learners are increasingly able to process more advanced content, and therefore the content presented on the test must reflect this. At the upper levels, passages are written by subject matter experts (SMEs) in the various academic fields appropriate to the testing population, whereas at the lower levels, passages and items are produced entirely by professional item writers. SMEs also record listening passages themselves, in order to maintain the voice of authority they possess when speaking about their field of expertise. They specify the content that they would expect someone at the level of a college freshman to glean from
the passages. Professional item writers then produce items in accordance with the SMEs' recommendations for important and salient targeted content. The poster will outline the test development process described above, emphasizing the distinctions between developing lower-level and higher-level content and designing test tasks to specifications.

Wednesday, June 22
16:30-17:45
Classroom Building Hall - upper level
University of Palermo, Scuola Politecnica - Faculty of Engineering

21) **Study on Comparability of Language Testing in Europe**

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

*Esther Gutierrez Eugenio, Cambridge English Language Assessment, University of Cambridge*

Following the “Conclusions on Multilingualism and the Development of Language Competences”, adopted in May 2014, the Council of the European Union invited the European Commission to explore the feasibility of assessing language competences across all the Member States by making use of existing national language tests. Cambridge English Language Assessment was commissioned to complete this task in what is officially known as the "Study on comparability of language testing in Europe". This study looked at 133 national language examinations at ISCED 2 and ISCED 3 levels (secondary education, approximately ages 14-15 and 17-18) from the 33 independent educational jurisdictions existing in the 28 EU Member States. The languages included were EU official languages other than the main language of instruction which were used in at least one other Member State, and which are studied by more than 10% of the students in secondary education in each jurisdiction. This study had two phases of data collection. The first phase regarded the collection of exam materials and supporting documents, which were obtained mainly from publicly available sources, in many cases with the help of in-country experts working for the national language assessment boards. The second phase of data collection involved the extraction of the actual data that would then be analysed and presented in the results. For this phase, the study adopted a mixed methods approach, which included the analysis of qualitative and quantitative data. The qualitative data was collected thanks to an online content analysis tool, which a group of experts in language assessment from throughout Europe were trained to use. The tool focused on elements regarding constructs, populations, interpretation of results and measurement characteristics. Additionally, quantitative data was also collected through an online comparative judgement exercise to determine the CEFR level of difficulty of samples of reading and writing tasks from all examinations included in the study. It emerged from the analysis of the data that the extent to which results of national language examinations can be compared depends on a number of
factors. First of all, comparisons of national results are only feasible when the data being compared have sufficient elements in common, both in the content and the format in which these results are expressed. However, and most importantly, this study has shown that language examinations across jurisdictions present a wide variety of features in terms of the constructs tested, the populations of test takers, the interpretations of the results, and the measurement characteristics of these examinations. These features importantly determine test quality, and in turn impact on the validity and reliability of the results obtained. The meaningful comparability of national results of language examinations across EU Member States will therefore depend not only on these results being expressed in a uniform format, but also on implementing measures at both national and European level that would increase the quality of current language examinations, and in turn ensure results are similarly valid and reliable across all jurisdictions.

Wednesday, June 22
16:30-17:45
Classroom Building Hall - upper level
University of Palermo, Scuola Politecnica - Faculty of Engineering

22) Learning-Oriented Assessment in and out of Class: Text and Context at Work

Angela Fenara, University for Foreigners of Siena

“Theorization” of classroom-based assessment (CBA) although perceived as long overdue (Davison & Leung 2009) does not seem to have found its own framework as yet (Turner 2012; 2014). Undoubtedly teaching, learning and assessment are so deeply intertwined to make investigation and documentation arduous, even more so in a second language acquisition context given the multi-layered nature of teacher-student interaction and the multiple types of information the teacher must be able to process. CBA, in short, is intrinsically reflective and cannot be easily psychometrically measured as is the case of large-scale testing. As Fulcher and Davidson (2007) point out though it might not be advisable to adopt paradigms we are familiar with and might need to look to validity and reliability as not norm-referenced when referring to CBA: upon assessment decisions are taken to aid students’ learning, the results of these decisions alone will prove the validity of the initial assessment (Moss 1994). CBA is not therefore a discrete activity which acts upon objective tests but being bound up with the teaching-learning process the teacher must make the most of the social context, the classroom, and design tasks whose content will enable students to positively engage and interact with the environment as a whole whilst at the same time provide a rich array of data relative to their linguistic communicative competence. Any theoretical framework, therefore, should take all of this into account to be able to develop a whole new set of relevant conceptual terminology to support teachers, training providers and underpin pedagogical materials.
This presentation reports on an attempt to use in university Italian foreign language classes tasks which make text and context fully interdependent thus extremely relevant both for the language learner and the assessment process, being the context construct-relevant in CBA (Fulcher & Davidson 2007), in an effort to map out, from a teacher’s perspective, the complexities underlying internal assessment which might provide useful insights albeit, admittedly, limited in scope. The tasks, linked to the CEFR and mindful of the tension between CBA and summative assessment and the consequent need for coherence, are interconnected and are designed so as to take into account the evolving research agenda concerning classroom assessment (Moss 1996; Black & Wiliam 1998; Purpura 2004; Fulcher & Davidson 2007; Lantolf & Poehner 2011; Turner 2012) and focus in particular on the theoretical and empirical studies on LOA (Carless 2007; Purpura & Turner 2014; TCCRISLS 2014). Internal assessment is a process in which teachers have access to both declarative and procedural knowledge, which enable them to consider and analyse a wide range of qualitative and quantitative data elicited from task–student interaction which must be interpreted and fed through the teaching-learning cycle. The presentation is an opportunity to conceptualise such process which is at the core of successful teaching but still in need of a coherent, consensually accepted paradigm.

Wednesday, June 22
16:30-17:45
Classroom Building Hall - upper level
University of Palermo, Scuola Politecnica - Faculty of Engineering

23) Developing a Speaking Rubric for the Vietnamese Standardized Test of English Proficiency

Huu Nguyen, Vietnam National University
Hoa Nguyen, Vietnam National University
Tuan Huynh, Vietnam National University
Phuong Tran, Vietnam National University
Ha Do, Vietnam National University
Thao Nguyen, Vietnam National University
Thuy Pham, Vietnam National University
Duyen Can, Vietnam National University
Anh Nguyen, Vietnam National University
Fred Davidson, Illinois University

Scoring rubrics are integral components of a test toolkit because they tell how stakeholders understand test scores and how raters evaluate test-takers’ performance. This poster presents the development of the scoring rubric for the speaking sub-test of the Vietnamese Standardized Test of English Proficiency (VSTEP), the Vietnam’s official four-skill English test which aims to measure learners’ English language proficiency from B1 to C1 levels, the expected levels for
post-secondary English learners in Vietnam. The VSTEP is an important outcome of the National Foreign Language 2020 Project of Vietnam which aims to enhance the English foreign language proficiency among Vietnamese people to prepare for the country’s economic growth and regional integration. Theoretically, the VSTEP speaking rubric was developed applying the criterion-referenced approach where oral performance of the test-takers was evaluated with respect to the CEFR-VN, the adapted version of the CEFR to suit the Vietnamese context. The constructs of the VSTEP speaking sub-test were investigated in the formation of the features of oral performance of the test-takers. To develop the descriptors of each level of performance corresponding to each feature, the CEFR-VN was applied. Empirically, the speaking performance of hundreds of test-takers was recorded to provide data for analysis. The results of the empirical data analysis helped improve the descriptors of the rubric, facilitated the rater training process, and supported in assuring the qualities of the speaking sub-test of the VSTEP. Since March 2015, the VSTEP test and its scoring rubrics have been applied in various test batches and have tested thousands of test-takers in Vietnam, the results of which have provided substantial information for the VSTEP development team to further review the VSTEP test and its components including the speaking rubric.
Investigating the Prompt Effect on Test Scores and Score Interpretations of a Group Oral Test

Zhouyang Lu, Zhejiang University

Performance-based second-language oral tests have become increasingly prevalent in recent decades, due to the enhanced interpretability of test scores, greater theoretical and construct validity (Kane, et al., 1999) and positive washback of such assessment tools (Bonk & Ockey, 2003). The group oral test, in which candidates are tested in groups but observed and assessed as individuals, has gained increasing popularity. Despite a body of studies conducted to examine its validity, a limited amount of research has systematically investigated under what circumstances and to what extent different prompts of the same task can be regarded equivalent. Given the importance of the comparable treatment received by every individual test-taker, which raises concerns not only about validity and reliability, but more importantly, about fairness, the effect of prompt variation requires close scrutiny. The present study attempts to investigate the validity of a group oral test by examining whether prompts of the same discussion task, have a differential effect in consistency of the scores and impartiality of the score interpretation (Bachman & Palmer, 2010). The setting for this study is a subtest of a newly developed large-scale English proficiency test in one major university in China, which is used as an exit requirement for undergraduates. Thirteen prompts in total, which slightly differ in length, topic domain, and how personal a response is invited, were used in rotation and got an approximately equal chance of being assigned. The Data comes from 1249 test-takers and 20 raters. A many-facet Rasch analysis is conducted utilizing FACETS 3.71 (Linacre, 2014) to research the comparability of the prompt difficulty and the interaction between prompts and raters, as well as the test-takers of different subgroups, i.e., test-takers of different genders, majors and English proficiency. A discourse analysis of the transcribed oral responses is further conducted to explore whether levels of accuracy, fluency and complexity differ in responses elicited by different prompts. The completed Rasch analysis suggests that 13 prompts, designed to be equal, virtually display eight levels of difficulty and can exert an undue impact on scores assigned on three categories of the rating scale which nevertheless can be reduced to an ignorable extent with the exclusion of one outlier and another two easiest prompts. While the bias analysis between prompts and raters indicates an interactional impact evidenced by the fact that four raters separately exhibit consistently opposite rating behaviors on two prompts, the bias analysis between prompts and test-takers’ background characteristics shows that test-takers’ genders, majors and English proficiency do not cause any differential prompt difficulty. The discourse analysis being processed aims to investigate whether test-takers’ responses are unfairly affected by the randomly assigned prompts, which
can therefore lead to spurious scores. The results of the study will provide validity evidence of the test by systematically revealing the prompt effect in a range of aspects and shed some light on prompt design and test development of L2 group orals, with the intention to reduce the potential threats to the validity, reliability and fairness.

