LTRC 2006

LANGUAGE TESTING AND GLOBALISATION
ASIA PACIFIC PERSPECTIVES

28TH ANNUAL LANGUAGE TESTING RESEARCH COLLOQUIUM
OF THE INTERNATIONAL LANGUAGE TESTING ASSOCIATION

UNIVERSITY OF MELBOURNE
AUSTRALIA
JUNE 28 - JULY 1, 2006
Acknowledgements

LTRC 2006 Organising Committee
Annie Brown (Ministry of Education, UAE)
Cathie Elder (University of Melbourne)
Liz Hamp-Lyons (Hong Kong University/University of Nottingham)
Kathryn Hill (University of Melbourne)
Tom Lumley (Australian Council for Educational Research)
Tim McNamara (University of Melbourne)
Gillian Wigglesworth (University of Melbourne)

Conference Organisation co-ordinated by
Kerry Ryan

We would like to thank the following for their financial support:
International Language Testing Association (ILTA)
Educational Testing Service
University of Cambridge ESOL Examinations
Ordinate
Australian Council for Educational Research
National Centre for English Language Teaching & Research
Elsevier
The University of Melbourne

We would like to thank the following individuals for their generous assistance with organisational matters:
Sally Jones
Janna Fox
Glenn Fulcher
Dan Douglas
Meredith Izon
All the student volunteers

Proposal reviewers:
Lyle Bachman, Jay Banerjee, Vivien Berry, Geoff Brindley, Annie Brown,
Nathan Carr, Liying Cheng, Christine Coombe, Fred Davidson, Alan Davies,
Craig Deville, Barbara Dobson, Dan Douglas, Cathie Elder, Neus Figueras,
Janna Fox, Liz Hamp-Lyons, Kathryn Hill, Ari Huhta, Dorry Kenyon,
Antony Kunnan, Tom Lumley, Sari Luoma, Tim McNamara, Kieran O’Loughlin,
Jim Purpura, Qi Luxia, John Read, Pauline Rea-Dickins, Steve Ross, Elana Shohamy,
Guenther Sigott, Bernard Spolsky, Lynda Taylor, Randy Thrasher, Carolyn Turner,
Sara Weigle, Gillian Wigglesworth, Xiaoming Xi, Jin Yan
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Message from the ILTA President</td>
<td>3</td>
</tr>
<tr>
<td>ILTA – UCLES Lifetime Achievement Award 2006</td>
<td>5</td>
</tr>
<tr>
<td>LTRC 2006 Program Overview</td>
<td>7</td>
</tr>
<tr>
<td>Program Details</td>
<td>8</td>
</tr>
<tr>
<td>Tuesday, June 27</td>
<td>8</td>
</tr>
<tr>
<td>Wednesday, June 28</td>
<td>8</td>
</tr>
<tr>
<td>Thursday, June 29</td>
<td>9</td>
</tr>
<tr>
<td>Friday, June 30</td>
<td>11</td>
</tr>
<tr>
<td>Saturday, July 1</td>
<td>13</td>
</tr>
<tr>
<td>Abstracts</td>
<td>15</td>
</tr>
<tr>
<td>2006 Samuel J. Messick Memorial Lecture</td>
<td>15</td>
</tr>
<tr>
<td>Symposium 1</td>
<td>16</td>
</tr>
<tr>
<td>Symposium 2 (parallel session)</td>
<td>17</td>
</tr>
<tr>
<td>Symposium 3</td>
<td>18</td>
</tr>
<tr>
<td>Papers and Posters</td>
<td>20</td>
</tr>
<tr>
<td>Session 1</td>
<td>20</td>
</tr>
<tr>
<td>Session 2</td>
<td>23</td>
</tr>
<tr>
<td>Session 3a</td>
<td>27</td>
</tr>
<tr>
<td>Session 3b</td>
<td>29</td>
</tr>
<tr>
<td>Session 4</td>
<td>31</td>
</tr>
<tr>
<td>Session 5</td>
<td>32</td>
</tr>
<tr>
<td>Session 6</td>
<td>36</td>
</tr>
<tr>
<td>Session 7</td>
<td>38</td>
</tr>
<tr>
<td>Session 8</td>
<td>41</td>
</tr>
<tr>
<td>Session 9</td>
<td>43</td>
</tr>
<tr>
<td>Research Networks</td>
<td>45</td>
</tr>
</tbody>
</table>
Message from the ILTA President

Dear Friends and Colleagues,

On behalf of the Executive Board and all the members of the International Language Testing Association, I would like to welcome you to the 2006 Language Testing Research Colloquium in Melbourne.

The conference organizers have put together an excellent programme for us, from the pre-conference workshops to papers, posters, and symposia. There are also a host of meetings and fun things to do, where we have the opportunity to meet with the people with whom we share our professional concerns. So much work goes into the organization of our annual conference behind the scenes, so I would also like to thank the conference organizers on behalf of us all.

This year there are a lot of issues to talk through at the Annual Business Meeting of ILTA. Of critical importance will be the discussion on how we should move forward with the ILTA Code of Practice, which Alan Davies will be leading. Alan has been at the forefront of the development of the Code of Practice as well as the Code of Ethics, and we hope that there will be a lively discussion that will inform the way forward with this important project.

ILTA is also proud to be able to support long established members of our profession and newcomers to the field. This year we are once again organizing the Lifetime Achievement Award in cooperation with Cambridge ESOL, which this year goes to John Clark. ILTA is also supporting a number of students to travel to LTRC through the travel awards, and will be giving the Lado Award for the best student paper at LTRC. Finally, the Award for the Best Paper in 2004 will be made at the conference banquet.

There are so many reasons to renew ILTA membership, or to become a member if you have not already done so. ILTA continues to support professional development activities, and this year will again make two Workshop Awards to support language testing activities that would otherwise not take place through lack of funds. The Electronic Newsletter and the new website (in the process of being updated) now enter their second year of operation and provide news about language testing to
members and the public. During the rest of the year we hope that the new electronic database of members and the voting system will finally be bedded in and provide a much easier and effective way for us all to maintain our membership and take part in the democratic life of the organization.

ILTA embodies the professional values that we hold dear, and which draw us to LTRC whenever we are able to attend. Membership is a sign of being involved in this professional community with which we are so proud to be associated.

I hope that you have an exceptionally enjoyable and productive conference here in Melbourne.

With best wishes,

Glenn Fulcher
President of ILTA, 2006
ILTA – UCLES Lifetime Achievement Award 2006

Awarded to
Dr John L. D. Clark
(emeritus Dean of Program Evaluation, Research, and Testing at the Defense Language Institute)

John Clark has been involved full time in language testing since his graduation from Harvard in 1967, where he carried out empirical studies related to the teaching of French pronunciation to American students at the Laboratory for Research in Instruction with John B. Carroll. Until 1981, he was an Examiner at the Educational Testing Service in Princeton, for the next five years he directed the Foreign Language Education Division at the Center of Applied Linguistics in Washington DC, and after that until his retirement in 1997 he was at the Defense Language Institute, an institution with a good claim to be the premier foreign language teaching organization. While thus busy in developing and administering language tests, he was constantly engaged in research in the field. During his career, he directed a number of major projects, being responsible, among other things, for the adaptation of the FSI Oral Proficiency Interview to use for the Peace Corps and the later development of the Test of Spoken English, the preparation of tests of Chinese and Japanese for English-speaking learners, and the creation of a procedural model and testing instruments for the diagnostic measurement of proficiency growth, maintenance and attrition for students in four major languages.

As a result of his skills and experience, he has been consulted by a wide range of testing bodies and educational institutions in the US and abroad, including Morocco, The Ivory Coast, The Philippines, Denmark, and England; he was a member of the advisory committee for the UCLES-TOEFL Comparability Study, and of many similar committees. He has spoken regularly at national and international meetings (the latest the plenary speech at ECOLT in 2005), and conducted workshops and training sessions internationally (including in a score of countries for the Peace Corps from 1970-80). In this way, his wisdom and experience in language testing have been widely spread.

His review of the Pennsylvania Foreign language project in 1969 made clear the high standards he advocated for educational research, and set the tone for his continued insistence on rigorous design and solid empirical testing of ideas in the field. His 1972 survey of language testing (Clark, John L.D. 1972, Foreign language testing: theory and practice, Philadelphia, PA: Center for Curriculum Development) remains a classic in the field. In 1973, he co-authored a Handbook on language evaluation which was published by the Central Institute of Indian Languages. His edited volume on Direct Testing of Speaking Proficiency led to the opening up of the FSI Oral Proficiency Interview to wider use and the development of the Simulated Oral Proficiency Interview. Another study, in 1981, showed the strong correlation between self-assessment and language measures. Similarly, his idea of describing proficiency on CAN-DO scales has formed the basis for European work on language proficiency,
as well as a being a key step in the development of the ACTFL proficiency scales and of the growth of the “Proficiency movement.” He has spoken and written on foreign language instruction, computer assisted instruction, and evaluation.

A practical language tester throughout his career, his practice has been constantly backed by strong theory. This was well captured in a note he sent to LTEST-L three years ago which reads in part: “It's not possible easily to summarize the twenty-plus years of effort on the part of both ACTFL and the federal agencies in developing and administering their interview-based testing procedures, and I would suggest not bashing them inordinately unless and until one has walked around the block a time or two in their research and operational shoes. This having been said, there remain a multitude of additional empirical studies that can and should be carried out on oral interview testing in all of its diverse aspects. I would suggest that these efforts would be more suitably focused and would have a greater empirical yield if we could manage to 'lose the definite article' in favor of the descriptively more accurate and much more highly productive gerund. 'The oral proficiency interview is dead; long live oral proficiency interviewing!'”

John has also served and guided the language testing community through his membership on national and international professional committees and as a President of the International Language Testing Association.

This tribute was written by Bernard Spolsky.
<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tuesday, June 27</strong></td>
<td></td>
</tr>
<tr>
<td>9:00 - 12:00</td>
<td>Workshop 1. Rasch...the basics</td>
</tr>
<tr>
<td>12:00-1:30</td>
<td>Lunch</td>
</tr>
<tr>
<td>1:30 - 4:30</td>
<td>Workshop continues</td>
</tr>
<tr>
<td><strong>Wednesday, June 28</strong></td>
<td></td>
</tr>
<tr>
<td>9:00 - 12:00</td>
<td>Morning session</td>
</tr>
<tr>
<td>12:00-1:15</td>
<td>Lunch</td>
</tr>
<tr>
<td>1:15 - 4:15</td>
<td>Workshops continue</td>
</tr>
<tr>
<td>4:30 - 6:00</td>
<td>Symposium 1. Assessing Indigenous Language</td>
</tr>
<tr>
<td>6:00-8:00</td>
<td>Welcoming Reception</td>
</tr>
<tr>
<td><strong>Thursday, June 29</strong></td>
<td></td>
</tr>
<tr>
<td>8.30-8.50</td>
<td>Welcome</td>
</tr>
<tr>
<td>8.50-9.50</td>
<td>2006 Samuel J. Messick Memorial Lecture</td>
</tr>
<tr>
<td></td>
<td>Professor Mark Wilson</td>
</tr>
<tr>
<td>9:50-10:00</td>
<td>Presentation of the Messick Award</td>
</tr>
<tr>
<td>10:00-10:25</td>
<td>Break</td>
</tr>
<tr>
<td>10:25-12:10</td>
<td>Session 1 (3 papers + 3 posters)</td>
</tr>
<tr>
<td>12:10-1:40</td>
<td>Lunch Lang. Ass't Quarterly EAB Meeting</td>
</tr>
<tr>
<td>1:40-3:30</td>
<td>Session 2 (3 papers + 4 posters)</td>
</tr>
<tr>
<td>3:30-3:50</td>
<td>Break</td>
</tr>
<tr>
<td>3:50-5:50</td>
<td>Session 3a (4 papers) Parallel Session</td>
</tr>
<tr>
<td>5:50-6:20</td>
<td>Session 3b (4 papers) Parallel Session</td>
</tr>
<tr>
<td>6:45-8:30</td>
<td>Poster Displays</td>
</tr>
<tr>
<td><strong>Friday, June 30</strong></td>
<td></td>
</tr>
<tr>
<td>8.30-10.00</td>
<td>Symposium 2. Social Responsibilities of Language Testers</td>
</tr>
<tr>
<td></td>
<td>Session 4 (3 papers) Parallel Session</td>
</tr>
<tr>
<td>10.00-10.20</td>
<td>Break</td>
</tr>
<tr>
<td>10.20-10.35</td>
<td>Group Photograph</td>
</tr>
<tr>
<td>10.35-11.50</td>
<td>Research Networks Session</td>
</tr>
<tr>
<td>11.50-1.15</td>
<td>Lunch Language Testing EAB Meeting</td>
</tr>
<tr>
<td>1.15-3.05</td>
<td>Session 5 (3 papers + 4 posters)</td>
</tr>
<tr>
<td>3.05-3.45</td>
<td>Poster Displays and Afternoon Tea</td>
</tr>
<tr>
<td>3.45-5.45</td>
<td>Session 6 (4 papers)</td>
</tr>
<tr>
<td>7.00-11.00</td>
<td>Banquet &amp; Awards Presentation</td>
</tr>
<tr>
<td><strong>Saturday, July 1</strong></td>
<td></td>
</tr>
<tr>
<td>8.30-10.20</td>
<td>Session 7 (3 papers + 4 posters)</td>
</tr>
<tr>
<td>10.20-11.00</td>
<td>Poster Displays and Morning Tea</td>
</tr>
<tr>
<td>11.00-12.30</td>
<td>Session 8 (3 papers)</td>
</tr>
<tr>
<td>12.30-2.30</td>
<td>Lunch ILTA Business Meeting</td>
</tr>
<tr>
<td>2.30-4.00</td>
<td>Session 9 (3 papers)</td>
</tr>
<tr>
<td>4.00-4.20</td>
<td>Break</td>
</tr>
<tr>
<td>4.20-5.50</td>
<td>Symposium 3. What role for native speaker norms in second/foreign language testing?</td>
</tr>
<tr>
<td>5:50</td>
<td>Best student paper award. Closing Remarks</td>
</tr>
</tbody>
</table>
Program Details

Tuesday, June 27

Workshop 1: Rasch...the basics
Tim McNamara

Location: Nanson Lab 212 - Mathematics and Statistics Dept
Richard Berry Building

9.00am-4.30pm

Wednesday, June 28

Workshop 2: ConQuest
Ray Adams & Margaret Wu

Location: Nanson Lab 212 - Mathematics and Statistics Dept
Richard Berry Building

