LTRC 2004
TEMECULA
CALIFORNIA

26th Annual
Language Testing Research Colloquium
of the
International Language Testing Association

Language Assessment in Schools

Program and Abstracts

Temecula Creek Inn
Temecula, California, USA

March 24-28, 2004
Acknowledgments

LTRC-2004 Organizing Committee

Co-Chairs

Lyle Bachman, *University of California, Los Angeles*

Antony Kunnan, *California State University, Los Angeles*

Nathan Carr, *California State University, Fullerton*

We would like to thank the following institutions for their financial support:

Charter College of Education, California State University, Los Angeles

Department of Applied Linguistics & TESL, UCLA

Cambridge University Press

English Language Institute, University of Michigan

TOEFL Program, Educational Testing Service

University of Cambridge ESOL Examinations

We would like to thank the following individuals for their assistance with this year’s LTRC:

**Publishers’ Exhibits**

Yoko Saito, Teachers College, *Columbia University*

**Program**

Yasuhiro Imao, Sun Young Shin, *UCLA*

**T-shirts**

Priya Abeywickrama, Tim Farnsworth, Zsuzsa Londe, *UCLA*

**Foam letters, folders, name tags, receipts, labels, catering, on-site registration**

Marely Cervantes, Danny Pittaway, Jillanne Thomas, *CSUF*

Tom Box, Victoria Byczkiewicz, Undarmaa Maamuujav, Nandini Patnaik, Kevin Welch, *CSULA*

Lorena Llosa, Yuko Haga, Hoky Min, Anna Shin, Sang Kuen Shin, Min Song, Olivia Tsai, Viphavee Vongpumvitch, *UCLA*
We would like to thank the following people who served as proposal reviewers:

Alison Bailey  University of California, Los Angeles
Geoff Brindley  Macquarie University
Frances Butler  CRESST, University of California, Los Angeles
Micheline Chalhoub-Devile  University of Iowa
Alister Cumming  Ontario Institute of Studies of the University of Toronto
Liying Cheng  Queen's University, Kingston
Fred Davidson  University of Illinois, Urbana-Champaign
Craig Devile  University of Iowa
Dan Douglas  Iowa State University
Glenn Fulcher  University of Dundee
Dorry Kenyon  Center for Applied Linguistics, Washington, D.C.
Liz Hamp-Lyons  University of Melbourne
Meg Malone  Center for Applied Linguistics, Washington, D.C.
Tim McNamara  University of Melbourne
Adrian Palmer  University of Utah
James Purpura  Teachers College, Columbia University
John Read  Victoria University of Wellington
Steven Ross  Kwansei Gakuin University
Nick Saville  University of Cambridge ESOL Examinations
Yasuyo Sawaki  Educational Testing Service, Princeton
Rob Schoonen  University of Amsterdam
Elana Shohamy  Tel Aviv University
Charles Stansfield  Second Language Testing Inc.
Diane Strong-Krause  Brigham Young University
Lynda Taylor  University of Cambridge ESOL Examinations
Carolyn Turner  McGill University
Xiaoming Xi  Educational Testing Service, Princeton
Yoshinori Watanabe  Akita University
Mikyung Wolf  University of Pittsburgh
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acknowledgements</td>
<td>ii</td>
</tr>
<tr>
<td>Site Map</td>
<td>1</td>
</tr>
<tr>
<td>Conference Program</td>
<td></td>
</tr>
<tr>
<td>Overview</td>
<td>2</td>
</tr>
<tr>
<td>Wednesday &amp; Thursday Workshops</td>
<td>3</td>
</tr>
<tr>
<td>Friday</td>
<td>4</td>
</tr>
<tr>
<td>Saturday</td>
<td>7</td>
</tr>
<tr>
<td>Sunday</td>
<td>9</td>
</tr>
<tr>
<td>Samuel Messick Memorial Lecture</td>
<td>12</td>
</tr>
<tr>
<td>ILTA Lifetime Achievement Awardee Lecture</td>
<td>13</td>
</tr>
<tr>
<td>Abstracts</td>
<td></td>
</tr>
<tr>
<td>Symposium</td>
<td>15</td>
</tr>
<tr>
<td>Closing Symposium</td>
<td>17</td>
</tr>
<tr>
<td>Papers</td>
<td>18</td>
</tr>
<tr>
<td>Posters</td>
<td>27</td>
</tr>
<tr>
<td>Works in Progress</td>
<td>38</td>
</tr>
<tr>
<td>List of Conference Participants</td>
<td>49</td>
</tr>
<tr>
<td>Advertisements</td>
<td>53</td>
</tr>
</tbody>
</table>
**LTRC 2004 Program Overview**