Thursday, June 23  
17:05-17:50  
Classroom Building Hall - upper level  
University of Palermo, Scuola Politecnica - Faculty of Engineering

2) Behavior in Speaking Tests: A Preliminary Model of Interaction

Jayanti Banerjee, Worden Consulting LLC  
India Plough, Michigan State University

Behavior in speaking tests: A preliminary model of interaction (9 words) In the past decade there has been a proliferation of studies addressing the description and definition of interaction as realized in the paired format testing context. Galaczi (2008) has identified general patterns of interaction (collaborative, parallel, asymmetric, blended); Ducasse and Brown (2009) have categorized the interactional features (non-verbal interpersonal communication, interactive listening, interactional management) attended to by raters; Nakatsuhara (2006), Brooks (2009), and Davis (2009) have investigated the effects of the proficiency of one’s partner on the co-construction of discourse (amount of talk, topic initiation, and topic continuation). More recently, May (2009, 2011) has suggested the possible influence of communicative characteristics inherent in other languages and cultures on the co-construction of interactional effectiveness. Our understanding of interaction is much richer as a result of this research. However, we still lack a model of interaction to inform the speaking construct. This Work In Progress (WIP) focuses on the interaction component of the speaking construct. It proposes a model of interaction that maximally benefits from the work already completed and applies it to a test of advanced proficiency in spoken English. During the session we will show how we have synthesized key findings from a review of empirical and theoretical research on interaction in the field of language assessment to develop a preliminary model of interaction. We will discuss how we are applying this model to 55 paired, face-to-face speaking tests. Previous comparative analyses of these tests indicated that the paired format elicits a wider range of functions. The specific interactional features produced during the test have not been explored but the presence of interaction-based functions suggests that these occur. The test video-recordings and transcripts are initially analysed using interaction features shown to be significant in studies like those cited above as well as the interaction features described in the scoring rubric of the test. The aim of this empirical application is to test the robustness of the model and to identify modifications. We expect that features that have not been discussed
in research to date will be identified and included in the model being created. In addition to the attention to empirical findings, the model addresses a significant issue that comes to the fore when reviewing the research on the assessment of speaking proficiency. Key, and sometimes debated concepts (e.g., authenticity), are not always defined or different terminology is used for the same object (e.g., paired format/method/task), leading to a vagueness that can hinder the progress of studies informed by and building on prior work. The model of interaction proposed in this WIP includes, and calls for an adoption of, definitions consistent with those used in seminal research in this area. Colleagues attending the session are invited to evaluate our proposed model of speaking interaction. We also look forward to discussing how we might validate the model through application to speaking tests employing formats other than paired, face-to-face.

Thursday, June 23
17:05-17:50
Classroom Building Hall - upper level
University of Palermo, Scuola Politecnica - Faculty of Engineering

3) Justifying the Use of Computer Automated Scoring in Portfolio-Based Writing Assessment in EFL Classroom

Shangchao Min, Zhejiang University
Lianzhen He, Zhejiang University
Dajian Chen, Zhejiang University

The use of computer automated scoring of constructed responses is widely embraced in the field of language assessment (e.g., Carr & Xi, 2010; Weigle, 2010; Xi et al., 2012), yet surprisingly little empirical research (Chapelle, Cotos & Lee, 2015) has been conducted to investigate how this advanced technology can be incorporated into classroom context to facilitate language learning and teaching. To fill the gap, this paper reports on the preliminary results of a longitudinal study examining the effectiveness of Bingo English, an Automated Essay Scoring (AES) system, in assessing Chinese EFL students' writing ability and improving their writing performance with detailed descriptive feedback in portfolio-based writing assessment. The assessment use argument (AUA), which was first proposed by Bachman (2003, 2005) and then further developed by Bachman and Palmer (2010), will be adopted as the theoretical framework of this study. The data collected were 640 essays, consisting of 320 first drafts and 320 revised drafts written by 80 university freshmen in response to each of four essay prompts over a period of four months. All the essays were rated by both Bingo English and two experienced Chinese EFL teachers. Interview and questionnaire data were collected from eight students as well as two teachers to gain insight into their perceptions of the usefulness of the feedback generated by Bingo English in supporting learning and teaching. Using these multiple sources of information, both quantitative and
qualitative data analyses will be conducted. Quantitatively, Many-Facet Rasch Modeling will be utilized to co-calibrate automated and human ratings so as to explore the accuracy and reliability of the automated scores in comparison to human scores. Qualitatively, the various feedback generated by Bingo English on essays written in response to different prompts in the four months will be compared in conjunction with students’ and teachers’ perceptions in order to see whether the automated feedback leads to improvements in learners’ performances. It is hoped that, this study, by justifying the use of computer automated scoring in portfolio-based writing assessment in EFL classroom, helps to promote the application of this advanced technology into classroom context to transform and enhance language learners’ learning experiences and shed light on the teaching, learning and assessment of EFL writing in China and elsewhere.

Thursday, June 23
17:05-17:50
Classroom Building Hall - upper level
University of Palermo, Scuola Politecnica - Faculty of Engineering

4) Incentivizing Students to Reach Stated Proficiency Goals

Bill VanPatten, Michigan State University
Susan Gass, Michigan State University
Paula Winke, Michigan State University
Koen Van Gorp, Centre for Language and Education, University of Leuven

In both teaching and testing for proficiency, an important issue in understanding outcomes is the motivation that students have for achieving any particular level of ability. In most university settings, the only requirement for language students is seat time (i.e., X number of credit hours completed). But if some level of proficiency is a goal of a language program, what motivation is there for students to reach the stated program goal if the only requirement is seat-time (Glisan, 2013; Moser, 2014; Schmitt, 2015; Thompson, 2014)? In this work in progress we report the initial findings of a study conducted with 300 students completing second year Spanish courses at a large U.S. university. All students had a seat-time requirement of 12 credit hours (four semesters of study). The program in which they are enrolled has a stated outcome goal of Intermediate-Mid in speaking, listening, and reading on the ACTFL proficiency scale. However, the program does not require that students meet this goal. In our study, we incentivized two groups of students: for one-half (n = 100) 10% of their final grade was based on meeting pre-established program goals: for a second half (n = 100), 5% of their grade was based on the same criteria. With a third, control group (n = 100), there was no incentivization: they only received 5% for taking the proficiency-based tests. We report the outcomes on the three proficiency measures (listening, reading, speaking) for the two incentivized groups, then we report outcomes from the seat-time-only group (control) vs. the seat time + final grade incentive groups (experimental). Our hypothesis is that the addition of a final grade incentive will
cause the experimental groups to outperform the control group. The entire study was replicated with a smaller group (n = 60) of Chinese language learners. We report preliminary data from the Chinese learners as well. If the incentivized groups outperform the control group, several questions will arise. First, were the incentivized groups actually motivated to perform better? And second, was that motivation long term (present during instruction), or was it short term and situation specific (present during test taking), or both? If the incentivized motivation was only present for the incentivized students while they took the test, and if they outperformed the control group on the test, it may mean that the test-score difference is construct-irrelevant; the test-specific incentivized motivation may have exploited the standard error inherent on the test. Or, if the incentivized motivation is long term (present during instruction; motivating the student to learn more in class), then the question is whether programs should enact proficiency-based exit-test requirements that articulate the programs’ stated goals (see Chalhoub-Devile, 1997). In this case the research would suggest yes, exit-test requirements should be put in place.

Thursday, June 23
17:05-17:50
Classroom Building Hall - upper level
University of Palermo, Scuola Politecnica - Faculty of Engineering

5) Mapping Language Use to the CLB and CELPIP-General LS within Workplace Contexts for New Immigrants

Christine Doe, Mount Saint Vincent University
Scott Douglas, University of British Columbia
Liying Cheng, Queen’s University

This work in progress reports on a research study that examined the interface between the workplace context for new Canadian immigrants in non-professional positions and two language constructs: (1) the Canadian Language Benchmarks (CLB) and (2) the Canadian English Language Proficiency Index Program General Listening and Speaking (CELPIP-General LS). To do so, we collected spoken data to map actual indicators of language use and communication challenges from Canadian immigrants working in non-professional positions referenced against the above two commonly used benchmark and proficiency criteria in Canada. Theoretical Framework. Within a socio-cognitive understanding of how additional languages are acquired, we draw on a model of English language proficiency that views additional language acquisition as developing across highly contextualized and cognitively undemanding contexts to increasingly decontextualized and cognitively demanding contexts (Cummins, 1981; Roessingh, 2006). This view places an emphasis on language learners’ abilities to express their lived experiences as an important stage of language development. In this sense, the function of language use is intimately connected to communicative purposes across various contexts, including the workplace. Stated another way,
language use represents the communication tasks (Bachman and Palmer, 2010) that new immigrants need to complete to be successful in their jobs, such as, taking a food order, negotiating break times with fellow employees, or talking to a customer. 

Methods. This study draws on multiple methods to (1) explore the functional language use and communicative challenges that new immigrants face in their lives within workplace contexts and (2) map them to given CLB levels and the CELPIP-General LS proficiency levels. We collected data from 45 participants across three stages, from CLB level 3 to 6. In Stage One, participants were interviewed to identify the various tasks they perform in order to be successful in his or her job; they also completed the CLB self-assessment. In Stage Two, participants were interviewed again to discuss his or her communication challenges when completing the various language use tasks at his or her workplace. In Stage Three, the participants took the CELPIP-General LS. An open-ended approach was taken to analyze the interview data. The codes were then mapped onto the CLB and the CELPIP-General LS level descriptors. At the time of LTRC 2016, we will be in the process of developing narratives of participants whose data illuminate language use tasks and functions within a particular benchmark and language proficiency levels and how the challenges manifest across workplace contexts (Creswell, 2008). Presentation. In the work in progress presentation, we will discuss the current literature and findings from the study. This study is a pilot study for a larger-scale study that will further explore the intersection of language use in non-professional positions and language constructs through a mixed methods design. We welcome comments and suggestions from the audience regarding the interpretation of the initial findings and design of the larger study.

Thursday, June 23
17:05-17:50
Classroom Building Hall - upper level
University of Palermo, Scuola Politecnica - Faculty of Engineering

6) Investigating the Validity of the Multiple-Choice Method for Assessing Comprehension of Conversational Implicature

Stephen O’Connell, University of Maryland, College Park

The ability to understand conversational implicature is linked to the upper levels of many language proficiency frameworks and scales (e.g., the Proficient levels, or C1/C2, in the CEFR, the Distinguished level in the ACTFL Guidelines), but how understanding implicature is assessed using multiple-choice items is an under-researched area. Building on work from the philosophy of language (Grice, 1975) and pragmatics (Sperber & Wilson, 1995), researchers such as Bouton (1988, 1994, 1999), Roever (2006, 2013), and Taguchi (2005, 2008, 2009) have made important contributions to understanding how different types of implicature have different levels of difficulty for learners of English. But a closer examination of how the ability to comprehend conversational implicature is assessed in standardized testing
contexts and how the ability is linked to a widely used proficiency framework such as the CEFR is missing. That is, clearer validity evidence for this method of assessing this important listening skill is needed. This study in progress aims to investigate these questions by conducting an experiment with more than 100 learners of English. The participants will be administered a test of conversational implicature (developed in accordance with AERA standards: i.e., multiple rounds of review by experienced test developers followed by piloting with L1 and L2 speakers of English). The items, designed to test conversational implicature at three levels of implicature strength, rather than by type, will be administered in both a multiple-choice and constructed-response formats. Participants will also have their proficiency measured with an independent language proficiency test that has been linked to the CEFR through a formal standard setting procedure. Additionally, participants will be assessed for their working memory capacity. Uncovering a link between working memory capacity and comprehending conversational implicature may shed light on the question of why there is an ongoing effect for working memory capacity even with higher intermediate and advanced L2 speakers, for whom real-time processing of linguistic information is generally far less of an issue. The method for assessing the degree of strength of implicature (by taking into account the general context, specific context, and number of indicators necessary for processing the implicature tested) will be discussed along with pilot results from both the multiple-choice and constructed response formats. Additionally, the rationale for including working memory capacity as a covariate and the planned analysis will be discussed.