9.00am-4.15pm

Workshop 3: The development and implementation of scale-based assessment in Australian schools
Helen Moore, Angela Scarino & Sophie Arkoudis

Location: Room 509, Arts Centre (Cnr Grattan & Swanston Sts)

9.00am-4.15pm

Symposium 1. Asssessing Indigenous Languages
Convenor: Gillian Wigglesworth

Location: University House

4.30pm-6.00pm

Welcoming Reception – SPONSORED BY Cambridge ESOL

Location: University House

6.00pm-8.00pm
<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.30-10.00</td>
<td>Opening Session</td>
<td>Law Building – GM15</td>
</tr>
</tbody>
</table>
| 8.30-8.50  | Welcome                                      | Gillian Wigglesworth, University of Melbourne  
              |                               | Glyn Davis, Vice-Chancellor, University of Melbourne |
| 8.50-9.50  | 2006 Samuel J. Messick Memorial Lecture      | Building out the measurement model  
              |                               | to incorporate complexities in language testing  
              |                               | Professor Mark Wilson, UC Berkeley |
| 9.50-10.00 | Presentation of the Messick Award by Cathie Elder, Chair, TOEFL Committee of Examiners, Educational Testing Service (ETS) |
| 10.00-10.25| Morning Tea sponsored by NCELTR              |                           |
| 10.25-12.10| Session 1                                    |                           |
| Papers 1 - 3|                                              |                           |
| 1. Raters’ Perceptions of Scoring Criteria in Writing and Speaking Performance Assessments | Thomas Eckes |
| 2. Exploring rater and task variability in second language oral performance assessment | Youn-Hee Kim |
| 3. Re-training writing raters online: how does it compare with face-to-face training? | Ute Knoch, John Read & Janet von Randow |
| Posters 1 - 3|                                              |                           |
| 1. This Test is Unfair: I’m Too Old. He’s Too Young | Jeff S. Johnson |
| 2. Introducing audit trails to the world of language testing | Jinshu Li, Hyeong-Jong Lee, Jung Tae Kim & Carsten Wilmes |
| 3. Examining Generalizability of Speaking Performance Measures Across Tasks | Rie Koizumi |
| 12.10-1.40 | LUNCH                                       | Law Building – Foyer      |
|            | Language Assessment Quarterly EAB Meeting    | Law Building – 101        |
|            | Graduate House                              |                           |
**Papers 4 - 6**

4. A longitudinal approach to the validation of a standards-based classroom assessment of English
Lorena Llosa

5. Effect Sizes in Using Test Accommodations To Improve ELLs' Achievement Test Performance
Maria Pennock-Roman & Charlene Rivera

6. Comparing Dichotomous and Polytomous Scoring Procedures on Multiple-Choice and Gap-Fill Tasks of Grammatical Knowledge
James E. Purpura

**Posters 4 - 7**

4. The TEPOLI test: construct and parameters to assess EFL teachers' oral proficiency
Douglas Altamiro Consolo & Vera Lúcia Teixeira da Silva

5. Developing a test for interpreter certification in Norway
Diane Mortensen

6. Development and Pilot Testing of an Interview Test in Multiple Languages
Nathan T. Carr, Hélène Domon, Janet L. Eyring, Jack (Jinghui) Liu, Michael S. Ray, Setsue Shibata, Marjorie Tussing

7. A Pre-Instruction EFL Test for the Airline Industry
Fred Davidson, Hyeong-Jong Lee, Jinshu Li, Jungtae Kim, Carsten Wilmes & Sarah Blayer

---

**3.30-3.50**

<table>
<thead>
<tr>
<th>Afternoon Tea</th>
<th>Law Building – 1st Floor</th>
</tr>
</thead>
</table>

**3.50-5.50**

<table>
<thead>
<tr>
<th>Session 3a (parallel)</th>
<th>Law Building – GM15</th>
</tr>
</thead>
</table>

**Papers 7 - 10**

7. DIF in a test of ESL pragmatics
Carsten Roever

8. The impact of test delivery modes and test-taker characteristics on speaking test performance
Eunjyu Yu

9. Task Difficulty in Semi-direct Speaking Tests
Jessica R W Wu

10. Intra-task comparison in a monologic oral performance test: the impact of task manipulation on performance
Tomoko Horai
### Session 3b (parallel)  
Law Building - Theatre 102

#### Papers 11 - 14

11. English language exams reform project in India: From achievement to proficiency testing  
   Antony John Kunnan & Rama Mathew

12. A model for investigating the impact of language assessment within a national educational context  
   Nick Saville

13. Applying the ‘diffusion-of-innovation’ framework to an examination reform in China  
   Qi Luxia & Liu Qingsi

14. An impact study of Taiwan’s General Proficiency English Test (GEPT)  
   Viphavee Vongpumivitch

#### Poster Displays  
Law Building – 1st Floor

#### Wine Tasting  
Graduate House

---

**Friday, June 30**

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
</table>
| 8.00am-6.00pm | Registration:  
Publishers’ Exhibits: | Law Building – Foyer  
9.00am-6.00pm | Law Building – 101 |
| 8.30-10.00 | Symposium 2 (parallel)  
The Social Responsibilities of Language Testers  
Dan Douglas  
Tim McNamara  
Mike Milanovic  
Discussant | Law Building – GM15 |
| 8.30-10.00 | Session 4 (parallel) | Law Building – Theatre 102 |

#### Papers 15 - 17

15. The Impact of ESL Placement Test Decisions  
   Young-Ju Lee

16. Predictors of Item Difficulty in an Automated Spoken English Test  
   Don Salting & Jennifer Balogh

17. Typical Error Patterns for Different L1 Groups in ESL Test-Takers’ Essays  
   Yong-Won Lee, Martin Chodorow & Claudia Gentile

10.00-10.35 | Morning Tea & Group Photograph | Law Building – 1st Floor |
<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.35-11.50</td>
<td>Research Networks Session</td>
<td>Law Building – G27 and G29</td>
</tr>
<tr>
<td>11.50-1.15</td>
<td>LUNCH – Volunteer-led trip to Victoria Market Language Testing EAB Meeting</td>
<td>Graduate House</td>
</tr>
<tr>
<td>1.15-3.05</td>
<td>Session 5</td>
<td>Law Building – GM15</td>
</tr>
</tbody>
</table>

**Papers 18 - 20**

18. The Trait Structure of Oral Language Ability in a Computer-based Speaking Test  
Sutthirak Sapsirin

19. Learner strategies and knowledge in TOEFL Listening Tasks: A validation study  
Dan Douglas & Volker Hegelheimer

20. Validating the GEPT: Lexico-Grammatical Ability as a Predictor of L2 Listening Ability  
Yen-Fen Liao

**Posters 8 - 11**

8. Academic English for School-age Learners: Research and Standards-informed Test Development  
Alison L. Bailey, Becky Hsuan-Hua Huang, Christine A. Ong & Frances A. Butler

9. Quantitative investigation into task type effects on listening test performance  
Yo In’nami

10. Validating the revised IELTS speaking test  
Annie Brown

11. What’s the problem? Technology and communication as mediating factors in test implementation  
Janna Fox

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.05-3.45</td>
<td>Poster Displays &amp; Afternoon Tea</td>
<td>Law Building – 1st Floor</td>
</tr>
<tr>
<td>3.45-5.45</td>
<td>Session 6</td>
<td>Law Building – GM15</td>
</tr>
</tbody>
</table>

**Papers 21 - 24**

21. De-construct-ing individual and paired oral proficiency testing  
Lindsay Brooks

22. Validating the group oral discussion task: The effects of personality type  
Gary J. Ockey

23. “Effective interaction” in a paired candidate EAP speaking test  
Lyn May

24. An empirically-based rating scale for 'interaction' in a paired oral test  
Ana María Ducasse

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.00-11.00</td>
<td>Banquet &amp; Awards Presentation</td>
<td>Bluestone Restaurant 349 Flinders Lane, Melbourne</td>
</tr>
</tbody>
</table>
Saturday, July 1

Registration: 8.00am-6.00pm  
Law Building – Foyer

8.30-10.20  Session 7  
Law Building – GM15

Papers 25 - 27

25. Extracting meaningful speech features to support diagnostic feedback: an ECD approach to automated scoring  
Xiaoming Xi, Klaus Zechner, and Isaac Bejar

26. Washback in an embedded assessment task  
Robyn Spence-Brown

27. Some implications of student reading assessments in PNG and the Philippines  
Prue Anderson

Posters 12 - 15

12. Helping ESOL teachers transcend borders: the case of Teaching Knowledge Test  
Hanan Khalifa & Nick Saville

13. Recognition Speed for Basic Vocabulary  
David Coulson

14. The effect and validity of a multiple-choice and open-ended listening test  
Akiyo Hirai

15. Potential ‘inequalities’ in language tasks and human brain imaging techniques  
Toru Kinoshita & Harumi Oishi

10.20-11.00  Poster Displays & Morning Tea  
Law Building – 1st Floor

11.00-12.30  Session 8  
Law Building – GM15

Papers 28 - 30

28. Testers’ intention, specialists’ presumption, and Stakeholders’ perception of reading test Items  
Hossein Farhady & Masoumeh Shahmohammadi

29. Criterial parameters for establishing writing proficiency levels within the Common European Framework: a retrospective analysis  
Cyril Weir & Sara Gysen

30. Group dynamics in language test development  
Jung Tae Kim, Hyeong-Jong Lee, Jinshu Li & Carsten Wilmes
<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.30-2.30</td>
<td>Lunch &amp; ILTA Business Meeting</td>
<td>Law Building – Theatre 102</td>
</tr>
<tr>
<td>2.30-4.00</td>
<td>Session 9</td>
<td>Law Building – GM15</td>
</tr>
<tr>
<td></td>
<td><strong>Papers 31 - 33</strong></td>
<td></td>
</tr>
<tr>
<td>31.</td>
<td>Indigenous assessment criteria: Refining the concept</td>
<td></td>
</tr>
<tr>
<td></td>
<td>M. Usman Erdősy</td>
<td></td>
</tr>
<tr>
<td>32.</td>
<td>The use of IELTS for university selection: an Australian study</td>
<td>Kieran O’Loughlin</td>
</tr>
<tr>
<td>33.</td>
<td>Planes, Politics, &amp; Oral Proficiency: Testing International Air Traffic Controllers</td>
<td>Carol Lynn Moden &amp; Gene B. Halleck</td>
</tr>
<tr>
<td>4.00-4.20</td>
<td>Afternoon Tea</td>
<td></td>
</tr>
<tr>
<td>4.20-5.50</td>
<td>Symposium 3</td>
<td>Law Building – GM15</td>
</tr>
<tr>
<td></td>
<td><strong>What role for native speaker norms in second/foreign language testing?</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pauline Foster</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Micheline Chalhoub-Deville</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Luke Harding and Gillian Wieglesworth</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cathie Elder</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Discussant: Alan Davies</td>
<td></td>
</tr>
<tr>
<td>5.50</td>
<td><strong>Best Student Paper Award</strong></td>
<td>Law Building – GM15</td>
</tr>
<tr>
<td></td>
<td>Closing Remarks</td>
<td></td>
</tr>
</tbody>
</table>
Abstract
This presentation gives an overview of a new and integrated way to think about item response models in language testing. This new framework allows the domain of item response models to be coordinated and broadened to emphasize their explanatory uses beyond their standard descriptive uses. The basic explanatory principle is that item responses can be modeled as a function of predictors of various kinds. The predictors can be (a) characteristics of items, of persons, and of combinations of persons and items; they can be (b) observed or latent (of either items or persons); and they can be (c) latent continuous or latent categorical. Hence, a broad range of models can be generated within the framework of explanatory item response models, including many extant item response models as well as many new ones. Note that the "item responses" that we are referring to are not just the traditional "test data," but are broadly conceived as categorical data from a repeated observations design. Hence, data from studies with repeated observations experimental designs, or with longitudinal designs, may also be modeled. The presentation starts with a language testing data set, and the explanatory approach is illustrated by describing several different research questions, involving both person and item predictors. These examples are used to substantiate the different sorts of measurement models that embody the different types of explanations.

Background
Mark Wilson is a professor in the Graduate School of Education at UC Berkeley. His interests focus on measurement and applied statistics. His work spans a range of issues in measurement and assessment from the development of new statistical models for analyzing measurement data, to the development of new assessments in subject matter areas such as science education, patient-reported outcomes and child development, to policy issues in the use of assessment data in accountability systems. He has recently published three books: one, entitled Constructing measures: An item response modeling approach (Erlbaum), is an introduction to modern measurement; the second (with Paul De Boeck of the University of Leuven in Belgium), entitled Explanatory item response models: A generalized linear and nonlinear approach (Springer-Verlag), introduces an overarching framework for the statistical modeling of measurements that makes available new tools for understanding the meaning and nature of measurement; the third, entitled Towards coherence between classroom assessment and accountability (University of Chicago Press--National Society for the Study of Education) is an edited volume that explores the issues relating to the relationships between large-scale assessment and classroom-level assessment. He has also recently chaired a National Research Council committee on assessment of science achievement—Systems for state science assessment (with Meryl Bertenthal of the NRC). He is founding editor of a new journal: Measurement: Interdisciplinary Research and Perspectives.
Symposia

Symposium 1

Wednesday, June 28
4.30pm – 6.00pm
University House

Assessing Indigenous Languages

Convenor: Gillian Wigglesworth

Discussant: Nick Evans

National assessment of Indigenous language vitality
Nick Thieberger, University of Melbourne

Within the past 15 years there have been a number of projects aiming to enumerate speakers of Australian Indigenous languages, ranging from the local to the national level, including: the ABS Censuses; the ABS National Indigenous Surveys; the State of Environment Report; the National Indigenous Languages Survey; and several regional surveys. In this talk I will summarise their findings and show that, despite considerable shortcomings, such surveys can be cross-correlated to be slightly less blunt instruments than they initially appear to be.

Measuring receptive skills for traditional languages
Caroline Jones, University of New South Wales

In Aboriginal communities of the central Northern Territory, where traditional languages are endangered, children and younger adults appear more often to be listeners than speakers of traditional languages. In assessing children's existing proficiency in traditional language for the purposes of guiding implementation of traditional language programs, this paper suggests that measurement of receptive rather than productive language skills can be a practical and informative option. The paper discusses data from a range of comprehension tasks requiring little or no verbal response, collected from primary school children from a central Northern Territory Aboriginal community. Relative strengths and weaknesses of each assessment task are discussed, with reference to apparently inherent aspects of the task and also with reference to what can be concluded from the response data collected.