### WEDNESDAY, MARCH 24

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 - 12:00</td>
<td>Workshop 1 (by pre-registration only)</td>
</tr>
<tr>
<td></td>
<td>LUNCH</td>
</tr>
<tr>
<td>1:30 - 5:00</td>
<td>Workshop 1 continued</td>
</tr>
</tbody>
</table>

### THURSDAY, MARCH 25

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 - 12:00</td>
<td>Workshop 2 (by pre-registration only)</td>
</tr>
<tr>
<td>12:00 - 1:30</td>
<td>LUNCH</td>
</tr>
<tr>
<td>1:30 - 5:00</td>
<td>Workshop 2 continued</td>
</tr>
<tr>
<td>4:00 - 6:30</td>
<td>Registration</td>
</tr>
<tr>
<td>6:30 - 8:30</td>
<td>Welcoming Reception</td>
</tr>
</tbody>
</table>

### FRIDAY, MARCH 26

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:30 - 10:30</td>
<td>Opening Session</td>
</tr>
<tr>
<td>8:30 - 9:00</td>
<td>Welcome</td>
</tr>
<tr>
<td>9:00 - 10:00</td>
<td>Keynote Address</td>
</tr>
<tr>
<td>10:00 - 10:15</td>
<td>Presentation of the Messick Award</td>
</tr>
<tr>
<td>10:15 - 10:30</td>
<td>Open Discussion of issues raised in Keynote Address</td>
</tr>
<tr>
<td>10:50 - 12:20</td>
<td>Paper Session 1</td>
</tr>
<tr>
<td>12:20 - 1:50</td>
<td>GROUP PHOTOGRAPH</td>
</tr>
<tr>
<td></td>
<td>LUNCH BREAK</td>
</tr>
<tr>
<td></td>
<td>Language Assessment Quarterly EAB meeting</td>
</tr>
<tr>
<td>1:50 - 3:20</td>
<td>Paper Session 2</td>
</tr>
<tr>
<td>3:40 - 5:10</td>
<td>Paper Session 3</td>
</tr>
<tr>
<td>5:15 - 6:45</td>
<td>Poster Break-out Session</td>
</tr>
<tr>
<td>6:45 - 8:30</td>
<td>Student organized session</td>
</tr>
<tr>
<td>7:30 - 9:00</td>
<td>ILTA Executive Board meeting</td>
</tr>
</tbody>
</table>

### SATURDAY, MARCH 27

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:30 - 10:00</td>
<td>Paper Session 4</td>
</tr>
<tr>
<td>10:20 - 11:50</td>
<td>Paper Session 5</td>
</tr>
<tr>
<td>11:50 - 1:25</td>
<td>LUNCH BREAK</td>
</tr>
<tr>
<td></td>
<td>Language Testing EAB meeting</td>
</tr>
<tr>
<td>1:25 - 3:25</td>
<td>Symposium</td>
</tr>
<tr>
<td>3:45 - 5:15</td>
<td>Paper Session 6</td>
</tr>
<tr>
<td>7:00 - 11:00</td>
<td>Banquet &amp; Awards Presentations</td>
</tr>
</tbody>
</table>

### SUNDAY, MARCH 28

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:30 - 10:00</td>
<td>Works in Progress Break-out Session</td>
</tr>
<tr>
<td>10:15 - 11:15</td>
<td>ILTA Lifetime Achievement Awardee Lecture</td>
</tr>
<tr>
<td>11:15 - 1:00</td>
<td>LTRC/ILTA Business Meeting</td>
</tr>
<tr>
<td></td>
<td>LUNCH BREAK</td>
</tr>
<tr>
<td>1:00 - 4:00</td>
<td>Closing Symposium</td>
</tr>
<tr>
<td>1:00 - 2:10</td>
<td>First Session</td>
</tr>
<tr>
<td>2:25 - 4:00</td>
<td>Second Session</td>
</tr>
<tr>
<td>4:00 - 4:30</td>
<td>Closing Remarks</td>
</tr>
</tbody>
</table>
Program Detail

WEDNESDAY, MARCH 24

Workshop 1
Statistical Analyses for Language Assessment
Lyle F. Bachman, University of California, Los Angeles
Antony John Kunnan, California State University, Los Angeles

Facilitators
Priyanvada Abeywickrama and Viphae Vongpumivitch
Location: Yula 2

9:00 - 12:00  Workshop (by pre-registration only)
12:00 - 1:30  LUNCH
1:30 - 5:00   Workshop continued

THURSDAY, MARCH 25

Workshop 2
Designing and Developing a Useful Language Test
Adrian Palmer, University of Utah
Mary Ann Christison, University of Utah

Facilitators
Priyanvada Abeywickrama and Viphae Vongpumivitch
Location: Yula 2

9:00 - 12:00  Workshop (by pre-registration only)
12:00 - 1:30  LUNCH
1:30 - 5:00   Workshop continued