7) Investigating the Usefulness of the TOEFL® Primary™ Reading and Listening Tests to Measure Young EFL Students’ English Learning Progress

Ian Blood, ETS
Yeonsuk Cho, ETS

The TOEFL® Primary™ Step 1 and Step 2 Reading and Listening tests measure the English ability of young students ages 8 and above who are learning English as a foreign language (EFL). Step 1 is designed for students who are at the beginning stages of learning English, and Step 2 is designed for students who have acquired some communicative English skills. Scores on TOEFL Primary Step 1 and Step 2 are vertically scaled, making them potentially useful for monitoring test takers’ English learning progress over time. Indeed, some institutions have already opted to use the tests for this purpose by administering them to students twice or three times in a single year. However, because TOEFL Primary is a relatively new testing program,
no empirical evidence has yet been gathered to examine the degree to which the tests are able to detect changes in students’ language ability when deployed two or three times in a year. Therefore, this study evaluates the TOEFL Primary tests’ ability to detect growth in young EFL learners’ language ability over time. In addition, we will attempt to investigate background variables (such as age, hours of English instruction per week, and years of English study) that are expected to impact young learners’ English language growth as measured by TOEFL Primary. To address these research questions, we will analyze operational TOEFL Primary scores and background questionnaire responses from young learners who have taken the tests two or three times in a one-year period. As the data have a nested structure (students within institutions and countries), and the time intervals between testing occasions vary, we plan to test a multilevel model in which changes in students’ TOEFL Primary scores are a function of the time interval between testing occasions. Moreover, as sample size permits, we intend to include background variables (e.g., hours of weekly English instruction) as covariates in our analysis to investigate their impact on TOEFL Primary score gains. This presentation of work in progress will provide preliminary evidence related to the utility of the TOEFL Primary tests to monitor young EFL learners’ English-learning progress. The study will also help advance understanding of the English language development of young EFL learners in general by investigating the relationship between growth in English language proficiency and student-, classroom-, and institution-level background variables that are hypothesized to impact language learning.

---

Thursday, June 23
17:05-17:50
Classroom Building Hall - upper level
University of Palermo, Scuola Politecnica - Faculty of Engineering

8) The Discursive Construction of “Security”: An Analysis of the UK Secure English Language Testing policy

Johann Unger, Lancaster University
Luke Harding, Lancaster University
Tineke Brunfaut, Lancaster University

Current conceptualizations of language assessment would hold that the central concern of assessment should be construct validity: the adequacy of inferences based on test scores for decision-making. From this perspective, the quality, fairness and ultimate usefulness of a language test for a specific purpose or context should be determined through careful, evidence-based reasoning. However in many policy contexts other considerations might overshadow this approach to the judgement of whether a test is fit for purpose. One such example is the emergence in recent years of policy surrounding “Secure English Language Tests” (henceforth SELTs) in the United Kingdom (UK), which are recognised, and condoned, by the UK Home Office as suitable tests for immigration and visa purposes. In place of discussions
of test quality and usefulness, the notion of “security” has emerged as a significant feature in the discourse of politicians, in official documentation on UK Home Office websites, and subsequently in the news media. At a basic level, “test security” is connected with vigilance against cheating and fraud, the provision of biometric data through which to determine identity, and with secure processes for data handling and reporting of results. However discourse surrounding SELT policy has frequently linked test security at the technical level with issues of national security, in effect drawing language testing into a broader narrative of security in UK immigration policy. This work-in-progress presentation will report on an ongoing project which seeks to apply a critical discourse studies approach to a language testing issue. A range of documents relating to the SELT policy in the United Kingdom will be analysed through a discourse-historical lens (see Reisigl & Wodak, 2009), with a view to revealing the manner in which security became a prominent concern in the development of language proficiency requirements for immigration purposes. Documents will include UK Home Office website information, parliamentary records, news media, and the documents provided in the most recent (2014) tender round for selecting Secure English Language Tests (which have been acquired through the Freedom of Information Act). Background to the SELT policy will be provided, and a sample analysis will be presented to illustrate how security has been constructed across different government channels over time. We will argue that the proximity of questions of language test security with broader issues of national security in public discourse on language testing suggests that “security” is functioning as a topos – a generalised warrant that does not need detailed explanation to be persuasive in arguments – to support claims of the strength and fairness of immigration policy more broadly.

Thursday, June 23
17:05-17:50
Classroom Building Hall - upper level
University of Palermo, Scuola Politecnica - Faculty of Engineering

9) Investigating Language Assessment Literacy in Brazil

Gladys Quevedo-Camargo, University of Brasília
Matilde Scaramucci, State University of Campinas

In the last fifteen years, Brazil has witnessed the growth in the importance of high-stakes exams in its educational system. Consequently, teachers have faced the challenge of assessing their students so as to comply with the required levels in study programmes and job opportunities. In this scenario, assessment literacy (Stiggins, 1991, 1995) has become a highly relevant issue for the development of language professionals. However, it has been little explored in the Brazilian teacher education. Based on the three actions proposed by Stiggins (1991), namely (1) understanding the concept of assessment literacy; (2) distinguishing levels of
assessments literacy; and (3) developing programmes of assessment literacy, the

current work in progress has the general aim of building a body of knowledge related
to the assessment literacy of the Brazilian language teacher. Five specific objectives
have also been established: (a) to carry out a literature review of the state of art in
assessment literacy worldwide; (b) to identify, in the literature, the elements that
constitute language teachers’ assessment literacy; (c) to identify which elements
from (b) can be found in the language teacher pre- and in-service teacher education;
(d) to plan actions for the strengthening of Brazilian language teachers’ assessment
literacy; and (e) to implement the planned actions in language teacher education at
local, regional and national level. The general research question guiding this work
is Which elements should constitute the Brazilian language teacher assessment
literacy? In addition, three specific questions were also made: (1) In general, what
kind of knowledge is presented by the specific national and international literature
about language teachers’ assessment literacy?; (2) Which of the elements identified
in the previous phase can be found in the Brazilian language teacher pre- and in-
service teacher education?; and (3) Which actions can be planned and implemented,
at local, regional and national level, aiming at the strengthening of Brazilian language
teachers’ assessment literacy? This study is predominantly of a qualitative nature
and will include, apart from an extensive literature review, document analysis (official
educational guidelines), interviews and questionnaires, applied to pre- and in-service
language teachers. This study started in August 2015 and has been planned to be
developed along three years.

Thursday, June 23
17:05-17:50
Classroom Building Hall - upper level
University of Palermo, Scuola Politecnica - Faculty of Engineering

10) The Construct of EFL Academic Writing Ability as
Operationalized in the Context of Chinese Colleges

Cecilia Guanfang Zhao, Shanghai International Studies University
Jinsong Fan, Fudan University

The evaluation of students’ educational progress in Chinese schools still depends
heavily and almost exclusively on the use of timed tests. The extensive use of such
tests, however, leads to a general lack of writing ability among Chinese students.
In EFL teaching and learning, this problem is especially salient (Yan, 2012) and is
seen as “directly affecting [Chinese students’] research and learning ability at the
postgraduate level” (Zhang, 2010, p.311). To help students develop their writing ability,
however, we need to first define EFL academic writing ability and identify the major
challenges and difficulties students face, so that teachers and students know what
exactly needs to be addressed and learnt in and beyond the classroom setting. As
a response, this study aims to collect evidence from various channels that help
define EFL academic writing ability in the Chinese context and identify the types
of challenges students face when writing in this particular social and educational context. First, national English learning standards, widely-used writing textbooks, and high-stakes English tests (with a focus on the writing section and the rubrics used to score the writing section) for Chinese college students are examined to identify the most prevalent type of writing tasks as well as the components of English writing ability that are addressed, taught, and assessed in such a context. Results from such document analysis should be able to present a fairly accurate account of the operational definition of EFL academic writing ability in Chinese colleges. Meanwhile, results from such analysis are also examined vis-à-vis the Hayes-Flower (1981) writing model, the language ability model (Bachman & Palmer, 1996), and the writing instruction and assessment practices in North American universities. Such comparative analysis could reveal potential construct under-representation or misrepresentation in the operational definition of the construct that may be a source of Chinese EFL students’ writing difficulties in the first place. Additionally, insights into the nature of EFL academic writing ability, and the challenges associated with the teaching and learning of this type of writing, are also gained from teachers and students themselves through surveys and interviews. While the study is still under way, preliminary results from a pilot study showed that EFL academic writing is defined, in Chinese colleges, as prompt-based opinion writing that is decontextualized and non-analytical, which is inadequate for developing academic writing ability that requires critical reading and analysis of source texts, logical reasoning and presentation of sophisticated ideas, and awareness of the conventions of scholarly conversations within a particular discourse community. Analysis of survey data (N=60) further revealed that Chinese college students consider generating sophisticated ideas, articulating their ideas, and word choice, among other challenges, to be the most salient. Such findings are consistent with what Llosa et al. (2011) found in their study of American high school students’ writing difficulties. Implications of such results will be discussed, focusing particularly on the similarities and differences of writing in first and second languages, and how such similarities and differences may inform EFL academic writing instruction and assessment.

Thursday, June 23
17:05-17:50
Classroom Building Hall - upper level
University of Palermo, Scuola Politecnica - Faculty of Engineering

11) Investigating Teachers' Attitudes towards Portfolio Based Language Assessment (PBLA) in Language Instruction for Newcomers to Canada (LINC) Program

Fatemeh Mohammadian Haghighi, The University of British Columbia

This exploratory study will investigate the perceptions of English teachers about the usefulness (Bachman & Palmer, 1996, 2010), challenges and benefits of the Portfolio

Work-in-Progress Abstracts
Based Language Assessment (PBLA) in the Language Instruction for Newcomers to Canada (LINC) programme. Therefore, this study will investigate how the teachers who teach different basic levels of Canadian Language Benchmarks (CLB) – literacy level (below CLB 1) and CLB 1 to CLB 4 – perceive the use and implementation of PBLA in their classes. The reason for selecting these levels is that the minimum language requirement for citizenship application is CLB level 4 in listening and speaking. Due to the significance of achieving CLB level 4 in the lives of immigrants who aim to become Canadian citizens, in this study I aim to investigate the teachers’ perceptions of using PBLA only up to CLB level 4. I expect to have at least two teachers from each level (N=10). The proposed study will seek answers to the following research questions: 1. How do teachers perceive PBLA compared to other traditional assessment methods? 2. What are the benefits of using PBLA, based on teachers’ experiences? 3. What are the challenges of using PBLA, based on teachers’ experiences? 4. What challenges do LINC teachers in different levels face in assessing the portfolios? 5. What are teachers’ views on PBLA training? Method This qualitative study will employ semi-structured interviews as the main method of data collection. The study aims to gain a better insight into the experiences of instructors in utilizing PBLA in the LINC programme, and look into the challenges they face in using and assessing the portfolios. According to Corbin and Strauss (2014), qualitative research is used to “explore the inner experiences of the participants” (p. 5). As such, a qualitative approach best suits the purpose of this study. Data will be collected through semi-structured interviews to provide an adequate level of flexibility, where necessary, to modify the interview protocol in a justified manner. Interviews allow the participants to share their views in a dialogic manner and help me gain a deeper understanding of their perceptions of this method of assessment. Significance of the study Findings will have implications for the LINC service providers, LINC administrators and teachers, as well as for other policy makers who are directly or indirectly involved in this national project—e.g., Centre for Canadian Language Benchmarks (CCLB), Citizenship and Immigration Canada (CIC), Ministry of Citizenship and Immigration (MCI). More importantly, the results of this study will potentially impact the language learners, particularly those who need to reach the CLB level 4 in order to meet the language requirement for the Canadian citizenship application. By casting light on the perceptions of LINC teachers of the PBLA in the LINC programme, this study can potentially lead to the formation of support and scaffolding programmes aimed at a more effective implementation of the PBLA in the LINC programmes across Canada.
Italian legislation has established national guidelines to learning and assessment for students with specific learning disabilities (SpLDs) to guarantee their access to education. As a result of this greater awareness, increasing numbers of dyslexic students are continuing their studies at university level. For language testers, this raises the issue of providing fair and valid tests of English language for these students, while applying the guidelines imposed by law. Researchers at the University of Venice have for some time investigated English language assessment for Italian university students with dyslexia and SpLDs. In particular we have examined the issues of fairness and validity regarding the use of text-to-speech technology by these students in testing contexts. In this project we move to the next step in our research: exploring the cognitive validity of a high-stakes English language test for Italian university students with SpLDs. Cognitive validity involves investigating whether the mental processes a test elicits from the test-taker are similar to the processes the test-taker would use in real-world conditions: Is the test valid as a predictor of real-life performance? We examine the test results of some Italian university students with dyslexia and related SpLDs. The test is a multiple-choice entrance test administered to first-year students to meet mandatory B1 English language requirements. It assesses reading and listening skills, and knowledge of grammar and lexis. The students sit the test with accommodations: extra time and separate sessions. The test is computer-based and uses the open-source learning platform Moodle. The analysis of the results uses Moodle Item Analysis tables and is divided into two separate phases: statistical analysis and content analysis using expert judgment. The first phase identifies the items that appear to be most difficult for students with SpLD and/or items that do not discriminate between learners. The second phase carries out content analysis of the critical items identified in the first phase to investigate whether they are a representative sample of the content of objectives or specifications the test was originally designed to measure. The aims of this analysis are to verify whether content validity seems to be maintained; whether the test content, or part of it, is more critical for dyslexic students than other students; and whether there are similarities in dyslexic test-takers’ performances. Little research on L2 learning of students with SpLDs has been carried out, and fails to provide an adequate basis for definitive conclusions about the mental processes elicited by language tests. Cognitive validity therefore remains an almost unexplored feature of the overall test validity for these students. An additional challenge derives from the fact that learning difficulties are not homogeneous, and similar disabilities affect different people in different ways, making the identification
of a possible standard mental process problematic. Nonetheless, investigation of cognitive validity may be decisive as the understanding of mental processes in the foreign language learning of students with SpLDs is potentially the basis for the identification of the test construct on which to design and build valid tests.