Elicited production techniques for examining young indigenous children's speech.
Carmel O'Shannessy, Max Planck Institute and University of Sydney

Language environments in indigenous communities are often complex, with adults and children speaking several languages and often more than one variety of each. Assessment of children's language learning in these environments is necessary for educational purposes and so that adults in the communities have an accurate and objective account of how their children speak. I will show how elicited production methods can target particular language properties and provide oral texts from children that are useful for school assessments, language researchers and community members. I will demonstrate techniques that encourage children to speak in both their weaker and stronger languages, which is necessary for an accurate picture of the children's repertoires.
Symposium 2 (parallel session)

Friday, June 30
8.30am-10.00am
Main Lecture Theatre (GM15) - Law Building

The social responsibilities of language testers

Testing news in the United States and China and elsewhere is now dominated by discussions about test use. Language testers too no longer debate overall language proficiency, or the cloze, but now wonder about how tests are used. After a period in which we acknowledged the power of tests for evil (and for good), we now prepare statements of ethical standards and a code of good practice. Do we need to go further? Is it enough to be sure that the tests we work on are the best that can be produced? How responsible are we for the way that results are used? In this symposium, a group of experienced language testers will consider this question as it applies to high-stakes tests -- language tests which have serious consequences for the test taker. Are our social responsibilities to the test taker, or to the test user, or to society? Remembering well-known issues -- the regular use of language tests in order to enforce curriculum changes, or to exclude unwanted immigrants and asylum seekers, or to restrict job opportunities --, we may reasonably ask (as did the physicists who worked on the atomic bomb) if we are responsible for what uses others make of what we produce. Can we just stop writing tests? Or if we accept wider social responsibility, is the time ripe to take the further steps that the founders of ILTA dreamed of and organize our profession to attempt to influence the uses of language testing.

Dan Douglas (Iowa State University)

I will discuss the use of tests in academic settings for the overtly innocuous purpose of assessing the English proficiency of international teaching assistants (ITAs), using as an example a performance test administered to all new ITAs at an American university. Complex ethical issues are centered on unofficial test use, wherein some departments require that teaching assistants achieve a requisite proficiency level on the performance test in order to stay in their doctoral programs, thus putting enormous pressure on the students, who in turn put pressure on the test administrators and raters.

Tim McNamara (University of Melbourne)
Test use: beyond validity?

Recent discussions have focussed on whether test use is or is not part of validity, and hence part of the responsibility of testers, and have increasingly concluded (Messick, Kane, Bachman) that it is. However, it is questionable whether current discussions of test use within validity theory are adequate for conceptualising the social functions of tests, which go beyond the immediate, explicit reasons for their use. The paper argues that we need better social theoretical models of the context of testing in order to understand the issue, which is discussed in the context of the experience of a small group of Canadian Anglophone executive level civil servants who are required to take the Public Service Commission test of spoken French. Their perceptions are examined in the light of, first, Bachman and Palmer's framework of test usefulness, and then a broader social theoretical understanding of the role of examinations in the construction of identity, drawing on Foucault.

Mike Milanovic (University of Cambridge ESOL Examinations)

Concern for access to opportunity based on merit rather than patronage led to the establishment of examination systems in Britain in the nineteenth century. In the 1850s the Oxford Delegacy and the Cambridge Syndicate were set up to provide external verification of
standards being achieved in schools. Educational achievement, social justice and fairness have continued to be guiding principles in the British approach. The presentation explains Cambridge’s role in promoting good practice in developing tests which are “fit for purpose” and in researching the impact that such tests have on the educational and social environments in which they operate.

Symposium 3

Saturday, July 1
4.20pm-5.50pm
Main Lecture Theatre (GM15) – Law Building

What role for native speaker norms in second/foreign language testing?

Convenor: Cathie Elder

Discussant: Professor Alan Davies

In traditional testing practice the native speaker has served as benchmark both for standard setting and test validation purposes. This practice is now being queried for a number of reasons: first, the quality of native speaker performance varies widely raising questions about which model should serve as target for the language learner (McNamara 1996, Hamilton et al 1993); second, English has established itself as global lingua franca with its nonnatives now outnumbering the traditional native speaker owners (Crystal 1997, Graddol 1997) and using new modes of English communication in previously unimagined contexts; third, global communication flows have made the boundaries between different native languages more porous such that hybrid forms and codeswitching between languages is the norm rather than the exception (Canagarajah, 2005). This symposium addresses the issue of native/non-nativeness in language testing from a variety of perspectives, posing some of the following questions: Do NS and NNS judges differ in their rating behaviour? How do attitudes to and familiarity with non-native accents influence judgements of L2 oral proficiency? What are the implications of English as a lingua franca (ELF) discourse findings for language test design? Is the NS still the norm of choice for language tests? Are there other viable norms?

"They went to a coffee shop and they drank a coffee/ they stopped off at a pub and had a couple of beers" Assessing learners’ performance using native-speaker data.
Pauline Foster (St Mary’s College)

This paper discusses the findings of a research project comparing native- and non-native speaker performance on four narrative tasks. With no investment in task results and no particular desire to display language proficiency, native speakers can be laconic and even grammatically careless, so that a useful native baseline performance might well require some coaching. However, the native speaker data was valuable in showing clearly that the lexical selections of some learners (those based in Britain) were more like native speaker selections than those of other learners (those based in Iran). Additionally, very few of the native-speakers misread any of the picture stories, and the learners in London were far less likely to fall into the cultural booby traps hidden in them than were their counterparts in Iran. The performance of the learners in London could therefore be said to be enhanced both by the idiomatic language they had absorbed through exposure to the lexical selections of the target language community, and by their familiarity with British cultural artefacts. How should we judge the learners in Iran whose grammar was no worse, but whose lexical selections were unidiomatic and who couldn’t tell a drawing of a pub from one of a coffee shop?
More research to debunk the myth of the native speaker as a rater
Micheline Chalhoub-Deville (University of North Carolina)

The present study continues a line of research that investigates whether there is a common or shared perception of English language proficiency among ESL teacher raters. The basic question we seek to address is whether native speaking teachers with different educational background and teaching experience differ in their rating of the oral performance of English language students. The present study draws on TOEFL Test of Spoken English (TSE) data previously collected by Chalhoub-Deville and Wigglesworth (2005) and ratings of native speaking teachers from different English speaking countries. Findings, which show significant variation in rater judgement depending on level of education and teaching experience, are used to query the notion of an ideal native speaker rater as benchmark for second language assessment.

Is there a role for a diversity of accents in academic listening assessment?
Luke Harding and Gillian Wigglesworth (University of Melbourne)

While there has been considerable debate over whose “norms” should have primacy in second/foreign language assessment standards, the accent norms of speakers on listening tests have not been widely discussed. The default position for large-scale academic tests of English such as IELTS and TOEFL has been to maintain standard accents, or, more recently, to include a range of inner-circle varieties (Kachru, 1985; 1992). However, the consideration of which accent varieties are acceptable on a listening test feeds into larger issues of authenticity, fairness and construct validity. In this talk, we will give a brief overview of the central issues facing test designers in the use of non-native speaker accents on listening assessment. We will then provide a rationale for a broader representation of English accents including both native varieties and non native varieties in academic listening tests, and discuss some implications for future language testing research and practice.

English as a Lingua Franca: Imagining the test construct
Cathie Elder (University of Melbourne)

This paper proposes two alternative models for assessing English as a Lingua Franca (ELF). Tests based on the first model resemble existing approaches to assessing English as a foreign language offered by such tests as TOEFL, IELTS. This model is based on standard “native-like” English. What distinguishes tests of this model from existing international test of English is that it explicitly allows test accommodations. Such accommodations modify the test delivery system in order to make it accessible and fair for ELF users without changing the construct. Tests based on the second model assume that ELF (as used by non native speakers of English) may be regarded as a code in its own right and therefore that the native speaker English model is inappropriate. In tests based on the second model, strategic competence takes precedence over linguistic accuracy. While both models are somewhat problematic in practice, neither, it is argued, entails any radical reconceptualization of language testing beyond what has already been envisaged and/or enacted in the field. The paper concludes by cautioning against moving too quickly to assess ELF before it has been properly described.
Papers and Posters

Session 1
Thursday, June 29, 10.25am-12.10pm
GM15 – Law Building

Papers 1 - 3

1. Raters’ Perceptions of Scoring Criteria in Writing and Speaking Performance Assessments

Thomas Eckes
TestDaF Institute

Previous research on TestDaF essay scoring has shown that raters differed substantially from each other in the way they attached importance to routinely-used scoring criteria. Building on these findings, I carried out more detailed analyses of the differential perception hypothesis: First, I considered two more dimensions of criterion perception (besides importance); second, I extended the basic research design to cover the perception of scoring criteria in the speaking section. With each criterion, raters were asked to indicate, on a four-point rating scale, (a) the importance attached to it in general, (b) the overall ease of applying it, and (c) the confidence in using it correctly in operational scoring sessions. The criteria that were rated on these dimensions referred to various performance aspects, such as fluency and grammatical correctness (for writing), or comprehensibility and appropriateness of response to the communicative situation (for speaking). Ratings of essay scoring criteria were provided by 64 raters, ratings of oral proficiency criteria were provided by 53 raters. To test the differential perception hypothesis and its derivative, the rater type hypothesis, I first subjected each set of ratings to a series of many-facet Rasch analyses. Results confirmed that raters differed strongly on the importance, ease of application, and confidence dimensions for both writing and speaking. Then, employing a two-mode clustering approach, I identified rater types characterized by distinct ways of perceiving criteria along these dimensions. The findings have implications for rater monitoring and rater training in large-scale language performance assessments.

2. Exploring rater and task variability in second language oral performance assessment

Youn-Hee Kim
McGill University

While performance assessment has broadened and enriched the practice of language testing, ongoing questions have arisen as to whether complexity and variability in performance assessment influence a test’s usefulness. That testing tools and human factors must be involved in test-taking and rating procedures is inevitable, but these factors have long been recognized as potential sources of variance that is irrelevant to a test’s construct. This study continues the ongoing discussion about rater and task variability by comprehensively examining how second language oral performance is assessed by different groups of teacher-judges across different task types. The substantive focus of the study investigated whether native English-speaking (NS) and non-native English-speaking (NNS) teacher-judges exhibited internal consistency and interchangeable severity, and how they influenced task difficulty and the calibration of rating scales across different task types. It was also identified what the salient construct elements for evaluation were to the two groups of teacher-judges across different task types when no evaluation criteria were available for them to consult.
A Many-faceted Rasch Measurement analysis of 1,727 ratings and a grounded theory analysis of 3,295 written comments on students' oral English performance showed little difference between the NS and NNS groups in terms of internal consistency and severity. Additionally, the two groups were neither positively nor negatively biased toward a particular task type. The NS and NNS groups, however, did differ in how they influenced the calibration of rating scales, and in how they drew upon salient construct elements across different task types. The suitability of the NNS teacher-judges, the need for context (task)-specific assessment, the usefulness of the Many-faceted Rasch Measurement, and the legitimacy of mixed methods research are discussed based on these findings.

3. Re-training writing raters online: how does it compare with face-to-face training?

Ute Knoch, John Read & Janet von Randow
University of Auckland

Online rater training has become increasingly popular as a means of moderating writing ratings because of its obvious benefits in terms of practicality. However, whilst face-to-face rater training has been shown to be effective (Weigle, 1994; 1998), online training has had less clear-cut results. A preliminary study by Elder et al. (forthcoming) showed that the online training program worked better with raters who had a positive attitude towards using the program.

The current study aims to compare the effectiveness of online and face-to-face training in the context of a large-scale academic writing proficiency assessment used to diagnose the language support needs of students entering a major English-medium university. A team of 16 raters was divided into two groups of eight and initially rated a set of 70 scripts. Then, in the training phase, the online group rated 15 benchmark scripts online and received immediate feedback on their rating behaviour. The face-to-face group received individualised rating reports on their pre-training performance, rated the 15 scripts at home and then met for a face-to-face training session to discuss their ratings of the benchmark scripts. After the training, both groups re-rated the initial 70 scripts and filled in a questionnaire eliciting their attitudes towards the rating process. A subgroup of the raters were also interviewed. The results were analysed using multi-faceted Rasch analysis. The findings are discussed in terms of the factors influencing rater responsiveness and the refinements that are needed for future rater training programs.

Posters 1 - 3

1. This Test is Unfair: I'm Too Old. He's Too Young

Jeff S. Johnson
University of Michigan

One of the challenges for test item developers is limiting item bias, which is defined as a significant probability that particular items are easier or harder to answer correctly for a particular group due to factors irrelevant to the test construct. Reading passage topic familiarity has been shown to have an effect on EFL/ESL reading comprehension test scores. The genesis for this study is the possibility that large age differences may be significantly related to topic familiarity, which in turn may lead to age-biased scores. Young examinees (15 and younger) may be more familiar with certain topics due to cultural and educational factors, and older ones (30 and older), may have an advantage on other topics due greater experience and exposure to world events. Data for this study come from a large-scale EFL certification test which is taken by over 30,000 examinees each year. Examinees were divided by country to control for cultural differences. The reading section (k=100) in the test includes grammar items, vocabulary items, cloze items, and four reading comprehension passages,
each with five comprehension questions. SIBTEST was used to measure differential bundle functioning (DBF) for each set of reading passage items with examinee age groups matched for proficiency based on scores from the grammar, vocabulary, cloze, and three remaining reading comprehension sections. Several reading passages show significant DBF, some favoring young examinees and some favoring older examinees. The aim of this study is to highlight topic attributes that test developers should consider to create age-fair tests.

2. Introducing audit trails to the world of language testing

Jinshu Li, Hyeong-Jong Lee, Jung Tae Kim & Carsten Wilmes
University of Illinois at Urbana-Champaign

Although it is widely used by business for quality insurance, an audit trail is an unconventional idea to language testing and there is little published research in this area. Our concern with validity - how to achieve it, and how to enhance it - forms the backdrop of our studies. We believe that it is instructive to explore the transfer of audit trails to language testing.

An audit trail is a documentation system that encourages critical reflection on the evaluation process as well as documenting evaluation decisions and their justification (Guba & Lincoln, 1989; Whitmore & Ray, 1989). In this poster session, we review how we implemented audit trails in various language test development projects. Our team adopts the criterion-referenced framework, which has four characteristics: iterative, feedback-laden, consensus-based, and specification (spec)-driven (Davidson & Lynch, 2002). These characteristics reveal the evolutionary nature of specs, the blueprints to generate test items/tasks. We keep track of changes in the specs and develop validity narratives in areas such as consistency, authenticity, alignment and bias-check. Each narrative documents an important change and its justifications, cross-referenced with feedback we received from stakeholders.