4:00 - 6:30  Registration  
Patio outside Yula 1 & 2

6:30 - 8:30  Welcoming Reception
Co-hosted by Cambridge ESOL, ELI, University of Michigan, and TOEFL
Lower Plaza (near the waterfall)
FRIDAY, MARCH 26

Registration: 8:00 a.m. - 6:00 p.m., Patio outside Yula 1 & 2
Publishers' Exhibits: 9:00 - 6:00 p.m., Kiva 1 & 2

8:30 - 10:30 Opening Session

8:30 - 9:00 Welcome: Antony Kunnan, California State University, Los Angeles
Nathan Carr, California State University, Fullerton
Carol Bartell, Dean, Charter College of Education, California State University, Los Angeles

Introduction of Keynote Speaker:
Lyle Bachman, University of California, Los Angeles

9:00 - 10:00 Keynote Address:
Language, learning, and assessment: Improving validity
Eva Baker, Director, National Center for Research in Evaluation, Standards, and Student Testing, University of California, Los Angeles

This is the Samuel Messick Memorial Lecture, Sponsored by the TOEFL Policy Board, Educational Testing Service, Princeton, New Jersey.

10:00 - 10:15 Presentation of the Messick Award: Mary Enright, Educational Testing Service

10:15 - 10:30 Open Discussion of issues raised in Keynote Address

10:30 - 10:50 Break

10:50 - 12:20 Paper Session 1, Chair: Abbas Mousavi

1. English language assessment and its relative importance in school accountability in California
   Simeon Slovacek, Laura Pantoja, & Wing Chi Lee
   p.24

2. English learners' performance on state-mandated content and English language assessments
   Jeff White & Lorena Llosa
   p.26

3. Linking teacher developed test results and action research findings to curriculum development
   Lucilla Lopriore & Guido Benvenuto
   p.21

12:20 - 1:50 GROUP PHOTOGRAPH
LUNCH BREAK
Language Assessment Quarterly EAB meeting

1:50 - 3:20 Paper Session 2, Chair: Chau Ho Fai

4. The lexical component of speaking test performance
   John Read
   p.23

5. Investigating the relative significance of vocabulary breadth and syntactic knowledge in the prediction of reading test performance: A structural equation modeling approach
   Toshihiko Shiotsu
   p.23

6. Measuring the knowledge of text structure in academic English as a second language
   Viphavee Vongpumivitch
   p.25
3:20 - 3:40  Break

3:40 - 5:10  Paper Session 3, Chair: Yula 1 & 2

7. Assessment of what? - Dimensions of language skills at individual and school levels  
   Lisbeth Åberg-Bengtsson, Gudrun Erickson, & Jan-Eric Gustafsson  p.18

8. The effects of visual chunks and planning time on the language quality of verbal descriptions of graphs  
   Xiaoming Xi  p.26

9. L2 group oral testing: The influence of shyness/outgoingness, match of interlocutors' proficiency level, and gender on individual scores  
   William J. Bonk & Alistair Van Moere  p.18

5:15 - 6:45  Poster Break-out Session  Kiva 3 & 4

1. Academic language-based assessment: A listening test  
   Micheline Chalhoub-Deville & Minhee Eom  p.29

2. Developing a user-friendly item bank for in-house placement and achievement/proficiency tests of university EFL learners  
   Akiyo Hirai  p.31

3. Technical innovations in the assessment of oral reading fluency in 4th graders  
   Isabella Barbier  p.27

4. Assessment of the initial stages of linguistic competence among children of immigrant families attending public schools in Italy  
   Monica Barni & Sabrina Machetti  p.27

5. Development and validation of a spoken Spanish test  
   Jared Bernstein, Elizabeth Rosenfeld, & Isabella Barbier  p.28

6. Assessing communicative competence in adult ESL  
   Daniel S. Pittaway & Aaron J. Kelly  p.33

7. The design, development, and validation of a placement test for the Community English Program at Teachers College, Columbia University  
   Kirby Cook, Hyunjoo Kim, Jee Wha Kim, Alick Liao, E. K. Park, Taejoon Park, James Purpura, Mary Regan, Yoko Saito, Elvis Waagner, Cynthia Wiseman  p.30

8. TOEFL influence in the ESL Placement Test (EPT) at UIUC  
   Hui-Jeong Woo  p.36

9. The role of cloze in a model of foreign language proficiency  
   Jeff S. Johnson, Amy D. Yamashiro, & Jing Yu  p.32

10. Developing a rating scale from the testing context  
    Lia Plakans  p.34

11. Systematic placement testing and its validity for university English language programs  
    Tomoko Fujita  p.30
12. Assessment of outcome in an advanced ESL writing course and its admission policy implications
   Doreen Bayliss