Thursday, June 23
17:05-17:50
Classroom Building Hall - upper level
University of Palermo, Scuola Politecnica - Faculty of Engineering

13) A Fresh Look at the Word Associates Format to Measure Depth of Vocabulary Knowledge

John Read, University of Auckland

An enduring distinction in second language vocabulary testing is that between breadth and depth of vocabulary knowledge. The construct of breadth, or vocabulary size, has been relatively easy to define, and to operationalise as tests of the form-meaning link. On the other hand, depth of vocabulary knowledge has proved to be more problematic, both to define as a distinct construct and also to measure, despite various efforts at construct definition over the years. (Chapelle, 1998; Henriksen, 1999; Nation, 2001; Read, 2004). A recent review article by Schmitt (2014) makes it clear that uncertainty about the depth construct persists to this day. In practical terms, Read’s (1993, 1998) two word associates tests have been widely used as research tools and the 1998 version in particular is regarded by many writers as the standard measure of depth of vocabulary knowledge. However, the 1998 version has limited value for any operational purpose because it is designed specifically to measure knowledge of adjectives as target words and the actual words are taken from an ill-defined domain of high-frequency adjectives. The starting point for the present project, then, is to develop a new test which samples from an established list of academic vocabulary across the major lexical word classes. The design of the test takes account of research since the 1990s on the word associates format, notably by Schmitt, Ng & Garras (2011); Schmitt’s (2013) comprehensive review of the literature on breadth and depth of knowledge; and the ongoing work by Fitzpatrick (2006) and her colleagues on the nature of word associations in a second language. It also draws on advances in corpus linguistics in identifying high-frequency collocations, since the syntagmatic relationship between words is a key component of the word associates format. The current aims of the project are to develop two versions of the test: a general academic version based on Gardner and Davies’ (2014) New Academic Vocabulary List; and a discipline-specific version designed for Business students. An appropriate measure of vocabulary size will also be used as a reference measure to help establish the distinctiveness of the contribution of the depth measures. In this session, the rationale for the design of the new test will be presented, together with the results of initial trials with students taking courses in English for Academic Purposes both before and after admission to a major English-medium university.
14) Exploring the Relationship between TOEFL Scores and Academic English Performance on Reading, Speaking and Writing Tasks at a Colombian University

Isabel Tejada-Sanchez, Universidad de los Andes

Today, English has an undisputable global role as being the most frequently used language of scientific publication. Thus, it comes as no surprise that many institutions of higher learning in non-English-speaking contexts place a premium on learning the language in English for Academic Purposes (EAP) programs, such that students may have a better footing within the international academy (Hamp-Lyons, 2011). In Colombia, the positioning of English as being important certainly holds true. The national government has recently implemented the program Colombia Bilingüe, which foments the development of “human capital” in the country through the teaching and learning of English. At our university, English is similarly viewed as being important, as there is specific EAP coursework offered both at the undergraduate and graduate levels of instruction. While research evidence shows extensive work on different aspects of EAP program development related to our university’s graduate population to date (Janssen, Ángel, & Nausa, 2011; Janssen, Nausa, & Rico, 2012;; Janssen, Meier & Trace, 2014, 2015), little systematic program development has taken place at the undergraduate level. The present work-in-progress derives from a larger program evaluation project intended to examine language performance at this level (N≈3000) and its implications in the students’ academic outcomes. Concretely, this study seeks to explore the extent to which TOEFLiBT scores are associated with i) performance in the regular English program and ii) with the participants’ overall academic performance that requires especially reading, writing and speaking in English. The English program at this university implies a completion requisite for the students and it is organized in seven levels divided in two cycles: a ‘basic cycle’ focusing almost exclusively on developing reading comprehension skills and strategies needed for academic success, and a ‘skills cycle’ following content-based courses intended to promote writing, pronunciation and oral skills. Focus on listening is not given at a specific part of the program. The requisite compels students to either complete the whole program or validate it with a global TOEFL iBT score of 85. The study examines associations between TOEFL and in-house tasks for each of the skills considered. The main motivation for this study is twofold: 1) the necessity to bolster the EAP program as the university wishes to certify a competitive level of English across all fields, empowering students to pursue postgraduate studies in English-medium higher-education institutions and 2) the necessity to attain English outcomes that are internationally comparable. Evidence of using the TOEFL test as a research tool at university level has been attested and validated for a myriad of purposes such as placement and linguistic performance (Ockey, Koyama, Setoguchi...
& Sun, 2014; Rosenfeld, Leung, & Oltman, 2001). Interestingly, these studies have addressed the importance of validating the role of TOEFL iBT scores as predictors of real language use, specifically in academic environments. Findings from this study are expected to shed light on the usefulness of the TOEFL test and how non-English-medium institutions achieve language proficiency in order to trigger both competitive academic performance across fields and internationally comparable EFL outcomes.

Thursday, June 23
17:05-17:50
Classroom Building Hall - upper level
University of Palermo, Scuola Politecnica - Faculty of Engineering

15) Diagnostic Writing Assessment of First-Year Engineering Students: Providing Profile-Based Pedagogical Support to Develop Academic Language Proficiency

Janna Fox, Carleton University
Maryam Wagner, Carleton University

Engineering students need to be effective communicators and proficient technical writers to ensure their success in university, but research suggests that some entering undergraduates need supplementary support to develop to their full potential (Browne & Doyle, 2010). Without early support, they are more likely to drop out of their programs. Furthermore, the writing support needs to identify both learners’ strengths and their areas for improvement to provide a pathway for remediation (Alderson, 2005). This work-in-progress reports on research in which: 1) A diagnostic assessment was designed to identify the characteristics of entering students’ engineering writing at a university; and 2) An in-progress investigation in which pedagogical intervention is designed based on students’ individual writing profiles and delivered to best support their writing needs. Two writing tasks were designed to elicit diagnostic information across three dimensions of writing: engineering genre expectations, logic, and language. Raters evaluated 489 engineering students. These students shared gaps in several key areas: academic language proficiency, mathematical background; and scientific text literacy (Authors). These students were invited to seek pedagogical support at a first-year Engineering support Centre served by Scholars who provide support. Scholars are drawn from graduate students who are experts in writing. Despite possessing background knowledge and experiences that identify them as suitable writing tutors, Scholars do not necessarily have shared knowledge, opinions or beliefs about how to best provide writing support to students. There is no systematized or principled approach to providing writing support, rather the approach is ad hoc and addresses the immediate needs of students. Alderson (2005) and others have argued that assessment is not diagnostic unless it is connected with pedagogical support. In this case there is a connection between diagnosis and pedagogical support, but it is not yet sufficiently targeted in relation to the individual needs, which the
diagnosis highlights. This research aims to fill this gap by developing an intervention for students requiring support to advance their academic writing proficiency by drawing on their individual writing profiles. At-risk students’ diagnostic writing assessments were evaluated using the analytic rubric and will be used to build a cognitive diagnostic model that characterizes their academic writing skill profiles. Students’ estimated skill writing profiles will be used to determine the range and type of the most frequently occurring profiles. Using students’ writing samples, we will identify salient features of the writing that can be used to characterize each writing profile by engaging writing experts, Scholars, and raters using the Delphi technique (Dalkey & Helmer, 1963) to develop pedagogical interventions that: a) provide support specific to students’ writing profiles; b) contribute to the development of students’ self-regulatory skill behaviours; and c) are also characterized by students’ background variables. Scholars will be trained to provide pedagogical support using the developed and individualized plans; the implementation will be tracked over a three month period and specific students’ (with varying profiles) development will be tracked. Participating students will be invited to write a post-diagnostic assessment. It is anticipated that this study will be completed in May 2016.

Thursday, June 23
17:05-17:50
Classroom Building Hall - upper level
University of Palermo, Scuola Politecnica - Faculty of Engineering

16) Authorial Voice in Integrated Writing Assessment Tasks
Atta Gebril, The American University in Cairo

Integrated writing tasks are increasingly used to make inferences about L2 students’ academic writing ability. While the writing assessment literature has focused on the comparison between integrated and independent tasks and also on issues related to source use, to the best of the researcher’s knowledge, little research, if any, has looked into the relationship between source materials and authorial voice in writing assessment. According to Matsuda, voice is defined as “the amalgamative effect of the use of discursive and non-discursive features that language users choose, deliberately or otherwise, from socially available, yet ever changing repertoires” (2001, p. 40). In integrated tasks, it is important to look into writer identity and how it is affected by external sources. When writers attempt to synthesize information from different texts, some of them seem to struggle with their voice, especially when they overuse the sources. For this reason, it would be useful to investigate how the interaction with external sources affects authorial voice. The current study addresses this issue by attempting to answer the following question: a) How does the use of external sources in integrated tasks affect authorial voice? b) How is authorial voice represented at different proficiency levels in integrated tasks? The author will collect data from students who take ESL courses in a university context. The students...
will be completing an integrated task which includes a prompt on a certain topic, followed by reading sources discussing this topic. They will be required to use the sources in their writing to support their opinion and to provide examples. Since the project looks into how authorial voice is represented across different proficiency levels, the researcher will attempt to select a representative sample that includes both advanced and low-proficiency students. Proficiency in this context will be operationalized based on students’ performance on the writing task. The task will be scored using an integrated writing analytic scoring rubric. In addition, Hyland (2008)’s model of voice will be used to code authorial voice features. Hyland’s interactional model of voice includes two major areas or ‘systems’ as he calls them: The writer-oriented stance and reader-oriented dimension. A qualitative coding scheme will be developed based on this model and two coders will be used to analyze the data. Before starting operational data analysis, the researcher will pilot the coding scheme and will provide necessary training for raters. As for quantitative data analysis, multiple regression will be used to look into how proficiency contributes to authorial voice in integrated tasks. The project could provide a number of implications for both language testers and writing instructors.