We found that an audit trail is a reliable approach of monitoring and enhancing the quality of language tests. It enables our team or a third-party examiner to systematically review the test development process. It also has a positive impact on validity by assembling pieces of information making up the record of validity evidence (Messick, 1989), and providing anticipatory preparation for later challenges.

3. Examining Generalizability of Speaking Performance Measures Across Tasks

Rie Koizumi
Tokiwa University

Generalizability (i.e., the degree to which interpretations based on test performance can be generalized across target groups, tasks, and other settings) is an essential aspect of validity (e.g., Messick, 1996) but is problematic in assessing speaking (Fulcher, 1997), which is affected by many factors other than language ability (e.g., tasks, interlocutors; e.g., Fulcher, 2003). This poster aims to investigate the generalizability of speaking performance measures across tasks. Speaking performance measures are "discourse analytic measures," which "provide counts of specific linguistic features occurring in the discourse" (Ellis, 2003; e.g., "No. of error-free clauses" divided by "No. of clauses" for accuracy). They have been substantially utilized in SLA research (e.g., Skehan & Foster, 2001) without much examination of their qualities. This study can help researchers and teachers select appropriate measures, as well as understand the importance of using multiple measures and tasks.

The participants were 225 Japanese EFL learners at the novice level. They took a tape-mediated speaking test comprised of monologues. Speaking performance derived from five tasks was assessed from four aspects using multiple measures: fluency, accuracy, syntactic complexity, and lexical complexity. Principal components analysis was conducted on each aspect to investigate whether a measure across tasks loads on the same component. Among the findings is that, concerning fluency, measures associated with speaking speed (e.g., No. of words per minute) are generalizable across tasks but measures related to dysfluency
markers (i.e., repetitions, self-corrections, and false-starts) are not. Other results using
test reliability analyses and implications for language teaching and assessment are discussed.

Session 2
Thursday, June 29, 1.40pm-3.30pm
GM15 – Law Building

Papers 4 - 6

4. A longitudinal approach to the validation of a standards-based classroom
assessment of English

Lorena Llosa
New York University

Many accountability systems not only focus on student achievement or mastery of standards,
but also on the amount of progress students make each year. Title III of the No Child Left
Behind Act, for example, requires states and school districts to show annual increases in the
number and percentage of students who become proficient in English, as well as in the
number and percentage of students who make progress toward that goal. But in order to
interpret changes in test scores from one year to the next as reflecting differences in
underlying language ability rather than as variations in the measurement, the language
assessments used should be measuring the same constructs over time.
The purpose of this study was to determine the extent to which a standards-based classroom
assessment used to make high-stakes decisions about students actually measures English
proficiency consistently over time. Three years of data were collected for 400 elementary
school students (Grade 2-4) and analyzed using a structural equation modeling approach. To
examine factorial invariance, equality constraints were imposed on parameters in the factor
structure across occasions and the resulting difference in the $\chi^2$ values for the two models
was tested for significance. Findings will be explained and discussed in term of their
implications for the use of standards-based classroom assessments to monitor student
progress within a high-stakes accountability system.

5. Effect Sizes in Using Test Accommodations To Improve ELLs’ Achievement Test
Performance

Maria Pennock-Roman
MPR Psychometric & Statistical Research & Consulting
Charlene Rivera
The George Washington University Center for Equity & Excellence in Education

Despite the limitations of studying mean differences to evaluate the psychometric validity of
test accommodations (National Research Council, 2004), it is the predominant evaluative
approach, largely for pragmatic reasons. The objective of our review is to examine to what
degree different types of accommodations available for English language learners (ELLS) lead
to possible improvements in the test performance of ELLs on achievement tests. We
reviewed and summarized all the studies to date that randomly assigned school-aged ELLs to
test accommodation conditions and reported means for the groups. We created summary
tables with a separate row for each type of accommodation included in the study, and
columns for type of content (e.g., science, mathematics), grade level, effect sizes, and
sample sizes per condition, per grade level, age, and other variables. We calculated Cohen’s
d (effect size—the difference in means divided by the control group or total standard
deviation depending on the context) for various contrasts. The contrasts included, where
possible: ELLs receiving an accommodated version versus ELLs receiving the control
(standard English-language) version of the same test booklet; ELLs versus non-ELLS receiving
the control version; ELLs receiving an accommodation versus non-ELLs receiving the same accommodation. Most effect sizes for the first type of contrast fell into the range considered "small," but we highlight exceptions. Other contrasts also are discussed. In addition, we discuss design features that future studies should incorporate, such as using English language proficiency not only as a covariate, but also as a predictor that possibly interacts with the accommodation type.

6. Comparing Dichotomous and Polytomous Scoring Procedures on Multiple-Choice and Gap-Fill Tasks of Grammatical Knowledge

James E. Purpura
Columbia University

Methods used to score multiple-choice and gap-fill tasks in L2 assessments have traditionally been based on practicality concerns rather than on theoretical or empirical rationales. As a result, both MC and GF tasks of L2 grammatical knowledge have traditionally been scored in a binary fashion, one for a correct and zero for an incorrect response. However, binary scoring often fails to consider the developmental information being measured in different types of multiple-choice distractors (Haladyna, 2004) or in varying types of "wrong" gap-fill responses. If language testers wish to increase score precision, especially in the lower half of the test score distribution, and potentially provide more accurate information to developing test-takers, testers must remain open to polytomous scoring techniques for MC and GF items. In an effort to determine which method provides the most meaningful and precise assessment of the examinees’ grammatical knowledge, this paper compares three methods (dichotomous scoring; partial credit with a two-point and partial credit with a three-point scale) used to score two different tasks (MC and GF). The data in this study involved 525 students responding to 15 MC items and 15 GF items in Paper Three of the First Certificate in English Exam (Cambridge ESOL, 2002). Four experienced ESL teachers independently scored the items based on developmental criteria. Three separate analyses were performed on both tasks using classical and IRT procedures. Results showed that similar to studies not involving language (Haladyna & Sympson, 1988; Schultz, 1995), option weighting generally perform better than simple dichotomous scoring.
4. The TEPOLI test: construct and parameters to assess EFL teachers’ oral proficiency

Douglas Altamiro Consolo – UNESP (Brazil)
Vera Lúcia Teixeira da Silva – UERJ (Brazil)

In this poster presentation we report on results from an ongoing research project (henceforth PROJECT) that aims at analyzing the oral proficiency in English of EFL teachers-to-be in teacher education courses - ‘Letters courses’ - in Brazil. The focus is on the process of revising a theoretical background that supports the definition of language aspects to be considered in the establishment of standards for English teachers’ oral proficiency (henceforth ELT-OP) and data collected by means of an oral proficiency test - the TEPOLI (in Portuguese, *Teste de Proficiência Oral em Língua Inglesa*), a research instrument being developed for collecting data but which may evolve into a test in ELT. The issue of EFL-OP has also been investigated in the PROJECT with reference to data collected by means of questionnaires, audio-recorded interviews, classroom observation, recorded lessons and other oral tests. These data have indicated that further investigation in the direction of parameters to assess teachers’ language is necessary. Hence, within the PROJECT, we aim at contributing for the development of a valid and reliable oral proficiency test for those future teachers. The construction of TEPOLI and its application to students graduating in Letters will be illustrated, followed by a discussion of the descriptors that constitute the rating scale for the test, which vary from a level of proficiency minimally adequate for such professionals, to higher levels of performance that can be attained by students who graduate from those courses in Brazil.

5. Developing a test for interpreter certification in Norway

Diane Mortensen
University of Oslo

More than a decade ago, Norwegian authorities recognized the need to establish fixed standards of quality for interpreting services in the public service domain. This resulted in the development of a nationwide interpreter certification scheme based on a practical written and oral test of interpreting.

Responsibility for the development of the Norwegian Interpreter Certification Examination (NICE) was delegated to the Interpreter Certification Project at the University of Oslo. Since the mid-1990’s, the NICE has been held in 16 languages, and more than 100 candidates have become State Sworn Interpreters.

Due partly to a lack of adequate training opportunities for both working and would-be interpreters, the pass-rate for candidates taking the NICE has been consistently low (15%). This has led to pressures from the public service community and others to lower the standards, thus allowing for a greater number of interpreters to become certified. The Interpreter Certification Project maintains, however, that a high level of professional competency should be required for certification, and that the criteria upon which evaluation of a candidate’s performance on the NICE is presently based, are justifiable.

This poster presentation will provide a brief description of the NICE and present some of the challenges that the Project has faced in developing and monitoring the exam. The visual presentation will be primarily problem-oriented, designed to stimulate discussions about interpreter testing in general, and in particular, how to ensure that candidates’ results are in fact true indicators of their interpreting skills.
6. Development and Pilot Testing of an Interview Test in Multiple Languages

Nathan T. Carr, Hélène Domon, Janet L. Erying, Jack (Jinghui) Liu, Michael S. Ray
Setsue Shibata, Marjorie Tussing
California State University, Fullerton

This presentation relates to the new university-wide second language graduation requirement at California State University, Fullerton, a large, ethnically diverse campus in Southern California. With the entering freshman class of 2006, all incoming undergraduate students will be required to complete three semesters of foreign language instruction, or demonstrate an equivalent level of competence in two of the four “skills.” An unknown but large number of students are expected to attempt to test out of this requirement; most of them are likely to be heritage speakers of various Asian languages and Spanish, and as heritage speakers, are likely to request tests in speaking and listening. For this reason, initial test development efforts have focused on developing an interview test to assess speaking and listening ability, particularly in languages not taught at the university. Assessment in these languages will be performed by faculty and “community partners” who are native speakers of heritage languages not taught on campus. To this end, a generic interview test was developed to be used in multiple languages. It includes interview prompts at the novice, intermediate, and advanced levels in a variety of topics, as well as an analytic rating scale. This poster presentation will address the development of the test specifications; the construction of the interview test; the rating scales; the results of pilot testing conducted in Chinese, French, Japanese, Spanish, and Vietnamese; and procedures for training prospective interviewers.

7. A Pre-Instruction EFL Test for the Airline Industry

Fred Davidson, Hyeong-Jong Lee, Jinshu Li, Jungtae Kim & Carsten Wilmes
University of Illinois at Urbana-Champaign (UIUC)
Sarah Blayer, Berlitz

English is the international language for pilot-to-controller communication in the airline industry. A number of tests have been developed to certify pilots and controllers for communication in English, and these tests are in use in many settings around the world. Professionals in this industry acknowledge the fundamental importance that such tests yield valid and reliable inferences; the alternative could be a catastrophic accident due to miscommunication.

Various English language training programs have arisen to help pilots and controllers improve English language ability, prior to any language certification exam. Airlines or their employees often ask: if we invest in a pre-certification course of instruction, what is the proper course placement, and how long can the student reasonably expect to spend in the instructional sequence?

We report on the development of a low-stakes, non-certifying, pre-instruction screening test for that placement purpose; we emphasize that our test does not ‘sign off’ on readiness for a test-taker to work in English, either as a pilot or a controller. Our test utilizes a semi-adaptive telephone-delivered oral interview, in which the rater follows a semi-structured evaluative scheme. Our poster details some of the evolution and design decisions that governed the rollout of the test, most notably the feedback from, and training for the raters.
7. DIF in a test of ESL pragmatics

Carsten Roever
The University of Melbourne

Research on test bias and differential item functioning (DIF) in language testing has been commonly applied to traditional test batteries covering grammar, vocabulary, and listening / reading (e.g., Bachman & Ryan, 1993; Elder, McNamara, Congdon, 2003; Henning & Chan, 1985; Kunnan, 1990) as well as to writing (Hunter & Lee, 2004) and speaking (Kim, 2001). This study applies DIF detection techniques to a test of ESL pragmalinguistics, which consists of three sections of twelve items each, covering implicature, speech acts, and routines. The study investigates whether test takers of European language background performed differently on the test than test takers of Asian language background and to what extent such differences could be accounted for by differential cultural background knowledge. The European language group consisted of 165 test takers with L1s German and Polish, the Asian language group of 123 test takers with L1s Chinese, Japanese, Thai, and Korean. Following Muniz et al. (2001), two DIF procedures were used: the Mantel-Haenszel procedure (Camilli & Shepard, 1994; Dorans & Holland, 1993) and the standardization procedure (Dorans, 1989; Dorans & Holland, 1993).

A total of 2 implicature items, 5 routines items and 2 speech act items were identified as possibly showing DIF, most in favor of the European group but generally with minor DIF according to the ETS classification (Dorans & Holland, 1993). Possible reasons for DIF and differences between the two detection procedures will be discussed.

8. The impact of test delivery modes and test-taker characteristics on speaking test performance

Eunju Yu
Ohio State University

In the era of globalization, a variety of spoken English tests have been developed to measure the communicative ability of non-native English speakers with different backgrounds. While Taylor, Jamieson, Eignor and Kirsch (1998) reported no significant correlation between examinees' computer ability and their computerized TOEFL test scores, Jeong (2003) found a positive relationship between their electronic literacy and test results. This study attempts to resolve this contradiction.

This study investigates the effects of test delivery modes and test-taker characteristics (i.e., computer familiarity and English learning experience) on test-taker performance during oral proficiency assessments. For the purposes of this research, approximately 160 international graduate students whose native language was not English were recruited in the US. The main data sources include the results of a computerized spoken English test, a conventional audio-taped speaking test, as well as replies to questionnaires. A 2 x 2 factorial research design is used for data analysis.

This presentation reports interactions between test-taker characteristics and test delivery modes during oral proficiency assessments. This study discusses both the benefits and disadvantages of the use of technologies in oral proficiency assessments. The reactions of test-takers to different speaking test delivery modes are reported. The findings enable test developers, language educators and other stakeholders to develop valid testing instruments. By interpreting test results appropriately, language educators can customize language
programs which correspond to the need of the particular population. The findings also help to provide testing and assessment formats based on students' preferred learning styles.

9. Task Difficulty in Semi-direct Speaking Tests

Jessica R W Wu
The Language Training and Testing Center

In Taiwan, as in many other EFL countries, a semi-direct testing approach is considered a more appropriate choice of speaking test format due to a lack of qualified interlocutors and assessors as required in a large-scale direct speaking test. As the semi-direct format becomes widely adopted in Taiwan, the need to check test quality in relation to reliability and validity continues to be pressing.

This study investigates the effect of intra-task variations in association with performance conditions on task difficulty in a semi-direct speaking test. The variables identified for investigation are linguistic demand of task input in relation to code complexity; amount of time allowed for performance in relation to communicative demand; type of pre-task planning in relation to communicative demand; and test-takers' familiarity with the non-verbal propositional content of the task input in relation to cognitive complexity. In addition, learners' proficiency in English is treated as a moderator variable to investigate the extent to which learners' English proficiency interacts with the effects.

The study takes a multi-dimensional approach in measuring the effects of the variables based on three different sources of data, including task scores, learners' responses to the post-task questionnaires, and interlanguage measures. The results indicate that the variations affect task difficulty and the effect varies in some cases due to learners' proficiency in English. In general, the more proficient learners tend to react to the variables more strongly than the less proficient learners. Significant implications for reliability (i.e., parallel tasks/forms) and content validity are also suggested.

10. Intra-task comparison in a monologic oral performance test: the impact of task manipulation on performance

Tomoko Horai
Roehampton University

'Monologue (individual long turn)' is advocated as one of the ideal elicitation tasks for assessing speaking ability in low and high stakes testing. Its practical advantages are that it can be delivered in a variety of forms, e.g. live or recorded. Moreover, as a single speaker produces a long turn without interacting with other speakers, it does not suffer from the contaminating effect of the co-construction of discourse, seen by McNamara as potentially resulting in construct irrelevant variance.

This paper explores how EFL learners' performance on a monologic speaking task is affected by deliberately manipulating tasks, in terms of planning time, the amount of support, and response time. Four monologic speaking tasks, which were demonstrated in an earlier phase of this project to be equivalent from a quantitative and qualitative perspective, were first manipulated and then given to 120 EFL participants. The resulting performances were analyzed using candidates' scores for their performance, results from a cognitive processing behaviour questionnaire (in pre-planning, in planning and during speaking stage), and through a linguistic analysis of the resulting spoken texts.

The results show how manipulating the tasks significantly impacted on the scores and cognitive processing while planning, as well as on the language generated. These contribute to our understanding of how task performance can be more accurately predicted and assessed.
Session 3b  (parallel session)
Thursday, June 29, 3.50pm-5.50pm
Theatre 102 – Law Building

Papers 11 - 14

11. English language exams reform project in India: From achievement to proficiency testing

Antony John Kunnan, California State University, Los Angeles
Rama Mathew, Delhi University

English language exams in India are administered by national or state secondary boards of examination at the end of Class 10 and Class 12. These exams are typically achievement type tests where reading and writing tasks are based on school texts and the focus of interest is knowledge and understanding of set reading texts and written responses to these set texts. As a result, students' test scores mainly reflect their ability to memorize and present stock responses to both reading and writing tasks. This situation has led to the problem that after six to eight years of instruction in English (depending on the school system), students are unable to demonstrate reading and writing abilities that are appropriate for their level (without previously reading and understanding the material).

In this paper, we will present a report of a reform project conducted with the intent of moving away from achievement to proficiency testing for two grade levels (Class 5 and 8) using the Delhi State curriculum and prescribed textbooks. The intent is to assess students' reading and writing abilities with new unseen materials (that is, without set texts). The report will include: 1) the articulation of proficiency levels through can-do statements (similar to the study abilities listed in ALTE 2002) that are appropriate for the multilingual context of India; 2) a demonstration of proficiency tasks developed on the basis of the can-do statements; 3) a discussion of test scores of school students (about N=250) based on new proficiency test tasks and the traditional achievement test tasks; and 4) a discussion of the correlational relationship between the two types of tasks. The presentation will also include a discussion of the implications of this project for teaching, testing, teacher training, curriculum, and policy.

12. A model for investigating the impact of language assessment within a national educational context

Nick Saville
Cambridge ESOL

This paper proposes a model for investigating the impact of language assessment within a national educational context.

It sees the investigation of impact not as a discrete or one-off activity, but as an essential component in establishing the overall validity of an assessment system in terms of its fitness for a specific purpose and context of use. The proposed model locates the study of test impact as one of a set of research and development tools within an iterative approach to validation.

A number of impact studies carried out by the speaker over the past 10 years have informed the proposed model and these will be briefly summarized. The first is a world-wide survey of the impact of an international English language testing system. The second is a study of the impact of language tests within a national reform project in Europe. The third, at a micro level, looks at a specific teaching/learning environment within a single language school in Italy. The case study data from each of the contexts have been used as meta-data leading to a meta-analysis and a comprehensive model for application in other educational contexts. The final part of the talk looks at current impact research based on the emerging model, and in particular its use in the context of Asset Languages, a British government-funded assessment project covering more than 20 languages. Will this implementation of the
Languages Ladder recognition system deliver the intended impacts as a key element of the National Languages Strategy?

13. Applying the 'diffusion-of-innovation' framework to an examination reform in China

Qi Luxia
Guangdong University of Foreign Studies
Liu Qingsi
National Education Examinations Authority

Innovation theory has been adopted in a number of studies on innovations in English Language Teaching (Stoller, 1994). The findings from these studies have shed light on the process of change and reform in the field. By applying some concepts from innovation theory to an analysis of a washback study conducted in Sri Lanka, Wall (1996, 2000) found a plausible explanation for the lack of washback expected in that context. Innovation theory seems applicable to research on language testing reform although few such studies have been reported in the literature. The present paper reports a case study which applied the framework of "diffusion-of-innovations" (Markee, 2001) to an evaluation of the implementation of an English listening test. This test is an important component in the reform of the national university entrance examination on the subject of English in China. Data collection and analysis were based on document analysis, interviews, and participant observation. It was found that logistics of test administration is a crucial factor contributing to success or failure of the reform. This is because test reform is deeply rooted in the social-political setting where the test is used and logistical problems could cause a threat to social stability, which would result in ending the test altogether. The implications for innovators interested in reform of large-scale language tests are that sufficient considerations must be given to logistical problems which are directly related to the social character of testing discussed by McNamara (2000).

14. An impact study of Taiwan’s General Proficiency English Test (GEPT)

Viphavee Vongpumivitch
National Tsing Hua University

Taiwan’s first large-scale standardized test of English, the General English Proficiency Test (GEPT), is developed with the goal to “promote the idea of life-long learning and to further encourage the study of English”. The test’s intended use is extremely broad, ranging from diagnostic to selection, placement, and graduation tests. With at least 910,000 test takers since 2000, the impact of GEPT on society is large, as mock-exam books filling bookstores and countless GEPT courses being offered.

With all the enthusiasm about GEPT, some questions remain: Is GEPT really promoting the idea of life-long learning? Could it be that GEPT is yet another target of cramming? In Taiwan, is a standardized test like GEPT needed because it is the only way people can be forced to learn English? These questions are relevant to the value implications of GEPT and its consequential evidence of validity in Messick’s (1989) framework.

This study reports in-depth interview results with different groups of stakeholders, namely, university students (n = 20), university professors (n = 10), elementary school teachers who enrolled in a GEPT brush-up program (n = 5), an ESL program administrator, and teachers of GEPT courses (n = 2). Participants discuss their views on the test and the role of GEPT as an agent to promote life-long learning of English in Taiwan. The findings in this pilot study reveal the impact of GEPT in Taiwan and also serve as an example that may be relevant to other countries.
Session 4 (parallel session)
Parallel with Symposium 2
Friday, June 30, 8.30am-10.00am
Theatre 102

Papers 15 - 17

15. The Impact of ESL Placement Test Decisions

Young-Ju Lee
Sookmyung Women's University

This study examines the impact of an ESL placement test on international graduate students at a large public university in the U.S. The test was administered to incoming students and was used to place them into appropriate ESL writing courses. Some test takers might be misclassified as non-masters and are required to take ESL courses. Therefore, they do not benefit from a full registration of content courses. Students who are unhappy with their test results are operationally defined as "malcontents" in this study. As case study participants, three students were selected based on their perception that they were not happy with their test results; they were tracked for the subsequent two semesters of their academic program. A researcher conducted in-depth, face-to-face interviews with them, and audiotaped and transcribed verbatim each interview. This study presents narratives from the test-taking experience to the first year of graduate program for these three malcontents. Three malcontents reported that they benefited from taking ESL courses and appreciated the benefit, which showed that adverse consequences were minimized at this academic context. The inclusion of malcontents' points of view contributes to the understanding of construct validity in several aspects. First, the effect of the tests on students addresses the consequential validity. Second, it focuses on test use, especially negative attributes, which have not been reported explicitly in any validation study. Third, the perspectives of students, the ultimate stakeholders in testing, need to be reflected in test validation.

16. Predictors of Item Difficulty in an Automated Spoken English Test

Don Salting & Jennifer Balogh
Ordinate Corp.

Versant for English is a 10-minute spoken English test that is administered over the telephone and scored automatically using speech processing technologies. Previous research has identified factors that predict item difficulty for a Sentence-Repeat task (Nair, 2005). The question addressed in this research is whether similar factors predict item difficulty in a different task called Sentence-Builds, or whether other syntactic features are more predictive. In the Sentence-Build task, test-takers hear a recording of a sentence that has been divided into three relatively coherent phrases. In the prompt, the phrases are spoken in an incorrect, random order. The test-taker is asked to produce a sensible and correct sentence from the phrases. In Nair's Sentence-Repeat research, the strongest predictors of difficulty included log number of phones and a vocabulary-based measure (average trigram log likelihood). While there must be differences between sentence repetition and phrase-based sentence production, the results corroborate previous findings, in that, of 22 different potential factors, log number of phones ($r^2=.19$, $p<.01$) and average trigram probability ($r^2=.13$, $p<.01$) were the strongest predictors. Other syntactic and morphological factors were also analyzed and found not significant. The results are couched within an interactive alignment account of language processing (Pickering and Garrod, 2004). Within this framework, speakers do not remember surface structures but rather remember the meaning of sentences and can repeat them back verbatim because of automatic processes such as priming. The importance of vocabulary, phonemic content and syntactic structure will be described within this model.
17. Typical Error Patterns for Different L1 Groups in ESL Test-Takers’ Essays

Yong-Won Lee
Educational Testing Service
Martin Chodorow
City University of New York
Claudia Gentile
Write to Change (Consultant)

The automated error-detection and feedback system is becoming an important component of online writing practice services based on retired prompts of large-scale ESL writing assessments. In refining such a system, one promising idea is to identify typical error patterns for ESL learners with different first language (L1) backgrounds and use such information to improve the accuracy of detection and feedback. The main purposes of the current study were: (a) to have ESL learners’ essays (from a large-scale assessment) error-coded and corrected by trained human coders and (b) examine patterns of writing errors that are unique to each L1 group or common across L1 groups.

Approximately 480 essays written for four different prompts (3 independent, 1 integrated) were error-coded and corrected by human coders. Five different L1 groups (Arabic, Chinese, Japanese, Korean, and Spanish) were represented in the essays. A total of eight coders were recruited and trained to code each essay for writing errors. For each L1 group, each essay was coded by two coders using a restructured version of Dagneaux et al.’s (1996) coding scheme.

Writing errors in adjudicated versions of essays were classified into major types and counted for each type. For different L1 groups, frequencies of different types of errors were examined at different score levels. Our presentation will show patterns of unique and common errors for different L1 groups, types of errors that are more easily captured by the automated system or humans, and their implications for automated error detection and feedback in diagnostic writing assessment.

Session 5
Friday, June 30, 1.15pm-3.05pm
GM15 – Law Building

Papers 18 - 20

18. The Trait Structure of Oral Language Ability in a Computer-based Speaking Test

Sutthisrik Sapsirin
Chulalongkorn University

Current technological advances have made it possible to deliver a speaking test via computers with various sources of input such as text, visual and audio (Flewelling & Snider 2001; Kenyon & Malabonga, 2001). However, the implementation of computer-based speaking tests (CBSTs) has raised concerns about the construct validity of such tests (Chapelle, 2001; Norris, 2001). This is because little is known about the extent to which the test methods of such tests may alter the candidate’s performance, which in turn may affect the assessment of their oral language ability. The purposes of this study were to examine the trait factors of oral language ability in a CBST and to investigate the processes the participants report using while taking the test. The CBST used in this study was developed to present test tasks with text and audio input and record the responses, which would be rated by human raters afterwards. The participants were approximately 200 first-year university
students in Thailand. A sub sample of the participants was asked to provide retrospective verbal reports on their test taking processes. Participants’ scores on the CBST will be analyzed using confirmatory factor analysis (CFA) of the multivariate-multimethod data. The hypothesized factors in the CFA model include five traits: knowledge of pronunciation, syntax, vocabulary, cohesion and functions, and four test methods: test tasks eliciting the functions of narrating, expressing opinions, describing an imaginary situation and persuading. The research results will be discussed.

19. Learner strategies and knowledge in TOEFL Listening Tasks: A validation study

Dan Douglas & Volker Hegelheimer
Iowa State University, USA

This presentation is based on the final report of an investigation of strategy use and the use of language and content knowledge by test takers while completing the New TOEFL listening test tasks. We employ a three-pronged approach for data collection and analysis: verbal protocol analysis of participants’ test-taking processes, screen capture of their online behavior, and analysis of their handwritten notes to investigate the following research questions:
1) What strategies do test takers use to respond to the tasks?
2) What linguistic and content knowledge do test takers use to respond to the tasks?

We contend, following previous research (Green 1998, Cohen 2000), that the processes of listening comprehension (Buck 2001, Rost 2005) are accessible to researchers by means of the analysis of introspective verbal protocols, online behavior, and handwritten notes. The focus in this project is on integrating these three sources of data to more fully understand test-taking and comprehension strategies and the use of language and content knowledge in responding to online listening test tasks. Participants in the study, 36 undergraduate and graduate test takers enrolled in social sciences, science, technology, and the humanities, completed tasks based on the listening passages in the Language (ETS 2003) materials. Quantitative and qualitative analyses focus on strategies and language and content knowledge employed and how handwritten notes are integrated into the response process. Findings will provide insights for the interpretation of test performance as evidence for language knowledge and communicative language use as defined in the test construct.

20. Validating the GEPT: Lexico-Grammatical Ability as a Predictor of L2 Listening Ability

Yen-Fen Liao
Teachers College, Columbia University

This study attempts to investigate the construct validity of the listening section and the grammar-vocabulary (GV) tasks in the reading section of the General English Proficiency Test (GEPT), a relatively high-stakes criterion-referenced test (CRT) widely used in Taiwan. A review of the literature on the role of lexico-grammatical knowledge in L2 listening comprehension reveals an enormous gap in the testing field and the strong need for further research. The present study thus focuses attention on exploring the issue of the relationship between lexico-grammatical ability and L2 listening ability in the context of the GEPT. The discussion first centered on the factorial structures of the GEPT listening section and the GV tasks, and then turned to the role of lexico-grammatical ability in predicting listening performance on the GEPT. Given the criterion-referenced nature of the GEPT, both the norm-referenced test techniques and the CRT approaches were employed to obtain information on item statistics and reliability. Exploratory factor analysis was then used to probe the underlying factors of the tests, followed by the implementation of structural equation modeling (SEM) to examine the degree to which the models represented the data. A model composed of two traits with lexico-grammatical ability and listening ability was confirmed. Subsequently, correlational analysis, hierarchical multiple regression, discriminant
analysis, and SEM were performed in light of two different constructs (form vs. meaning, vocabulary vs. grammar). The statistical analyses provided ample evidence in support of lexi-co-grammatical ability as a critical predictor of L2 listening ability.