Thursday, June 23
17:05-17:50
Classroom Building Hall - upper level
University of Palermo, Scuola Politecnica - Faculty of Engineering

17) ACTFL Proficiency Levels and TOEFL Scores of Students Entering US Universities

Daniel Reed, Michigan State University
Susan Gass, Michigan State University
Erwin Tschirner, Universität Leipzig

The ACTFL English Research Study is designed to provide information on initial proficiency levels of international students upon matriculation as well as on the relationship between TOEFL scores and English language proficiency levels as defined by the ACTFL Proficiency Guidelines. The results of the study are expected to aid institutions in determining more reliable cut scores for international tests of English such as the TOEFL for various academic purposes and to determine more appropriate courses of action to helping students of varying proficiency levels to attain appropriate levels for both academic purposes and student integration into US society. This study reports on the results from one of 20 institutions that are expected to participate. All students in this study take the computer-adaptive Listening and Reading Proficiency Assessment (L&Rcat), and some take the Internet-
delivered Oral Proficiency Interview (OPIc), and/or the Writing Proficiency Test (WPT). Prior to taking one or several ACTFL proficiency assessments, each student is asked to take a brief demographic survey to evaluate their language background, their probable levels of English in listening, reading, speaking, and/or writing as well as their exposure to English. This work in progress will report a statistical analysis to establish the proficiency levels of international students admitted to one major Midwestern US university, to determine correspondences between TOEFL scores and proficiency levels in listening and reading, to determine correlations between skills, and to determine instructional needs depending on international English test scores. Although the logistics of administering large numbers of proficiency tests are challenging (requiring language labs with up to date equipment and knowledgeable personnel), the particular university in this work-in-progress study had already conducted large-scale testing of ACTFL proficiency levels in French, Spanish, Russian, and Mandarin and was thus well positioned to gather and analyze some of the preliminary data in the larger ACTFL English study. The logistical problems and solutions, along with the preliminary results of proficiency score distributions and the comparisons with TOEFL will be presented and discussed along with implications for the remainder of the project and curricular changes that should be considered.

Thursday, June 23
17:05-17:50
Classroom Building Hall - upper level
University of Palermo, Scuola Politecnica - Faculty of Engineering

18) Investigating the Validity of TOEFL iBT Integrated Writing Assessment: Do the Tasks and the Rating Scheme Used Require any Revision?

Maryam Homayounzadeh, Shiraz university
Alireza Ahmadi, Shiraz university

With the renaissance in the integrated assessment of academic writing, the Test of English as a Foreign (TOEFL iBT) adopted a variety of tasks in its Internet-based version, which involves reading and listening comprehension skills in the evaluation of writing. The tasks are claimed to provide a more authentic and valid account of the test-takers’ academic writing ability, which entails writing by drawing upon external sources (Cumming, 2013). In spite of the bulk of evidence supporting the claim, there are still questions requiring further investigation. One such issue pertains to the impact of source characteristics on the quality of the examinees’ writings. Although characteristics of source texts, including their genre, topic, propositional density, and lexico-grammatical difficulty (Bon-Gettler & Kendeou, 2014; Jiuliang, 2014; Yu, 2009) have been proved significant in affecting the quality of test-takers’ summaries, no studies have investigated this effect among the various TOEFL iBT tasks. To address the stated problem, the present study will select two pairs of authentic read-listen-
write tasks from among those offered by the ETS TOEFL preparation program, one
pertaining to the compare-contrast and the other to the information-synthesis
genre. In each pair, the tasks will differ in being domain-specific or domain-general
and in the summarizability levels of their reading and listening materials, estimated
based on measures of readability, passivization, and lexical diversity (Yu, 2009)
through Cho–Metrix computational tool. Sixty TOEFL candidates will respond to the
tasks under simulated TOEFL iBT testing conditions and will fill out a questionnaire
about their perceived difficulty of the prompts. Borrowing Weigle and Parker’s (2012)
framework, the particular textual borrowing strategies employed by the examinees
(quoting, referencing, copying, modifying, or reformulating) will be specified and
counted per t-unit in each text. The essays will be further scored for the overall
quality of their summaries based on Yu’s (2007) model and for their quality of
language use assessed through Cho–Matrix computational tool. The effect of the
genre, summarizability, domain-specificity, and perceived difficulty of the tasks will
be investigated on the writers’ source-borrowing strategy use, quality of summaries,
and language use. Implications will be provided for integrated writing task design and
assessment rubric development.

Thursday, June 23
17:05-17:50
Classroom Building Hall - upper level
University of Palermo, Scuola Politecnica - Faculty of Engineering

19) ECD for MSA - Developing a Comprehensive Construct
Definition
Bjorn Norrbom, NCA
Yong Lou, NCA
Abdulrahman Alshamrani, NCA

The present paper describes the review and validation of a test of Modern Standard
Arabic (MSA) for general purposes covering listening, reading, writing, and grammar
with the purpose of describing the test construct in sufficient detail through domain
analysis and modelling to support an Evidence Centered Design (ECD) approach.
The CEFR is also used along with a learner corpus-based resource currently under
development that provides CEFR Reference Level Descriptors (RLDs) for Arabic.
The need to develop a clear construct definition for MSA stems not only from the
present test, but from the field at large as MSA construct definitions are typically
expressed in very general terms. Developing a clear construct definition for any
language is fraught with a number of difficulties. However, in the case of Arabic some
of these difficulties are augmented by Arabic diglossia; in particular for listening and
speaking where various colloquial forms of Arabic, at the expense of MSA, can be
said to form the larger part of most Target Language Use (TLU) domains. There are
considerable challenges involved in defining each different form and consequently
creating tasks and scoring procedures. Furthermore, there is an ongoing discussion
among Arab linguists as to whether MSA itself can be defined in any meaningful way. Perhaps due to the difficulties involved in defining the MSA construct, test providers, including those of the present test, tend to rely more on psychometric evidence than a solid construct definition for validation purposes. From an ECD point of view, this is considered as problematic. At the core of ECD is the Conceptual Assessment Framework (CAF) which consists of the Student Model (SM) describing what we measure, the Evidence Model (EM) describing how we measure, and the Task Model (TM) describing where we measure. For the CAF to be effective in validation procedures, it must rest on a solid foundation of domain analysis and modelling. In the case of MSA, such a foundation is typically lacking and we cannot therefore draw a valid inference between the TM and the SM. Hence, any psychometric evidence quoted in the EM also comes into questioning. The ECD “sources of information” were taken as a starting point for domain analysis and modelling. Sources included academic and professional real-life tasks, national and international Arabic L2 programs, high-quality text books, available tests of Arabic, and CEFR descriptors and RLDs. Preliminary results indicate that the parts of the Arabic construct that are possible to define more clearly and comprehensively are Reading Comprehension, Writing, Grammar, and a highly restricted form of Listening Comprehension. No attempt was made at domain modelling for speaking. To the best of our knowledge, no similar comprehensive research for testing purposes has previously been reported elsewhere and we also feel that it will require a strenuous and continuous effort on part of members of the Arabic language testing community. We hope that the present paper will be a starting point in coordinating such research efforts, particularly for listening and speaking.
Who'er doth enter this horrid tomb.
Here sees the realm of cruelty severe;
Wherefore 'tis writ upon the walls of gloom,
Abandon hope all ye who enter here.
Down here we know not if 'tis dark or day.
But tears we know and pain and cruelty;
And here we know not if we ever may
Hear struck the hour of long-crave liberty.

A mid-17th century verse, translated from Sicilian, believed to be written by Simone Rao, and found on the wall of a cell in the Inquisitor's Prison in Palermo. Cited in Seeking Sicily by John Keahey.
**Using Web-Technology and Spoken Dialogue Systems to Assess L2 Pragmatic Ability**

*Veronika Timpe Laughlin, ETS*
*Heidi Han-Ting Liu, Teachers College, Columbia University*
*Jennifer Wain, ETS*
*Keelan Evonini, ETS*

Recent advances in automated speech technology have made it possible to implement interactive speaking tasks into computer-mediated learning and assessment contexts to evaluate pragmatic performance. We have explored the use of automated spoken dialogue systems in a prototype capability for a self-access, web-based, interactive platform that is intended to help adult English language learners increase awareness of English pragmatics in the context of the workplace. Embedded in the scenario of a real life job cycle are nine learning modules, each of which focuses on a specific pragmatic phenomenon or speech act that is important for successful communication in the workplace, such as implicatures, making requests, suggestions, and small talk. Each learning module is complemented by an assessment module, designed as an end-of-unit (achievement) test. In this presentation, we will demonstrate research into the development and use of a sample assessment module focusing on the speech act of requests. After briefly outlining the Evidence-Centered Design framework and structure of the end-of-unit assessment, we will focus in particular on the interactive speaking tasks included that use automated spoken dialog systems to assess the L2 learners’ pragmatic performance in a multi-turn conversation. First, we will describe the software tool that is used for authoring speaking items with branching conversations and the automated dialog system infrastructure that can process a test taker’s spoken responses in a real-time, interactive application. Then, we will discuss the types of evidence of request performance that these items can elicit and how they can be used to expand the construct coverage of a computer-mediated assessment of productive pragmatic ability. Finally, we will present user samples and findings from a study that investigated how the interactive speaking items and the spoken dialogue system actually performed in assessing L2 pragmatic skills.
Test-ify: Software for Assessment Design and Instruction

Adrian S. Palmer, University of Utah
Daniel Dixon, University of Utah (Asia Campus)

Test-ify: software for assessment design and instruction Authors: Adrian Palmer and Daniel Dixon

Bachman and Palmer (2010) described a justification-based approach to assessment development employing four documents: “Justification” (Assessment Use Argument and Backing), “Initial Planning,” “Design Statement,” and “Blueprint.” Five years of experience teaching the material in the book and using it to design language tests have shown the traditional book format to be challenging. Primarily, this challenge stems from the large amount of information that needs to be kept track of across the four documents. Early on in their work with the book, Bachman and Palmer realized that a well-designed software program might be useful in helping address the challenges associated with tracking information. Using software would allow designers to focus more on the development of their assessment rather than the organization of the information. However, their initial attempts to find competent software designers proved both very expensive and, ultimately, futile. In 2012, one of Palmer’s graduate students, Daniel Dixon, saw Palmer’s demonstration of the product of Bachman and Palmer’s search and thought that he could improve on what he had seen. For the next four years, Palmer and Dixon worked to design and create software that managed the document creation process and automated the organization of information. Further, the software provides an interface that is intuitive, easy to learn, and easy to use. Those familiar with a Word document and typical web-based form completion tasks can easily navigate and use this software. This demonstration shows how the software works and what it accomplishes that a paper-based approach cannot. The Test-ify software uses Microsoft’s InfoPath® application for designing, distributing, filling out, and submitting electronic forms containing structured data. Test-ify includes electronic versions of the four documents described in Language Assessment in Practice (Bachman & Palmer, 2010). It also instantly and continually organizes all of the hundreds of interactions among the information in these documents so that the test designer does not have to keep track of and update this information. In this demonstration we will first show how the Test-ify interface looks and how basic operations are used to create documents and manage interactions among them. Next, will show an example of how Test-ify is used to create a type of assessment task by modifying the characteristics of a target language use task. Finally, we will show how the Test-ify can be used to improve the teaching of assessment design and use. Each of these parts of the demonstration will take only a few minutes, so we should be able to give you a feeling how the software works within the allotted time while allowing time for questions. While some familiarity with Language Assessment in Practice would enhance the impact of the demonstration, this is not necessary in order to appreciate the usefulness of software of this type.
Using Technology to Help Learners and Teachers in a Learning-Oriented Approach to Assessment