Posters 8 - 11

8. Academic English for School-age Learners: Research and Standards-Informed Test Development

Alison L. Bailey, Becky Hsuan-Hua Huang & Christine A. Ong
Dept of Education UCLA/CRESST
Frances A. Butler
Language Testing Consultant

This study reports on the development of prototype test items to assess the academic English language proficiency of upper elementary students (Grades 4-6). The work is the outgrowth of research articulating the construct of academic English through analysis of language demands in content-area classrooms and standards (e.g., Bailey & Butler, 2002/3). Research questions included: Are the prototypes valid measures of academic language? Do they discriminate among students with known differences in level of English language proficiency? Following Davidson et al. (in press), the process includes an audit trail to document item modification. Sixty-seven reading items were created from test specifications based on linguistic analysis of math, science, and social studies textbooks. Authentic passages and items were rated for linguistic difficulty and item type familiarity by a focus group of ESL and content-area teachers (n=9). Three tryout phases were: 1) Initial tryout with a fourth grader to estimate completion time and clarify directions; 2) 77 students were administered prototypes in whole-class settings and performance was examined for consistency across items measuring the same type (e.g., vocabulary) and same complexity of language demand, correlated with concurrent reading assessments and background factors (e.g., grade, gender), and used to examine item discrimination between students with differing English proficiency; 3) Verbal protocols with 19 additional students revealed the differences in performance on certain items. The audit trail illustrates how test specifications and resultant prototypes should be considered working documents undergoing a process of revision and improvement as a result of multiphase tryouts before field-testing commences.

9. Quantitative investigation into task type effects on listening test performance

Yo In'nami
University of Tsukuba, Japan Society for the Promotion of Science

Despite determined efforts to investigate how task types affect test performance, past studies (e.g., Brindley & Slaty, 2002) remain uncertain of the magnitude of such effects, which masks whether statistically significant effects of task types are realistically meaningful. Furthermore, even if the magnitude of such effects is estimated, interpretation is difficult because a result based on a single investigation is often susceptible to, for example, chance variability, design quality, sampling error, and research settings (e.g., Cooper, 1998). This poster presents a quantitative investigation into the extent to which three task types (multiple choice, open ended, and summary gap filling tasks) affect listening test scores using meta-analytic and data-gathering (conducting and analyzing listening tests on study samples) studies. For the meta-analysis, the search for studies was conducted through relevant literature and journals and through computer searches of databases (e.g., ERIC). After the literature was retrieved, the effect size measure of unbiased estimates of $g$ (Hedges & Olkin, 1985) and its 95% confidence interval were computed for each task comparison. They were then weighted by sample size, and the combined mean effect size was computed and interpreted. For the data-gathering study, 450 EFL students took two listening tests to estimate their listening proficiency levels and also to investigate task type effects. Their
responses were analyzed for each task comparison at proficiency and item levels. Preliminary meta-analytic results show that the multiple choice format is easier than the open ended format. Other findings and implications for language assessment are discussed.

10. Validating the revised IELTS speaking test

Annie Brown
Ministry of Education, United Arab Emirates

The study reported here was implemented as part of the validation programme for IELTS, in which revisions to the test are monitored and evaluated. The aim of the study was to examine the validity of the assessment process used in the revised IELTS oral interview, in which a set of four analytic scales replaced the earlier holistic scale. The study addressed the following questions:

- How do raters interpret the scales and what performance features are salient to their judgments?
- How easy is it for raters to differentiate levels of performance in relation to each of the scales?
- How easy is it for raters to differentiate the four scales?
- What problems do raters identify when attempting to make rating decisions?

The study involved a combination of two methodologies: a large scale survey of examiners, and verbal report data collected from a group of IELTS examiners. The findings indicated that, in contrast with their use of the earlier holistic scale, examiners adhered closely to the descriptors when rating, and in general found the scales easy to interpret and apply. Specific problems related to overlap between the scales, a lack of clear distinction between levels, and the inference-based nature of some criteria.

The presentation will highlight the implications of the study not only for the revised IELTS speaking test, but for oral assessments more generally.

11. What’s the problem? Technology and communication as mediating factors in test implementation

Janna Fox
Carleton University

This poster documents the unanticipated technological and communication issues that impeded the introduction and use of a new screening and placement test. The English for Academic Purposes (EAP) test was commissioned at-a-distance by the College of the North Atlantic-Qatar, an English-medium institution experiencing rapid growth. The College needed a test, which would more efficiently select and group applicants. The new two-part test was designed for automated test assembly (ATA) (Parshall, Spray, Kalohn & Davey, 2002) from a task bank, and computer delivery. First, responses to the multiple-choice screening test were automatically downloaded into a database and scored, producing an initial test taker profile. The placement test was then administered, to those demonstrating a minimum level of proficiency on the screening test, with written and spoken responses scored by human raters. Unanticipated issues arose, however, with the introduction of the test: 1) communications were difficult as much as a result of disciplinary distances (testing vs. computer technology) as geographical ones; 2) test administration decisions were shaped as much by technological as testing concerns (Brown, 1996); 3) the sources of test problems were confounded by the computer interface (Gorsuch & Cox, 2000). This poster demonstrates not only the enhanced language testing potential of computer-based tests, but also their increased complexity. Addressing this complexity as part of initial test planning and development is an essential (and often overlooked step) in effective test development.
Session 6  
Friday, June 30, 3.45pm-5.45pm  
GM15 – Law Building

Papers 21 - 24

21. De-construct-ing individual and paired oral proficiency testing

Lindsay Brooks  
OISE/University of Toronto

This paper presents the differences in the features of interaction between adult ESL test-takers participating in an individual oral proficiency test, in which they interact with an examiner, and the same test-takers participating in a paired oral proficiency test, in which they interact with another student. The data are drawn from a larger thesis study comparing the two test formats in the context of high stakes exit testing from an Academic Preparation Program at a large Canadian university. All of the participants (N=80) were tested both individually and in pairs with comparable speaking prompts. The findings from the qualitative and quantitative analyses show that overall not only was the performance of the students better in the paired format (based on statistical results) but the language in this test format was much more complex. Actual differences in performance and test-takers’ perceptions of their performance will also be compared along with analyses of the complexity of the dialogue in the paired interaction. However, not all of the students did better in the paired test format; the results suggest there was a person by format interaction with some students performing differentially depending on the test format. The findings from this study have implications for test theory as analyzing the language in these test formats provides insight into the constructs and the contextual features that may advantage or disadvantage different test-takers.

22. Validating the group oral discussion task: The effects of personality type

Gary J. Ockey  
University of California, Los Angeles

Over the past decade, the group oral discussion task, in which three or more test takers discuss a topic with no interruptions or promptings from interlocutors or other non-test takers, has come into widespread use as a test of foreign language speaking ability. However, the extent to which the group oral yields valid score-based inferences is unclear. A number of personal characteristics, such as the personality types of the members in one’s group, could affect the performance of individual test takers, and hence the validity of score interpretations. This study is embedded in an assessment utilization argument framework (Bachman, 2005) and reports on the extent to which individuals’ scores on the group oral discussion test are affected by their levels of assertiveness, compliance, and self-consciousness.

First year native Japanese speaking students majoring in English at a Japanese university were administered the Revised NEO Personality Inventory (NEO-PI-R) (Costa & McCrae, 1992) and a group oral test. Based on Rasch adjusted group oral scores (Linacre, 2001), the degree to which individuals’ levels of assertiveness, compliance, and self-consciousness affect their scores was investigated using structural equation modeling techniques (Bentler, 1985-2005). Results of the analysis will be reported along with implications for group oral testing.
23. "Effective interaction" in a paired candidate EAP speaking test

Lyn May
National Institute of Education, Nanyang Technological University, Singapore

The need to define and operationalise "effective interaction" in paired candidate speaking tests is becoming increasingly important, given the growing popularity of this test format. Co-constructed discourse in speaking tests has been explored mostly in the context of language proficiency interviews. Brown (2004), for example, found that raters' perceptions of a candidate's oral proficiency, and hence the scores they awarded, were affected by the choice of interviewer. It has been suggested that the impact of the interlocutor on ratings may be more pronounced in the case of paired candidate speaking tests. In research on the paired candidate segment of the FCE, Galaczi (2004), for example, identified different patterns of dyadic co-construction, and features of the performance associated with higher scores on the "Interactive Communication" scale.

However, as Brown (2004) has so clearly demonstrated, for a more complete understanding of this issue there is a need to go beyond candidate discourse and score data to include the rater's perspective. This paper will, therefore, examine how interaction in a paired candidate EAP speaking test is perceived by raters, through analysis of rater notes, summary statements, stimulated verbal recalls and paired rater discussions in conjunction with scores given for "Effectiveness". The findings have implications for our understanding of the construct of effective interaction in paired candidate speaking tests, the development of appropriate rating scales and the vexed question of the separability of individual contributions to the joint performance.

24. An empirically-based rating scale for 'interaction' in a paired oral test

Ana María Ducasse
La Trobe University

Paired tasks in oral proficiency tests have been incorporated into various large-scale tests and consequently this has stimulated research into both their advantages (Taylor 2001, Csepes 2002, Galaczi 2005) and their disadvantages (Foot 1999, Norton 2005). These studies have largely focused on what affects candidate output and/or test outcomes; however, raters and rating criteria are in a crucial mediating position between the output and outcomes. This warrants further investigation into rater orientation to paired interaction in such formats and the implications for rating criteria.

This paper reports on empirical rating scale development using test discourse of 17 beginner dyads in Spanish as a foreign language context and Verbal Protocols from 12 Spanish teacher raters. The aim was to develop a rating tool using test performance and rater orientation as empirical input. The main theoretical approach adopted was Turner and Upshur's (1995) Empirically-based, Binary-choice, Boundary-definition (EBB) method, which is adapted for the context. Turner and Upshur (1995) applied this approach to the development of a rating scale for speaking on monologic tasks; this study focuses on a scale to measure paired interaction.

The findings of the study show that raters focus on three areas: body language, the supportiveness of the listener and the level of engagement of the speaker. This supports the notion that new criteria are required for rating in a paired context. The implications for notions of the listening construct in interactive contexts are also explored.
Papers 25 - 27

25. Extracting meaningful speech features to support diagnostic feedback: an ECD approach to automated scoring

Xiaoming Xi, Klaus Zechner, and Isaac Bejar
Educational Testing Service

Although operationally the TOEFL iBT speaking section is scored holistically, the evaluation of specific components of speech such as delivery (pace and clarity of speech) and language use (vocabulary and grammar use) might provide diagnostic information to learners.

This study used an evidence-centered design approach (ECD) (Mislevy, Almond & Lukas, 2004) to extract features that provide evidence about the quality of responses to TOEFL iBT speaking tasks through the use of speech and natural language processing (NLP) technologies. The computed features were identified through a detailed explication of the scoring rubrics, informed by the relevant second language acquisition literature, and confirmed by TOEFL Speaking test developers.

Trained human raters assigned holistic scores and analytic scores (delivery and language use) to responses to each of six tasks. Both the holistic and analytic scores were on a scale of 0-4.

Classification trees were developed to predict the human holistic scores and analytic scores. Multiple linear regression models were built to predict the human scores as well. The trees and regression models were modified to reflect test developers' input, and cross-validated on another set of samples. The classification accuracy on the independent samples used for cross validation from the classification trees and regression models will be compared.

We will discuss the feasibility of extracting meaningful speech features amenable to diagnostic feedback and the possibility of using automated scoring as a quality control for human scores.

26. Washback in an embedded assessment task

Robyn Spence-Brown
Monash University

Embedded assessment (assessment based on performance in pedagogic tasks) is advocated as a means of motivating learners, ensuring the relevance and 'authenticity' of assessment, and minimising its disruptive effect. It is often implied that the washback on learning will be entirely positive, or that negative impact will be minimal, due to the low stakes involved.

However, the impact of assessment on the behaviour of teachers and learners in low stakes SLA environments is rarely investigated, and may be more pervasive and potentially problematic than is generally acknowledged.

This paper examines the processes and products of an embedded assessment task in which intermediate students of Japanese were required to conduct interviews with native speakers, under conditions that were designed to be authentic. Through the use of retrospective interviews incorporating stimulated recall, as well as analysis of the discourse of the assessment products themselves, the study examines the ways in which students framed the task, the strategies that they employed and the resulting discourse produced for assessment.

Drawing on activity theory, the study explores the complex dynamic of task engagement and execution, and the personal, interpersonal, social and physical conditions that contribute to and are influenced by it. It highlights the crucial role of assessment in shaping student behaviour, problematizing the notion that assessment contexts can uniformly be assumed to be low or high stakes, and discusses the implications for both assessment and pedagogy of using learning tasks for summative assessment.
27. Some implications of student reading assessments in PNG and the Philippines

Prue Anderson
Australian Council for Educational Research

In Papua New Guinea and the Philippines English is a foreign language for most students, yet it is the main medium of instruction in schools after Grade 3. Knowing how well students can read in English is critical to any evaluation of the effectiveness of the education system and interventions funded through aid programs. This paper discusses the findings of two reading assessments as part of AusAID-funded educational initiatives: one administered to 7171 students in Grades 4, 5 and 7 in PNG in 2003; and the other to 18511 students in Grades 4 and 8 in Mindanao (Philippines) in 2004. Key issues were the match between curriculum expectations and student achievement levels and the diagnosis of weaknesses and strengths in students’ mastery of reading in English. Students’ achievement levels were markedly different between the countries, raising questions about the effects of high curriculum expectations and the place of English as the language of instruction in countries with different cultural contexts. In both countries weak readers were able to demonstrate a range of reading skills in response to accessible texts, however even the best readers were overly reliant on word matching strategies. High rates of guessing or missing responses to some items also revealed particular weaknesses in students’ skills such as generating and supporting opinions or locating evidence to support a point of view.

The findings have implications for teaching practices, curriculum design and aid-funded interventions in developing countries where the medium of instruction is a foreign language for most students.