Evelina D. Galaczi, Cambridge English Language Assessment
Sarah Unsworth, Cambridge English Language Assessment

Assessment is often conceptualised in binary terms: formative vs. summative, classroom vs. large-scale, assessment for learning vs. assessment of learning. While this conceptualisation is useful and intuitively appealing, it oversimplifies the assessment landscape and can become a false dichotomy, e.g. summative tests can be used formatively and formative tests can take on a summative role. As a result, the view of assessment as supporting learning in a systematic and integrated fashion – Learning Oriented Assessment – is gaining ground in theoretical discussions. It is also finding practical applications. In the practical realm, learners present some typical needs, as seen in these learner comments: ‘I want to speak like real people…I mean real language – not classroom language.’ and ‘I think my English is pretty ok. It is useful when you do some exams ‘cos it tells you how well you are and what you don’t know. Even though I don’t like tests so much.’ These statements raise, among other things, the issue of extending learning beyond the classroom, and of the ambivalent relationship that students often have with exams and assessment. At the same time, teachers are faced with their own challenges in the contemporary classroom setting, such as lack of time to prepare activities and assessments tasks, lack of digital support to keep detailed records of students’ achievement and progress, and the need to deal with large and/or mixed-ability classes and provide personalised feedback or suggestions for extension learning activities. In this demo, we will briefly present the concept of Learning Oriented Assessment and will discuss its key features and the synergies between learning and assessment tasks, learner involvement and targeted feedback, which have been shown to increase learner autonomy and learning gains. We will then illustrate the practical application of key Learning Oriented Assessment principles through a demonstration of the ELT course ‘Empower’ (Cambridge University Press), which provides a platform for learner development through integration of learning and assessment. We will specifically focus on demonstrating features of the course which assist learners with the development of their speaking skills through computer-delivered and assessed tasks which target micro-level phonetic features and tasks which target fluency. Through this demonstration we hope to illustrate how technology can play a fundamental role in integrating learning and assessment through helping teachers and learners in the collection of learning data, in structured record keeping, and in providing individualised and pedagogically useful feedback through automated scoring and feedback systems.
A growing number of stakeholders argue for the use of second language (L2) speaking assessments that measure the ability to orally communicate in real time (e.g., Butler, Eignor, Jones, McNamara, & Suomi, 2000; Fulcher, 2003; McNamara, 1996; Brooks & Swain, 2015). Unfortunately, using real time speaking tasks, which provide opportunities for test takers to demonstrate their ability to interact with an interlocutor, have been limited by a number of challenges in large-scale assessments, including the impracticality of getting test takers and examiners in the same physical location. As a possible solution to this challenge, a web-based virtual environment (VE) that allows live voice communication among individuals was developed. Others have achieved some levels of success with off-the-shelf VEs with live voice communication (e.g., Second Life) for language learning activities and to a much lesser extent L2 assessment of oral communication. However, there have been few attempts to design and create live voice communication VE software for the purpose of assessing L2 oral communication. The VE that was built was created explicitly for this purpose and is based on theories of L2 oral language assessment theory. The presentation will include a description of the technology accompanied by a demonstration of the interface and discussion of the language assessment theory that guided the design. The presentation will conclude with results from a small exploratory feasibility study on use of the technology. The technology is designed to be used to deliver a group oral discussion test. Test takers and a moderator are connected by web-based technology from remote sites. Test takers can see avatar images of the other test takers and a static image of the moderator. When a test taker speaks, the test taker’s avatar shows body language movements (arms and head move), and a bubble with the test taker’s name appears above the avatar. Sessions begin with brief introductions and an explanation of what the test takers will do. The test takers are then asked to discuss a topic, which is displayed through the user interface. The moderator and the user interface function together to guide the test takers through the test. In the feasibility study, eighteen university students, grouped in threes, used the VE to complete small group discussion tests. Each group of three had a student located at one of three university campuses. Initial results suggested that the technology functioned quite effectively. There were only a small number of instances when test takers could not hear each other, and the avatar movements led to quite effective turn taking. Interestingly, the VE was perceived quite differently than typical real-world face-to-face environments. Users felt some social presence, the feeling of being present with another person when socializing, but not to the
same extent as in face-to-face oral communication. This may limit the authenticity of the assessment but may also lead to less anxiety. Less anxiety could in turn lead to discourse which better represents what the test taker can produce in a non-testing situation.

---

**Demo Abstracts**

**Thursday, June 23**
14:30-16:00
Room 8 - Classroom Building Hall - 1st floor
University of Palermo, Scuola Politecnica - Faculty of Engineering

**Say Hello to TELL**

*Masanori Suzuki, Pearson*
*William Bonk, Pearson*
*Alistair Van Moere, Pearson*

Test of English Language Learning (TELL) is a newly-developed battery of interactive English language proficiency assessments designed for English language learners (ELLs) in Kindergarten through grade 12 (K-12). TELL is delivered on tablet devices (currently on iPads). Integrated video-based instructions guide students through the test, clearly illustrating how to answer each item type. TELL utilizes multimedia item presentation and elicits various modes of responding - touching, moving, handwriting, speaking, reading aloud, and writing. All test responses are scored automatically by custom-built advanced scoring systems that take advantage of artificial intelligence, machine learning, and natural language processing. TELL is designed to serve three assessment needs: screening, diagnostic, and progress monitoring. K-12 grades are divided into five different grade bands: Kindergarten, 1-2, 3-5, 6-8, and 9-12. A total of 22 item types were developed to measure the four language domains (e.g., listening, speaking, reading, writing). For each test, a subset of 9 to 12 item types are selected, depending on the grade band and assessment need. Some of the item types are quite interactive. For example, to assess the early literacy skills of Kindergarten and Grades 1-2 students, they are asked to write letters or words on the iPad screen using their fingers, and these handwritten responses are automatically scored for accuracy, or to assess print awareness, students see a set of three images and select the image that matches the spoken instructions (e.g., Touch the picture with a title of the book on it.). For grades 3-5 and above, students watch academically oriented video mini-lectures (e.g, a topic from Physics) and summarize them in speaking. Also for grades 3-5, students read a passage on screen and write a summary in writing using the iPad keyboard. Generally speaking, grade bands K and 1-2 have tasks that are more focused on foundational skills such as pre-literacy skills, print awareness, and sound-letter correspondence, as well as on spoken language tasks such as repeating sentences, retelling stories, and describing pictures. Grades 3-12 perform more complex tasks such as watching a short video on content areas and summarizing it in speaking, and reading a passage...
and summarizing it in writing. Furthermore, tablet delivery enables the presentation of interactive and engaging tasks such as touching or moving objects with a finger on the screen in response to spoken or written prompts. Testing times vary based on the type of assessment administered and the grade band of the assessment. In general, all tests are approximately 20 to 50 minutes in length. After the test is completed, the score reporting system allows teachers and schools to see the following score categories:

- Scrrener: overall
- Progress monitor: overall, listening, speaking, reading, writing
- Diagnostic: overall, listening, speaking, reading, writing, subskill scores (e.g., fluency, pronunciation, grammar, vocabulary, pre-literacy, etc)

During the demo, the presenters will demonstrate how the TELL app works, how items are presented in each of the 22 TELL item types, and discuss how the automated scoring system uses particular traits for various item types.

---

**Thursday, June 23**  
14:30-16:00  
Room 8 - Classroom Building Hall - 1st floor  
University of Palermo, Scuola Politecnica - Faculty of Engineering

**Educator Rater Training Materials for the ACCESS for ELLs 2.0 Speaking Test**

*Mark Chapman, WIDA at the Wisconsin Center for Education Research*  
*Meg Montee, Center for Applied Linguistics*

This demo presents the self-access educator training materials for the ACCESS 2.0 Speaking Test, a speaking assessment of academic English language proficiency. The speaking assessment is one section of a test designed for English language learners in grades 1-12 and used in the U.S. to assess over one million students annually for accountability and program purposes. In order to be able to administer and score the face-to-face version of the ACCESS for ELLs 2.0 Speaking Test, educators must become certified. This demo will present the training materials that educators must complete for scoring certification. These materials are developed collaboratively between The Center for Applied Linguistics (CAL) and the Teaching & Learning team of WIDA at The University of Wisconsin – Madison. The materials are designed to serve two main purposes: 1. To establish and enforce rigorous standards for speaking certification. Only educators who meet these criteria will be certified to administer and score the speaking test. 2. To make the training materials engaging and relevant to educators who may have varying levels of assessment literacy. The training is time consuming so the materials must be designed to be practical, accessible, and engaging for educators. Achieving the first goal entailed developing multiple certification quizzes to be completed by the educators after working through...
the training materials. The certification criteria are programmed into the online delivery platform and certification is only granted after the established criteria have been met. The second goal was more challenging and required extensive use of multimedia delivery to introduce and explain the assessment concepts (academic language standards, scoring criteria, rating scale design, and scoring rules) that are essential for reliable scoring. The demo will focus on how the training materials present and explain technical assessment content in an educator-friendly way. The use of a featured educator to guide users through the training materials will be emphasized in the demo, as will the use of animation and video to maintain educator interest. The demo will show how educators build familiarity with the academic language standards that underlie the assessment, learn how the test tasks relate to these standards, understand the types of responses that will be elicited by the tasks, and learn how to apply the scoring criteria to multiple authentic samples of students’ oral responses. Implications of the design of the educator training materials include how to build assessment literacy for ESL educators, with a focus on how to explain assessment-related terms and concepts in an educator-friendly manner. The training materials have been designed for educators and by educators, with the input of assessment professionals. The design and operationalization of these materials is relevant to the discussion of assessment literacy for English Language educators and provides an example of an engaging and rigorous self-access rater training program for large-scale use.

Friday, June 24
11:00-11:30
Room 8 - Classroom Building Hall - 1st floor
University of Palermo, Scuola Politecnica - Faculty of Engineering

**DELTA Writing Component: An Online Automated Diagnostic Test of Academic Writing**

*Alan Urmston, Hong Kong Polytechnic University*
*Michele Raquel, University of Hong Kong*
*Roxanne Wong, City University of Hong Kong*

This presentation is a demonstration of an online diagnostic test of academic writing developed in Hong Kong. DELTA (www.deltatesting.net) was originally developed as an online English assessment of listening, reading, grammar and vocabulary skills for university-level students. DELTA provides diagnostic information at the sub-skill level to enable students to focus their English language enhancement activities so as to foster improvement. In 2014, a writing component was proposed as an outcome of a research study into the writing needs of Hong Kong year one tertiary students across Hong Kong. The research study recommended that due to the large number of students that need to be assessed in a short amount of time, the diagnostic writing test should be administered online and should be automatically scored. The product
is an online diagnostic system with four modules: (i) an online writing test platform; (ii) an automated essay scoring system; (iii) a diagnostic report; and (iv) an online interactive diagnostic feedback platform. This demonstration will illustrate the workflow of the back-end system and end-user experience. We will first describe the design of the test delivery and script capture platform that students interface with. The platform delivers the writing task and provides a "writing pad" for the students to type their answers on. We will then explain the design of the task bank and of the enrolment and test administration system. We will also highlight the flexibility of the system in the scoring of the candidate's writing text so that the text can be scored automatically (by the AES engine) or can be stored for downloading and scoring by human markers. Finally, we will explain how the diagnostic reports are generated and how the interactive diagnostic feedback module can support teaching and learning.