Posters 12 - 15

12. Helping ESOL teachers transcend borders: the case of Teaching Knowledge Test

Hanan Khalifa & Nick Saville
University of Cambridge ESOL Examinations

Cambridge ESOL has a commitment to professionalism in all aspects of assessment, and in relation to professional conduct this can be illustrated by the newly developed “Teaching Knowledge Test” (TKT).

In recent years, English language learning and teaching have taken a central role in educational reform worldwide, with English introduced at an early stage of school education and more and more teachers without ELT qualifications being asked to teach English. In response to this situation, Cambridge has developed TKT to cater more closely to the needs of such teachers. It is aimed at teachers who have a minimum language competence of CEFR B1 level and its goal is to test candidates’ knowledge of concepts related to language, language use and the background to language teaching and learning.

The presentation will discuss fairness and professionalism by focusing on:
- Cambridge ESOL’s model-based approach to test development in which validation and research play a key role;
- The teaching knowledge construct and its operationalisation in the context of TKT;
- Research activities and quality assurance procedures followed during TKT development, trialling, and live administration phases with specific reference to findings from the trialling phase (n = 1020, test takers from Latin America, Asia and Europe).

The presenters will conclude with an overview of on-going validation activities and projected future research studies underlying TKT.
13. Recognition Speed for Basic Vocabulary

David Coulson
Swansea University

My research focuses on the development and validation of a computerized test of word recognition. The main hypothesis is that accessibility of words is a basic dimension of L2 lexicons, along with size and associational organization. It is known that the more advanced an L2 speaker becomes, the faster word recognition becomes. However, practical methods of measuring word recognition ability for use in classroom assessment have not been developed. One of the reasons for this is the difficulty in accurately measuring the typical recognition latency of words (less than 500 milliseconds) with non-specialist equipment. The test format which I will describe overcomes this problem by placing each item in a mask, for example: ‘asolfamlypedede’. The delay caused by searching the string allows reasonably accurate measurement on a personal computer. The benefit of this test is that each item has been normed against native speaker performance. Subjects score 1 point if they answer within this norm.

One large problem is that in repeated tests, items recognized by L2 learners are not necessarily recognized again within native norms later. This may be partly a characteristic of the test, or may reflect an underlying instability in lexical accessibility. Concerning the former possibility, I am investigating the validity of the test items by examining the effect of various kinds of masks on recognition facility. Once validity is established, I can check whether there is a basic threshold of competence and the nature of change in word recognition ability over time.

14. The effect and validity of a multiple-choice and open-ended listening test

Akiyo Hirai
University of Tsukuba

Multiple-choice (MC) test format is more commonly used for large scale testing due to its high practicality for scoring than open-ended (OE) format. However, whether it is a valid measure to assess learners’ communicative competence is often criticized (e.g., Weiping & Juan, 2005) along with a heightened trend of more authentic integrated-skills testing. It is also questionable whether an MC listening test can cover high proficiency learners’ listening abilities accurately (Hirai, 2005). This paper, based on Hirai’s pilot study, examines the effect and validity of MC and OE combined tests on listening performance. First, item difficulty and item discrimination of stem-equivalent MC and OE items were compared by using the two-parameter IRT model. For validation of these two formats, correlational and factor analyses were employed with reading, grammar, spelling, and receptive and productive vocabulary test scores. The results showed that the item difficulty and item discrimination values of most of the OE items were higher than those of the corresponding MC items, indicating that an MC and OE combined test can cover a wider range of learners’ listening abilities than an MC only test without increasing the number of items. The study further examines individual items with respect to question type as well as item difficulty and discriminatory power in order to suggest how we can combine MC and OE tasks to best measure learners’ listening abilities.

15. Potential 'inequalities' in language tasks and human brain imaging techniques

Toru Kinoshita, Nagoya University
Harumi Oishi, Gifu Shotoku University

Since the late 90s, researchers of L1/L2 acquisition have utilized technological breakthroughs in brain sciences focusing on imaging techniques of higher brain functions. However, studies are limited in number and research coverage. The issue of hemispheric specialization in bilinguals in relation to age of L2 onset is discussed in many studies. These studies generally overlook the issue of relative distance between learners’ native and target languages. Even
fewer studies incorporate brain science perspectives into language assessment. Such studies are limited to assessing abilities of participants with traumatic brain injury. Virtually no study has attempted to address issues of language assessment for learners with normal brains in consideration of L1 and L2 language relationships. Using the brain imaging technique of functional near-infrared spectroscopy, this study compares the cortical activation patterns in classical language areas of two groups of normal EFL learners. Blood flow to Wernicke's area and angular-gyrus/ supramarginal gyrus in twelve participants with native languages in Indo-European Languages (IEL) group and fourteen participants whose L1s belong to Asian-Pacific, non-IEL, language group was monitored as they engaged in listening tests in English. Results show the Asian-Pacific group needing significantly more blood concentration (i.e. more attention resources) than the IEL participants, although the two groups were equivalent regarding listening proficiency on the TOEIC test. This suggests the presence of another test-taker characteristic which brain science could reveal. The detected difference is a possible source of explanation for differences between the groups' performances in same-level programs and tests in various educational settings.

Session 8
Saturday July 1, 11.00am-12.30pm
GM15 – Law Building

Papers 28 - 30

28. Testers' intention, specialists' presumption, and Stakeholders' perception of reading test Items

Hossein Farhady
Masoumeh Shahmohammadi
Iran University of Science and Technology

Tests such as TOEFL, IELTS, MELAB, and CAEL are internationally used for certification and admission purposes. Therefore, the validity of the inferences made on the basis of the scores of such tests plays an important role in the future career of the stakeholders around the world in general, and in Asia/Pacific regions in particular. The purpose of this study was to investigate the validity of the reading section of the IELTS regarding the correspondence between what the test is claimed to measure and what it actually measures. To this end, a reading sub-skills checklist along with a copy of the test was prepared and distributed among 17 testing specialists around the world and 60 advanced EFL test takers. In addition, the information obtained from a cursory review of the test manual was taken into account to form the following three sets of data:
1. What the test developers claimed in the test manual that the test items are intended to measure,
2. What testing specialists presumed that the items measured, and
3. What the test takers perceived that the test items measured.
The triangulation of data made both qualitative and quantitative analyses possible. The findings revealed significant mismatches across and within the groups of participants calling for a more comprehensive research on the validity of the test.
29. Criterial parameters for establishing writing proficiency levels within the Common European Framework: a retrospective analysis

Cyril Weir
University of Luton
Sara Gysen
Universiteit Leuven/Centrum voor Taal en Migratie

A number of contextual parameters were identified as being potentially criterial in making distinctions between adjacent exams along a proficiency continuum after recourse to the literature and from a retrospective analysis of previous test papers across the proficiency range. Context validity addresses the particular performance conditions under which a task is to be performed (such as purpose of the task, time available, expected length, writer reader relationship, known marking criteria as well as linguistic demands: functional, lexical, structural and discoursal).

Cognitive parameters that might be useful for the surface analysis of ESOL examinations at different levels of ability were also identified. For example macro-planning, organisation and monitoring, editing and revision were considered potentially useful in distinguishing across the skilled /unskilled writer continuum (Kellogg 1994, 1996, Scardamalia and Bereiter 1987, Field 2004, 2005).

We report on pilot studies investigating the performance parameters that appear to help ground distinctions between different levels of proficiency in writing in ESOL examinations. A number of test tasks were examined across the proficiency continuum in 3 workshops with around 80 European language testers. They were asked to judge whether a variety of individual contextual and cognitive parameters enabled them to differentiate between adjacent exams along a proficiency continuum.

The presentation reports on consensus across these workshops as to where these previously established contextual and cognitive parameters appeared to help differentiate between different level writing tests.

30. Group dynamics in language test development

Jung Tae Kim, Hyeong-Jong Lee, Jinshu Li & Carsten Wilmes
University of Illinois at Urbana-Champaign

Other than Davidson and Lynch (2002, Chapter 6), there is little published work on the interaction of team members while developing a language test. The study addresses that gap. We believe that quality test development process is the backdrop of any language testing practice, and we believe that better understanding of test development team members’ interaction and roles is an important foundation of that quality.

We analyze several video recordings of interaction between members of a test development team. We investigate the group dynamics on language test development practice in terms of achieving validity and reliability of a language test. This is a highly experienced testing team (unlike the teams analyzed by Davidson and Lynch), and our work reveals how efficiencies of their relationships contribute to cyclical test development and improvement.

We illustrate the team’s workings through audits. An audit is a written and video record of a particular testing problem or challenge, showing how it was identified, debated, and resolved in team discussions and in interactions with external stakeholders. For example, one audit details mis-communication between the team and external stakeholders about the test purpose for a pre-instructional language placement exam. The stakeholders believed that the test purpose – placement – was the primary shaping force of the exam, whereas our team’s evolving test specifications also identified other elements shaping the test design, in this case, logistics and teaching philosophy.
Session 9
Saturday, July 1, 2.30pm-4.00pm
GM15 – Law Building

Papers 31-33

31. Indigenous assessment criteria: Refining the concept

M. Usman Erdősy
Carleton University

The concept of indigenous assessment criteria, "used by subject specialists in assessing ... apprentices in academic or vocational fields" (Douglas, 2000) is intuitively appealing and enjoys empirical backing. The general characteristics of indigenous assessment criteria are also coming into focus: they are context-specific, they are seldom articulated in the abstract, they focus on task fulfilment, and they allow raters to award extreme scores or attach disproportionate weight to the fulfillment of certain tasks.

However, several issues surrounding indigenous assessment criteria need to be resolved. This paper will address two of these with reference to empirical data. First, few indigenous assessment criteria emerging from empirical studies (Douglas & Myers, 2000; Jacoby & McNamara, 1999) appear restricted to a single community of practice. Second, and of particular interest to language testing, indigenous criteria are predominantly non-linguistic, yet linguistic competence is crucial to satisfying the content-based criteria that indigenous assessment overtly focuses on.

I explore these two issues – identifying guidelines for classifying criteria as indigenous, and assigning a role for linguistic criteria in indigenous assessment –through a study of the assessment of written work in an undergraduate history class. This study, based on data from interviews, verbal protocols, and analyses of written documents, followed two important guidelines also emerging from research into indigenous criteria: it situated assessment in the broader discourse of a community of practice, and was undertaken by a researcher with insider knowledge which allowed him to follow the discussions of substantive issues within which indigenous assessment criteria were embedded.

32. The use of IELTS for university selection: an Australian study

Kieran O’Loughlin
University of Melbourne

In recent years there has been growing interest in the consequential aspect of construct validity, especially in relation to 'high stakes' language tests. A major challenge for research in this area is to begin to understand how test scores are interpreted and used in specific contexts. This paper addresses this question by reporting on a study (funded by IELTS Australia) which examined the use of IELTS in selecting international, full fee-paying students to a large faculty at a major Australian university. Using institutional documents, questionnaires and interviews the study attempted to firstly, map how IELTS was used in the selection process and secondly, explore the knowledge and beliefs which staff (both administrative and academic) and students had about the test. Key findings included 1) a generally high level of procedural compliance with university policy and procedures on the part of staff directly involved in selection; 2) stronger knowledge about the IELTS (both the test and the scores it produces) amongst students than staff; 3) a wide range of views amongst both stakeholder groups about the appropriacy of minimum IELTS entry scores specified by the faculty; 4) a lack of understanding amongst both groups as to what different IELTS scores imply about a student’s readiness for university study, including their need for further English development; 6) a lack of clearly established equivalence between the IELTS and other acceptable measures of English proficiency in university selection policy. The theoretical and practical implications of these findings will be discussed in detail.
In specific purposes performance testing, best practices demand that test tasks be representative of those in the target language use situation, based on a needs analysis (MacNamara 1996, Douglas 2000). In the domain of Aviation English, Teasdale (1996) describes such procedures used in the development of a performance test for European Air Traffic Controllers. However, recent international policy changes by the International Civil Aviation Authority (ICAO) have imposed new requirements for Aviation English, which assume that performance on common English tasks related to aviation topics will accurately predict language performance on the job (ICAO Document 9835).

This study investigates the variation in oral proficiency demonstrated by 15 Asian Air Traffic Controllers across two types of testing tasks: radio telephony-based tasks and non-specific English tasks on aviation topics. The test-takers were from China, Korea, Mongolian, and Japan. Their performance was compared statistically in terms of ratings on the six ICAO skill areas (pronunciation, structure vocabulary, fluency, comprehension, and interactions), as well as qualitatively through a discourse analysis of the candidates’ communicative competence.

The results indicate that there were major differences in the performance of the test-takers across task types and that these differences were not predictable across subjects. Furthermore, the differences between “general English proficiency” and specific purpose proficiency were even greater than those we would expect for other LSP situations.

We discuss the implications of these findings for assessing Aviation English using ICAO standards and draw conclusions about the influence of political contexts on LSP testing.
Research Networks Session
Friday, June 30, 10.35am-11.50am
Rooms G27 & G29 - Law Building

1. Relaunching the Graduating Students’ Language Proficiency Assessment in Hong Kong

Alan Urmston & Mark Knight
Hong Kong Polytechnic University

GSLPA (Graduating Students’ Language Proficiency Assessment) is an assessment of the spoken and written English language proficiency of graduates from a university in Hong Kong. It was developed, trialed and administered in the late 1990s / early 2000s. However, the decision of the Hong Kong Government to endorse IELTS as a de-facto university English exit test led to the decision to withdraw the GSLPA. Changes in the language enhancement policies at the Hong Kong Polytechnic University have meant that the GSLPA is to be reintroduced and approximately 5000 students will take the Assessment each year, beginning in 2006/07. The graduates will be able to use their results in the GSLPA to aid their search for employment post-graduation.

Whilst the GSLPA has a solid grounding in terms of its validity and reliability, thanks to the work of the team which developed it, and tasks and procedures are already in place, there are new challenges for the current team to face. These include the implementation of on-screen marking of the written language test and the provision of administering the spoken language test in a language lab to a large population of test-takers. In addition, the GSLPA has to be ‘sold’ to its end users so that it can be established as the English language assessment of choice for university graduates in Hong Kong and beyond. In this session, the presenters will discuss these issues with interested colleagues.