Friday, June 24
11:35-12:05
Room 8 - Classroom Building Hall - 1st floor
University of Palermo, Scuola Politecnica - Faculty of Engineering

An Innovative Online Testing and Rating System for the Integrated Reading-Writing Assessment Research

Yu-Chen Lin, Indiana University
Sheng-Bo Chen, Ohio State University

Integrated writing tasks have been regarded as important test tasks in language assessments because using integrated skills is similar to authentic tasks in the target language use domain. However, the nature of reading-writing activities and the complexities of cognition, metacognition, schemata, L2 acquisition, and other aspects of summary writing become the overwhelming cognitive load on ESL/EFL students (Kirkland & Saunders, 1991). Since there are many possible factors influencing test takers' performances on integrated reading-writing tests, it is extremely challenging for researchers to collect large-scale qualitative and quantitative data and to analyze all aspects of the data in a detailed and efficient manner so as to examine the relationships among those factors and test performances. This demo will introduce how an innovative online testing and rating system can effectively collect large-scale data for related tasks and provide an objective, consistent, and efficient interface for human raters to assign analytic scores and analyze each test taker’s texts.

The 18-minute demo will include several steps listed below:

1. Introduce two interfaces of the Online Testing System for different roles: the login page for test takers and the login page for the proctor/researcher.
2. Demonstrate all steps a test taker should complete online: (1) timed text summary test, (2) untimed questionnaire about cognitive operations, (3) timed text reading comprehension test.
3. Present how a proctor can check test takers’ responses in real time.
4. Introduce two interfaces of the Online Rating System for different roles: the login page for raters and the login page for the researcher.

5. Demonstrate the rater’s rating interface: (1) how to check important idea inclusion and how total points are calculated automatically to determine “Faithfulness to the Main Idea and Supporting Ideas”, (2) how to fairly assign analytic scores for other categories “Language Use”, “Organization”, and “Verbatim Source Use”. These are four critical criteria for assessing a summary (Imao, 2001).

6. Present how a researcher can monitor all raters’ rating and the time they have spent on rating each written summary.

7. Demonstrate how a researcher can obtain all scores and raw data online.

After the actual demo, the presenters will spend about two minutes briefly discussing the advantages and challenges of developing this online testing and rating system and how the system has worked for testing and rating 400 students’ integrated reading-writing tasks. Some suggestions or modifications for future online testing, rating, and data collection will also be covered. The remaining 10 minutes will be for questions and comments from the audience.

Friday, June 24
12:10-12:40
Room 8 - Classroom Building Hall - 1st floor
University of Palermo, Scuola Politecnica - Faculty of Engineering

**Measuring Progress Using Adaptive and Automated Scoring Technologies**

*David Booth, Pearson*
*Paul Seddon, Pearson*

We would like to demonstrate a unique global test of English language proficiency which is designed to measure progress and give formative feedback to students and teachers. The test is delivered online via a learning management system (LMS) dedicated to English language learning, hosted on the Amazon Cloud and separately hosted in China on MyStack for global coverage. The test can be launched directly from outside of the LMS using LTI, a simple but standard method to establish a secure link from other applications to the test itself. Audio recording and automated scoring technologies are achieved via APIs and the test can be delivered to PCs, Macs and Android tablet devices. The test has twenty item types covering 4 language skills and 2 enabling skills and utilises an adaptive algorithm to optimise the assessment pathway. The item types are based on familiar tasks found in learning and assessment materials. They require the student to understand the language in context and process meaning in order to demonstrate language proficiency. Integrated skills tasks require students to process language using different language skills, for example listen and then write or read and then speak.
The test makes use of automatic scoring technologies in the scoring of speaking and writing. Questions are stored in a single IRT calibrated and tagged item bank. The tagging includes difficulty estimates, cando learning objective and topic. In short, the test operationalises a range of language constructs across a specified number of topics aligned to accurate measurement. There is a significant impact on learners, teachers and other stakeholders as the test provides the systematic tracking of progress against a global scale. Teachers and learners can drill down to find areas of strength or weakness or the data can be aggregated to show how groups of learners perform under different conditions. The test provides a benchmark which can help demonstrate the efficacy of learning materials and programmes. The demonstration is organised below. 2 minutes Introduction of the presenters and the aims of the presentation 5 minutes Description of the basic assessment and technology principles which underpin the test: calibrated and tagged item bank with 20 question types including integrated skills, adaptive algorithm functionality, automated scoring capabilities, interfaces of technology including APIs, scoring and grade book functionality. 8 minutes Demonstration of product using screen shots. Examples of item types which are used in the adaptive stage of the test alongside a specialised widget which shows the calculation of person ability and CSEM on the fly, the selection of the part 2 fixed form, examples of part 2 integrated test items which are scored by automated scoring engines, examples from the grade book which show score reporting for both learners and teachers. 5 minutes Demonstration of the product online. Demonstration of single sign on, teacher functionality, teacher and student information and data screens, formative feedback drill down for both students and teachers.
Castiddruzzu Miu

Oh Castiddruzzu miu ca ti scurdaru n capu na muntagnola abbannunata dun circu russu lu suli a lu scurari ti circunna e mpacci lu pasisi po taliari.

Seculi, dimmi, quantu nn a sfidatu cu ssi macigni di rocchi a sustintari, supirchjarii quantu nna vidutu nni ddru paiari ca ti voli scurdari.

Sicuru e fermu, livatu ni lantu, tu sienti lu vientu hiuhhiari e quarchi rocca chi ddraxsutta sempri cadi.

Ancora, bieddru miu, ca ncapu a stari comu n’aquila cu l’uocchji grifagni chi accuvacciata ncapu l’ova av’a cuvari.

My Little Castle

Oh, my dear castle that all have forsaken, sitting upon a mountain all alone, forlorn; at dusk a reddish circle glows around you and people from the town can gaze on you.

How many centuries have you withstood with your enormous rocks, the test of time? Tell me, how many wrongs, have you endured from that commune that wishes you were gone?

Solid and strong, raised high in our midst, You feel the hissing of the blowing wind And hear your stones that keep on crumbling down.

You must remain up there a long, long time, just like an eagle with rapacious eyes, sitting up on the eggs to hatch its brood.

Piero Carbone, Arba Sicula, Journal of Sicilian Folklore and Literature, Volume XXXIV, Number, 1 & 2, pp. 16-17.
Index of Presenters

Ahmadi, Alireza 35, 163
Alderson, Charles 29, 67
al-Shamrani, Abdulrahman 35, 164
Altamiro Conso, Douglas 30, 132
Argondizzo, Carmen 30, 135
Bachman, Lyle 28, 110
Baker, Beverly 28, 32, 76, 103, 105
Banerjee, Jayanti 34, 146
Batty, Aaron Olaf 33, 83
Beltrán, Jorge 27, 53
Benigno, Veronica 36, 37, 90, 98
Berry, Vivien 32, 79
Blood, Ian 34, 151
Bonk, William 33, 37, 98, 171
Booth, David 36, 175
Brau, Maria M. 37, 99
Bridgeman, Brent 32, 73
Broeker, Roderick 30, 134
Brooks, Rachel 37, 99
Brunfaut, Tineke 34, 152
Butler, Yuko 26, 50
Can, Duyen 30, 143
Carr, Nathan 30, 131
Chapelle, Carol 31, 69
Chapman, Mark 28, 33, 109, 172
Chen, Dajian 34, 147
Chen, Michelle Y. 30, 127
Chen, Sheng-Bo 36, 174
Cheng, Liying 33, 34, 111, 149
Cho, Yeonsuk 34, 35, 116, 151
Choi, Ikkyu 31, 33, 35, 68, 81, 116
Cox, Troy 30, 128
Csépes, Idikó 30, 136
Davidson, Fred 30, 143
Davis, Larry 34, 114, 117
de Backer, Fauve 29, 126
De Jong, John H.A.L. 12, 36, 37, 90
Decker, Laura 30, 128
D’Este, Claudia 34, 157
Deygers, Bart 28, 60
Dimova, Slobodanka 27, 51
Dixon, Daniel 29, 168
Do, Ha 30, 143
Doe, Christine 34, 149
Donovan, Anne 29, 64
Douglas, Scott 34, 149
Ducasse, Ana Maria 36, 87
Dunlea, Jamie 30, 129
Dunlop, Maggie 27, 37, 57, 100
Egbert, Jesse 30, 128
Elder, Catherine 36, 89
Elliott, Mark 29, 122
Evanini, Keelan 28, 167
Fairbairn, Judith 30, 129
Fan, Jinsong 34, 154
Fenara, Angela 30, 142
Flynn, Eleanor 36, 89
Fox, Janna 34, 160
Frost, Kellie 26, 49
Galaczi, Evelina D. 33, 169
Gass, Susan 29, 34, 35, 123, 148, 162
Gebril, Atta 34, 161
Germain, Joyce 28, 105
Grabowski, Kirby 37, 95
Grapin, Scott 29, 64
Green, Tony 30, 136
Gu, Lin 32, 33, 78, 170
Gutierrez Eugenio, Esther 30, 141
Haapakangas, Eeva-Leena 29, 67
Haghighi, Fatemeh Mohammadian 34, 155
Hamnes Carlsen, Cecilie 27, 55
Liu, Heidi Han-Ting 27, 28, 53, 167
Harding, Luke 34, 152
He, Lianzhen 34, 147
Hill, Kathryn 36, 87
Homayounzadeh, Maryam 35, 163
Hope, Amelia 32, 76
Hsieh, Ching-Ni 32, 35, 78, 117
Index of Presenters

Hsu, Annie 30, 138
Huhta, Ari 29, 67
Huisman, Annemiek 36, 89
Huynh, Tuan 30, 143
Ikeda, Naoki 36, 88
Inoue, Chihiro 27, 32, 53, 79
Isbell, Daniel R. 27, 54
Jacobs, Peter 28, 106
Jamieson, Joan 27, 54
Jang, Eunice Eunhee 29, 37, 65, 100
Janssen, Gerriet 28, 61
Jenkins, Jennifer 29, 63
Jenni, Barbara 28, 106
Jimenez, Jean 30, 135
Jin, Yan 33, 112
Kantor, Robert 35, 116
Kashimada, Yuko 30, 140
Kasper, Gabriele 25, 39
Keehner, Madeleine 33, 170
Kenyon, Dorry M. 25, 28, 45, 107, 108
Kim, Ah-young Alicia 28, 109
Kim, Hyun Jung 32, 75
Knoch, Ute 28, 31, 36, 69, 89, 104
Kremmel, Benjamin 36, 94
Kunnan, Antony 30, 33, 111, 131
Labben, Afef 37, 96
Lam, Wendy Y.K. 30, 127
Lau, Clarissa 29, 65
Lee, Jooseung 32, 75
Lee, Shinhye 229, 123
Lee, Yong-Won 33, 112
Leung, Constant 29, 63
Levi, Tziona 28, 62
Lewkowicz, Jo A. 29, 63
Lim, Gad 29, 122
Lin, Yu-Chen 36, 174
Llosa, Lorena 29, 64
Lockwood, Jane 32, 77
Lou, Yong 35, 164
Lu, Zhouyang 33, 145
Lucas, Ruth 30, 134
Ludbrook, Geraldine 34, 157
Lukácsi, Zoltán 30, 133
MacGregor, David 25, 28, 45, 107
Malone, Margaret E. 29, 64
Manias, Elizabeth 36, 89
May, L. D. Nicolas 27, 54
McIvor, Onowa 28, 106
McNamara, Tim 36, 89
Meier, Valerie 28, 61
Mieko, Debora 30, 132
Min, Shangchao 34, 147
Mollau, Pamela 35, 117
Montee, Meg 33, 172
Nakatsuahara, Fumiyo 27, 32, 52, 79
Nguyen, Anh 30, 143
Nguyen, Hoa 30, 143
Nguyen, Huu 30, 143
Nguyen, Thao 30, 143
Nieminen, Lea 29, 67
Norrbom, Bjorn 35, 164
Norris, John 10, 27
Norton, Jennifer 28, 109
Novák, Jakub 35, 116
O’Hagan, Sally 36, 93
O’Reilly, Tenaha 25, 35, 43, 115
O’Sullivan, Barry 30, 129
Ockey, Gary 33, 170
O’Connell, Stephen 34, 150
Oh, Saerhim 32, 74
Okabe, Yasuko 30, 140
Palmer, Adrian 29, 168
Papageorgiou, Spiros 33, 81
Patharakorn, Patharaorn 37, 97
Pellegrino, James W. 8, 31
Pham, Thuy 30, 143
Pill, John 36, 93
Plough, India 34, 146

Language Constructs, Contexts, and Content in Classroom and Large-Scale Assessments
Index of Presenters

Powers, Nancy 30, 134
Pulupa, Catherine 30, 140
Purpura, James E. 25, 27, 43, 53
Quevedo-Camargo, Gladys 34, 153
Ramineni, Chaitanya 32, 73
Raquel, Michelle 32, 36, 77, 173
Read, John 34, 158
Reed, Daniel 29, 35, 123, 162
Ricci, Roberto 25, 41
Rini, Danilo 29, 124
Robinson, Ian 30, 135
Rocca, Lorenzo 29, 121
Roever, Carsten 25, 39
Ross, Steven J. 25, 41
Sabatini, John P. 25, 43
Saif, Shahrzad 27, 59
Saville, Nick 30, 141
Sawaki, Yasuyo 33, 114
Scaramucci, Matilde 34, 153
Schedl, Mary 35, 115
Schmidgall, Jonathan 35, 117
Schmitt, Norbert 36, 94
Schmitt, Romain 27, 59
Seddon, Paul 36, 175
Sheehan, Kathleen 32, 72
Shin, Sun-Young 33, 82
Shohamy, Elana 28, 107
Sifakis, Nicos 30, 136
Sitajalabhorn, Woranon 27, 56
So, Youngsoon 31, 68
Stabler-Havener, Michelle
Stansfield, Charlie
Sullivan, Kristen 30, 130
Suzuki, Masanori 33, 37, 98, 171
Tan, May 30, 134
Taylor, Lynda 27, 52
Tejada-Sánchez, Isabel 34, 159
Timpe Laughlin, Veronika 28, 167
Trace, Jonathan 28, 61
Tran, Phuong 30, 143
Tsagari, Dina 30, 136
Tschirner, Erwin 35, 162
Tsutagawa, Fred 32, 74
Ullakonoja, Riikka 29, 67
Unger, Johann 34, 152
Unsworth, Sarah 33, 169
Urmston, Alan 36, 173
Van Avermaet, Piet 29, 126
Van Booven, Christopher D. 29, 64
Van Gorp, Koen 29, 34, 126, 148
Van Moere, Alistair 33, 171
van Naessens, Margaret 29, 119
VanPatten, Bill 34, 148
Vogt, Karin 30, 136
Voss, Erik 30, 138
Wagner, Maryam 34, 37, 100, 160
Wain, Jennifer 28, 167
Wang, Yuan 32, 36, 74, 117
Wei, Junli 30, 137
Wigglesworth, Gillian 28, 103, 104
Winke, Paula 29, 34, 123, 148
Wolf, Mkyung Kim 32, 74
Wong, Roxanne 36, 173
Woodward-Kron, Robyn 36, 89
Wu, Jessica 33, 84
Wu, Rachel Y. F. 33, 84
Wudthayagorn, Jirada 33, 113
Xi, Xiaoming 35, 114
Xu, Liu 29, 120
Yahalom, Sharon 36, 86
Youn, Soo Jung 36, 92
Yu, Xin 28, 110
Zhao, Cecilia Guanfang 32, 34, 70, 154
Zumbo, Bruno D. 30, 127

Language Constructs, Contexts, and Content in Classroom and Large-Scale Assessments
The Department of Languages and Culture at Universidad de los Andes, Colombia is pleased to welcome the 39th Language Testing Research Colloquium to Latin America.

July 17–21, 2017
Universidad de los Andes
Bogotá, Colombia
ISE: Integrated Skills in English

Assesses integrated skills in English – reading, writing, speaking and listening

- A contemporary multi-skills exam
- Focus on real language performance
- Promotes the development of transferable skills
- Builds communication skills for study and employability
- Flexibility to build qualifications over time
- Separate results for each skill shown on certificates
- Five levels from A2-C2 on the Common European Framework of Reference (CEFR)

ISE Foundation (A2)

ISE I (B1)

ISE II (B2)

ISE III (C1)

Reading & Writing

<table>
<thead>
<tr>
<th>ISE I (B1)</th>
<th>ISE II (B2)</th>
<th>ISE III (C1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 x Reading tasks</td>
<td>3 x Speaking tasks</td>
<td>3 x Speaking tasks</td>
</tr>
<tr>
<td>1 x Reading into writing task</td>
<td>1 x Listening task</td>
<td>1 x Listening task</td>
</tr>
</tbody>
</table>

ISE IV (C2) has a different format – see www.trinitycollege.com/ISE for details.
Check out the latest Language Testing titles from Routledge

Visit www.routledge.com/linguistics to learn more
Canada’s Leader in English Language Testing

As Paragon develops and produces its assessments, it strives for the following:

**INNOVATION**
**VALIDITY**
**FAIRNESS**
**ACCURACY**

Paragon encourages and financially supports collaborative research relationships with professional researchers and graduate students.

For more information, visit [www.paragontesting.ca/research](http://www.paragontesting.ca/research) or write to research@paragontesting.ca

A subsidiary of The University of British Columbia
IELTS Joint Funded Research Program 2016

Educational institutions and suitably qualified individuals are invited to apply for funding to undertake applied research projects in relation to IELTS for a period of one or two years. Projects will be supported up to a maximum of £45,000/AU$70,000.

Application deadline: 30 June 2016

Download an application form from: www.ielts.org/research-proposals
IELTS Research Grants – apply now!

IELTS Joint Funded Research Program 2016

Educational institutions and suitably qualified individuals are invited to apply for funding to undertake applied research projects in relation to IELTS for a period of one or two years. Projects will be supported up to a maximum of £45,000/AU$70,000.

Application deadline: 30 June 2016

Download an application form from: www.ielts.org/research-proposals

Recent Spaan Research projects are now available as CaMLA Working Papers

A Practical Guide to Investigating Score Reliability under a Generalizability Theory Framework
Chih-Kai Lin
Center for Applied Linguistics (CAL)
United States

The Characteristics of the Michigan English Test Reading Texts and Items and their Relationship to Item Difficulty
Khaled Barkaoui
York University
Canada

A Validation Study of the Reading Section of the Young Learners Tests of English (YLTE)
Paula Winke, Shinhye Lee, Irene Jieun Ahn, Ina Choi, Yaqiong Cui, Hyung-Jo Yoon
Michigan State University
United States

Linguistic Features in MELAB Writing Task Performances
YeonJoo Jung
Scott A. Crossley
Georgia State University
United States

Variability in the MELAB Speaking Task: Investigating Linguistic Characteristics of Test-Taker Performances in Relation to Rater Severity and Score
Geoffrey T. LaFlair
Northern Arizona University
United States

Shelley Staples
Purdue University
United States

Jesse Egbert
Brigham Young University
United States

Internship & Research Opportunities

Visit the CaMLA website to learn more about summer internships in language assessment. Funding is available on a limited basis for research related to CaMLA language tests. Grant funding is awarded under the Spaan Research Grant Program; project reports are published as part of the CaMLA Working Papers.
The British Council Assessment Research Awards and Grants Results for 2016

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. Winners for 2016 are:

- **Maria Georgina Fernandez Sesma** (University of Southampton, UK, supervisor Dr Ying Zheng)
- **Itthikar Haider** (University of Illinois at Urbana-Champaign, USA, supervisors Professor Emeritus Fred Davidson and Professor Melissa Bowles)
- **Benjamin Kremmel** (University of Nottingham, UK, supervisor Professor Norbert Schmitt)
- **Suh Keong Kwon** (University of Bristol, UK, supervisor Dr Guoxing Yu)
- **Heidi Han-Ting Liu** (Teachers College, Columbia University, USA supervisor Professor James E. Purpura)
- **Yueling Xu** (The University of Hong Kong, SARPRC, supervisor Professor David R. Carless)

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. Winners for 2016 are:

- **Stephen Bax & Prithvi Shrestha** (Open University, UK) for their project to explore lexical thresholds and lexical profiles across the Common European Framework of Reference for Language (CEFR) levels assessed in the Aptis test.
- **Nguyen Thi Thuy Minh & Ardi Marwan** (National Institute of Education, Nanyang Technological University, Singapore) for their project which seeks to analyse test-takers’ pragmatic performance and cognitive processing in the Aptis General Writing Test, Task 4.
- **Sally O’Hagan & Kevle Frost** (University of Melbourne) for their project which will examine test-taker processes and strategies and stakeholder perceptions of relevance of the Aptis for Teachers Speaking Test in the Australian context.
- **Parvaneh Tavakoli & Fumiyoshi Nakatsuura** (University of Reading) for their project which looks at the scorers validity of the Aptis Speaking Test: Investigating fluency across tasks and levels of proficiency.
- **Xun Yan, He Ram Kim & Ji Young Kim** (University of Illinois at Urbana-Champaign) for their project which explores the complexity, accuracy and fluency features of speaking performances on Aptis across different CEFR levels.

Innovation in Assessment Prize

The Innovation in Assessment Prize celebrates innovation in the area of language testing and assessment. The winner for 2016 is the:

**Language Academy, Universiti Teknologi Malaysia**, Johor Bahru, Malaysia, for their approach to developing an assessment instrument, the Test of English Communication Skills for graduating students (UTM-TECS), which measures university graduates’ readiness to communicate in English in the workplace. The tasks and assessment criteria of the test were derived from, and developed through, collaboration with industries at various stages of test development and implementation, including the validation and revision phases.

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 **Professor Emeritus Sauli Takala**.

Sauli Takala received his PhD from the University of Illinois at Urbana-Champaign in 1984. Sauli is now Professor Emeritus in Applied Linguistics at the University of Jyväskylä, Finland. He has extensive experience of research in language testing and assessment, in language policy and planning, curriculum development and teachers’ in-service education. For 15 years, he was on the Finnish Matriculation Examination Board. Sauli has published a large number of research reports and articles in Finnish, Scandinavian and international journals. He was editor of the Finnish Journal of Educational Research, co-editor of the Scandinavian Journal of Educational Research and is on the Editorial Board of Language Testing.

Sauli coordinated the IEA International Study of Writing in the 1980s and helped plan the EU-funded internet-based DIALANG diagnostic assessment project in the 1990s. For many years, he has been associated with the Council of Europe’s work on modern languages, most recently with the Common European Framework of Reference for Languages (CEFR), in particular the Manual for Relating Language Examinations to the CEFR (2009). He is a founding member of the European Association for Language Testing and Assessment (EALTA), served on its Executive Committee and was its second President in 2007. He is a consultant for the European Centre for Modern Languages (ECML).

Assessment Research Awards and Grants

Key dates for 2017

- **Call for proposals:** October 2016
- **Closing date:** 30 January 2017
- **Winners announced:** March 2017
Research for All

Studies in Language Testing

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

To find out more about our full list of publications:

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