2. The iterative process of construct definition, observed behavior, and scale development

Sarah L. Briggs and India C. Plough
University of Michigan

In an attempt to gain a better understanding of the dynamic interaction among the different components that comprise the speaking construct of an oral performance test currently used to assess the language competence of prospective university instructors, empirically-based and intuitive methods were employed to revise the rating scale to more accurately reflect those components. This Work in Progress is motivated by observations that the influence, or weight, of various components of the speaking construct on an evaluator’s final rating varies based, in part, on the task, the rater, and the relative presence or absence of other components in the oral interaction. Empirical methods have been adopted to analyze qualitatively and quantitatively the recordings and transcriptions of the oral performance tests of twenty candidates, ten of whom received ratings “approved for teaching” and ten of whom received ratings “not approved for teaching.” Initial coding focuses on the relative effects of specific pragmatic, textual, and phonological features of language required during tasks that require both transactional and interactional language use. Discriminant analysis is used to examine the extent to which these linguistic features contribute to final ratings. Results are compared to the evaluators’ “real-time” comments on the evaluation sheets of the aforementioned twenty candidates. Complementing this approach, intuitive methods have been adopted to explore those features of candidate performance attended to by evaluators. Interviews are being conducted with experienced evaluators to determine the salient features they judge necessary to communicate effectively in instructional settings.
3. Investigating Item Writers as Informants on Second/Foreign Language Development

University of Michigan, ELI

Alderson (2005) speculated that an investigation into the knowledge and experience of professional item writers who write language tests might be able to contribute to a better understanding of second/foreign language development. In this networking session, the researchers report on a pilot study on this topic.

The research questions for this preliminary study are:

1. To what extent can professional item writers predict item difficulty and discrimination?
2. To what extent can professional item writers predict distracter attractiveness?
3. To what extent can professional item writers predict the process test takers use to select answers?
4. What factors influence item-writer predictions?
5. What item features contribute to successful item-writer prediction?

Three professional item writers were asked to review 30 multiple-choice grammar items and two reading comprehension passages each with five to eight multiple-choice questions. The items were designed for intermediate to advanced level language learners. For each item, they predicted item difficulty, discrimination, and the attractiveness of each distracter. To answer the first three research questions, item-writer predictions were compared with classical item statistics from administering the items to the target population. A combination of focus groups and interviews were conducted to collect data on the processes test takers used for answer selection and the reasoning used by item writers when making predictions.

The researchers will report on the results of their preliminary study and seek input on how best to design future studies to tap item writers’ knowledge related to the process of second/foreign language development.

4. Validity of utilising group oral assessment in Japanese upper-secondary schools

Fumiyu Nakatsuha
Essex University

Assessments of oral production are sometimes not adapted due to anticipated validity/reliability concerns and practicality constraints. In the Japanese TEFL context too, although the course of study for secondary schools primarily aims at developing "practical communication ability" emphasising aural/oral ability, paper-and-pencil formats are mostly the only tool in university entrance examinations and even in low-stakes classroom tests. My ongoing PhD research, therefore, investigates validity evidence for utilising a group oral assessment for an upper-secondary school achievement-test purpose, so that school teachers and developers of university entrance examinations can be informed to what degree the format is valid in their contexts. The format is chosen because of its practicality and its theoretical and pedagogical desirability to assess communicative ability.

To provide an integrated evaluation of different but complementary sources of validity evidence, supporting a unitary validity property (Messick, 1989), this research examines content, scoring (Weir, 2005) and consequential validity. Data was collected from speaking tests in groups of 3-4 taken by 300 students and by questionnaires to students/teachers in upper-secondary schools with varied academic levels in different regions. Analyses are currently in progress. Firstly, for assessing content validity, language functions which
teachers want their students to achieve are compared with those observed in the test transcripts. Secondly, scoring validity is examined in relation to candidate characteristics as construct-irrelevant test variances. Finally, the reaction of students/teachers to the test and possible impacts of introducing such an oral component into classroom tests and university entrance examinations are discussed.

5. The washback of year 12 Japanese Second Language examinations on teaching

Akiko Ryumon
Monash University

Washback research has focused primarily on ESL, and research on other languages such as Japanese is underdeveloped. Moreover, the washback for the Japanese VCE (year 12) examinations in Victoria has not been examined previously, despite the high-stakes nature of the tests.

This paper presents the results of a small study which examined the Japanese VCE assessment in terms of its effects on teaching, and outlines plans for a larger study looking at the impact of assessment on both teaching and learning. The first study investigated teachers’ perceptions of how the VCE curriculum is implemented in classrooms, with specific reference to the relationships between curriculum, teaching and assessment. The study utilised interviews with VCE teachers and the director of an organisation supporting Japanese teachers, as well as an analysis of assessment and curriculum documents to explore factors that are affected by washback, and factors that affect washback. It identified a gap between externally prescribed curriculum and implemented courses and argued that this gap derives to a large degree from the washback of the VCE assessment on teaching.

In a larger follow-up project, it is planned to conduct a longitudinal classroom observation in conjunction with follow-up interviews with teachers and students, to investigate the washback on both teaching and learning. In addition, an analysis of relevant documents, such as worksheets, internal assessment tasks and students’ work, will be carried out to identify the mechanisms and extent of washback in the VCE context.

6. Exploring Native- and Nonnative-Teachers’ Decision-making Behaviour in the Assessment of CET-SET

Ying Zhang
Monash University

The question of whether the native speaker should serve as benchmark in the teaching/testing of English as a second/foreign language has been the subject of heated debate (Jenkins, 2000; Seidhoffer 2001, Lazaraton 2005) but there is thus far little research into the standards applied by NS and NNS raters in judging performance. The study explores these issues in relation to the CET-SET, a high level test of speaking proficiency in China designed to assess university students’ readiness to communicate in English in a variety of situations. It sets out to compare both the language attitudes and rating behaviour of four different groups --- non-native speaking official raters who routinely assess performance on the test (N=10), non-native speaking EFL teachers in Chinese universities (N=10), native speaking EFL teachers working in China (N=10) and native speaking EFL teachers working in Australia (N=10) — using speech samples drawn from previous administrations of the CET-SET. By analyzing the scores and verbal protocols of these different groups of raters, the author examines differences in rater orientations between and within groups and the extent to which the notion of native speaker competence is invoked in their decision-making. It is anticipated that findings of this research will provide further insight into both the nature of the CET-SET test construct and the consequences of using non-native rather than native speaker raters to assess performance.
7. Young learners in Japan: lost for words?

Miyoko Kobayashi & Nobuko Hasegawa
Kanda University of International Studies

This paper reports an attempt to develop a working word list as a basis for vocabulary tests for Japanese young learners of English. This project constitutes part of a larger-scale research programme on the relationship between the human brain and language development. Despite the heated debate about the introduction of English language teaching to Japanese primary schools, the government has so far produced no clear guidelines or objectives for teachers and materials developers, and research into what should be included in the syllabus is acutely needed. Some coursebooks and teaching materials for young learners claim that their selection of words is based on children’s first language development, but no clear warrant for such a claim seems to be evident. Recent development in corpus linguistics has offered insights into the range and type of vocabulary required for learners at different levels, and research suggests that frequency lists can inform language teaching and testing practices (Nation 2001, Read 2000). However, the vast majority of research has focused on adolescent and adult learners. This research is intended to address such questions as: how best can we develop a word list for young learners of English? To what extent do adult frequency word lists reflect children’s language development? Do formulaic expressions have a greater role for children? How can the selection of words reflect children’s social and cognitive development? Currently, we are evaluating widely-used international coursebooks and examination materials for young learners of English, as well as psycho-linguistic research into child language development.
It's the Test that Helps
Teach Communicative English

The New TOEFL® iBT Provides Broad Support for English-language Learners—and Their Teachers

The new Internet-based Test of English as a Foreign Language™ (TOEFL® iBT) utilizes tasks that integrate language skills and simulate real university life to determine if students have the English-language skills they need to succeed.

The test encourages positive washback by motivating students to learn the English language the way it's actually used in academic settings.

ETS supports English-language programmes with:
- Teacher Professional Development workshops which help programmes adapt or design an integrated skills curriculum.
- The NorthStar Series: Building Skills for the TOEFL® iBT instructional series, which help students develop the English-language skills they need to succeed.

Learn more about the TOEFL® test and other ETS resources that support English-language learning.

Web: www.ets.org/toefl
Phone: +1-609-693-2432
E-mail: internationaldevelopment@ets.org

Copyright © 2006 by Educational Testing Service. All rights reserved. ETS, the ETS logo and TOEFL are registered trademarks and TEST OF ENGLISH AS A FOREIGN LANGUAGE is a trademark of Educational Testing Service (ETS). 177988
Setting the Standard for
Measuring Communicative English

High Quality Testing and Objective Scoring Set
the New TOEFL® iBT Apart from the Rest

The new Internet-based Test of English as a Foreign Language™ (TOEFL® iBT) measures how well students have mastered the integrated language skills that communicative English requires.

TOEFL® test takers demonstrate their mastery of the English-language skills they'll use every day — in the classroom, at the registrar's office and in the bookstore. Because it simulates real university life, the TOEFL test helps admissions officers determine if students have the skills they need to succeed.

The TOEFL iBT also sets the standard by providing fair, unbiased and objective scoring that has made ETS the world leader in educational measurement:

- Multiple raters anonymously score responses.
- Raters score responses from more than one country.
- Constant monitoring by scoring leaders ensures quality control and reliability.

Learn more about the TOEFL® test.

Web: www.ets.org/toefl
Phone: +1-609-683-2432
E-mail: internationaldevelopment@ets.org

“Swinburne University of Technology
sees the new TOEFL iBT as an excellent
new way of assessing student ability in
all four macro-skills required for success
at university. Students are tested in a
way that closely reflects the way they
will use English in an academic setting.
The fact that coordination of marking for
all tests is done centrally...means this is
truly an international test of English.”

Jeffrey Smart, Director,
Swinburne International, Swinburne
University of Technology, Australia
Research Notes

Quarterly publication reporting on Cambridge ESOL's research, test development and validation

An invaluable resource with up-to-date information to support your teaching, teacher training or research activity.

Research Notes 23
Focus on technology

- Assessment systems: conceptual, human, technological
- The Cambridge ESOL Item Banking System
- ESOL Professional Support Network Extranet
- IELTS Writing: revising assessment criteria and scales (Phase 5)
- IELTS test performance data 2004
- ESOL Special Circumstances 2004: Upper Main Suite provision

Online
Articles matching your particular area of interest, searchable by:
- exam (e.g. KET, CAE)
- skill (e.g. listening, speaking)
- area (e.g. candidates, research methods)
Visit the website to find articles, subscribe or download back issues of Research Notes:

www.CambridgeESOL.org/researchnotes
Studies in Language Testing series
published jointly by: University of Cambridge ESOL Examinations / Cambridge University Press and available from booksellers
Series editors: Michael Milanovic and Cyril Weir

Volume 1 (1995)
An Investigation into the Comparability of Two Tests of English as a Foreign Language: The Cambridge-TOEFL comparability study
- Lyle F. Bachman, Fred Davidson, Katherine Ryan, Inn-Chull Choi

Volume 2 (1995)
Test Taker Characteristics and Test Performance: A structural modeling approach
- Antony John Kunnan

Volume 3 (1996)
Performance Testing, Cognition and Assessment: Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Amherst
- Michael Milanovic and Nick Saville (Eds)

Volume 4 (1996)
The Development of IELTS: A study of the effect of background knowledge on reading comprehension
- Caroline Clapham

Volume 5 (1998)
Verbal Protocol Analysis in Language Testing Research: A handbook
- Alison Green

Volume 6 (1998)
Multilingual Glossary of Language Testing Terms
- ALTE members

Volume 7 (1999)
Dictionary of Language Testing
- Alan Davies, Annie Brown, Cathie Elder, Kathryn Hill, Tom Lumley, Tim McNamara

Volume 8 (1999)
Learner Strategy Use and Performance on Language Tests: A structural equation modeling approach
- James E. Purpura

Volume 9 (2000)
Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida
- Antony John Kunnan (Ed)

Volume 10 (1999)
Issues in Computer-adaptive Testing of Reading Proficiency
- Micheline Chatthouph-Doville (Ed)

Volume 11 (2001)
Experimenting with Uncertainty: Essays in honour of Alan Davies
- C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara, K. O'Loughlin

Volume 12 (2000)
An Empirical Investigation of the Componentiality of L2 Reading in English for Academic Purposes
- Cyril Weir, Yang Huizhong, Jin Yan (Eds)

Volume 13 (2001)
The Equivalence of Direct and Semi-direct Speaking Tests
- Kieran O'Loughlin

Continuity and Innovation: Revising the Cambridge Proficiency in English examination 1913–2002
- Cyril Weir and Michael Milanovic (Eds)

Volume 16 (2004)
The Development of CELS: A modular approach to testing English language skills
- Roger Hawkey

Volume 17 (2006)
- Barry O'Sullivan

Volume 18 (2004)
- Michael Milanovic and Cyril Weir (Eds)

Volume 20 (2005)
Testing the Spoken English of Young Norwegians: A study of testing validity and the role of 'smallwords' in contributing to pupils' fluency
- Angela Hasselgreen

Volume 21 (2005)
Changing Language Teaching through Language Testing: A washback study
- Liying Cheng

Volume 22 (2005)
The Impact of High-stakes Testing on Classroom Teaching: A case study using insights from testing and innovation theory
- Dianne Wall

Impact Theory and Practice: Studies of the IELTS test and Progetto Lingue 2000
- Roger Hawkey

Forthcoming volumes:

Volume 19
IELTS Collected Papers: Research in speaking and writing assessment
- Lynda Taylor and Peter Falvey (Eds)

Volume 23
- Alan Davies

For more information: www.Cambridge.org

Research Notes - a quarterly publication reporting on matters relating to research, test development and validation within Cambridge ESOL - is available on-line at www.CambridgeESOL.org/researchnotes

www.CambridgeESOL.org
Socio-Cognitive Aspects of Second Language Learning and Teaching
The University of Auckland
April 12th–14th 2007

Is language learning essentially about cognition?
How important is the social context for effective language learning and teaching?

This conference offers three days of discussion and debate about the nature of language learning and the implications for language teaching.

Guest Speakers

Patricia Duff
University of British Columbia

Rod Ellis
University of Auckland

James Lantolf
Pennsylvania State University

Richard Schmidt
University of Hawaii at Manoa

Merrill Swain
Ontario Institute for Studies in Education

Elaine Tarone
University of Minnesota

Alison Mackay
Georgetown University

Further information:
Please see our website at http://www.arts.auckland.ac.nz/sociocog/
106  Law Building  
112  University House  
160  Richard Berry Building – Nanson Lab  
199  Arts Centre  
465  Graduate House  

Main conference venue
Symposium 1; Welcome Reception
Workshops 1 & 2
Workshop 3
Wine tasting, LAQ / LT EAB meetings