13. An adaptation of the Modern Language Aptitude Test for Spanish-speaking children
   Daniel J. Reed & Charles W. Stansfield

14. Training nonnative speakers of English to perform holistic scoring of same language group (Chinese) nonnative speakers speech samples in the LanguEdge™ Instructional Tool
   Marcia J. Reeves, Theresa Richter, & Terry Cryan

15. Empirical Bayes subscore augmentation: A micro and macro look
   Paul Jaquith

16. Using item response theory to create a common scale for comparing results from large-scale prefectural English proficiency tests
   Chisato Suida & Tamaki Hattori

17. Eigoquest: A game/test providing proficiency measurement and, hopefully, motivation
   Randy Thrasher & Katsumi Yuasa

18. Validation of speaking performance measures: The case of Japanese learners of English with low speaking ability
   Rie Koizumi

19. The Iowa Test of English Language Learning: Vocabulary, writing, and structure & expression
   Micheline Chalhoub-Deville, Craig Deville, Anne Cummings, & Shelley Fairbairn

20. Language test preparation over the Internet - High stakes interpreter testing de-mystified
   Kathy Tinios-Rose & Christian N. Jones

21. The relationship of performance on the UIUC English as a Second Language Placement Test (EPT) to scores on the Test of English as a Foreign Language (TOEFL)
   Youngmi Yun

22. Complementary role of two performance tests in schools
   Yuji Nakamura

23. WebLAS Japanese Project
   Shoichi Gregory Kamei, Asako Hayashi, Yuko Haga, & Yasuhiro Imao

6:45 - 8:30  Student organized session  Yula 1 & 2
7:30 - 9:00  ILTA Executive Board meeting  Temecula Creek Inn Restaurant
SATURDAY, MARCH 27

Registration: 8:00 a.m. - 5:00 p.m., Patio outside Yula 1 & 2
Publishers' Exhibits: 9:00 - 5:00 p.m., Kiva 1 & 2

<table>
<thead>
<tr>
<th>8:30 - 10:00</th>
<th>Paper Session 4, Chair:</th>
<th>Yula 1 &amp; 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.</td>
<td>The challenges of the Ontario secondary school literacy tests for ESL/ELD students</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Liying Cheng</td>
<td>p.20</td>
</tr>
<tr>
<td>11.</td>
<td>Portfolio assessment in practice</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nishita Prasad &amp; Jo Lewkowicz</td>
<td>p.22</td>
</tr>
<tr>
<td>12.</td>
<td>Revising assessment criteria and rating scales for assessing L2 writing: The contribution of quantitative analyses in the validation phase</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lynda Taylor, Stuart Shaw, &amp; Graeme Bridges</td>
<td>p.24</td>
</tr>
</tbody>
</table>

10:00 - 10:20 Break

<table>
<thead>
<tr>
<th>10:20 - 11:50</th>
<th>Paper Session 5, Chair: Bernadette Stoneman</th>
<th>Yula 1 &amp; 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.</td>
<td>What does it take to get a high score on a paired task? Examining discourse, scores and the construct of L2 proficiency</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Carolyn E. Turner &amp; Joanna White</td>
<td>p.25</td>
</tr>
<tr>
<td>14.</td>
<td>Insights into the construct(s): Paired oral proficiency testing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lindsay Brooks</td>
<td>p.19</td>
</tr>
<tr>
<td>15.</td>
<td>Modeling development of spoken language ability</td>
<td></td>
</tr>
<tr>
<td></td>
<td>John H.A.L. de Jong</td>
<td>p.20</td>
</tr>
</tbody>
</table>

11:50 - 1:25 LUNCH BREAK

Language Testing EAB meeting Stone House

<table>
<thead>
<tr>
<th>1:25 - 3:25</th>
<th>Symposium</th>
<th>Yula 1 &amp; 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NCLB and its implications for the testing practices of English language learners in U.S. schools</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Organizer: Craig Deville</td>
<td>p. 15</td>
</tr>
</tbody>
</table>

Presentations:

Overview of ITELL, the Iowa Test of English Language Learners
Micheline Chalhoub-Deville, Stephen Dunbar, Minhee Eom, Shelley Fairbairn, Dongmei Li, Lei Wan, & Anne Cumnings

From needs analysis to test development in a test for ELLs
Sari Luoma, Yeonsuk Cho, Gary Buck

Approach to test development for the ACCESS for ELLs test of the WIDA Consortium
Jim Bauman & Dorry Kenyon

Discussants: Alison Bailey & Frances Butler

3:25 - 3:45 Break
**LTWC 2004**  
**Program:** Saturday

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Chair</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>3:45 - 5:15</td>
<td><strong>Paper Session 6, Chair:</strong> Yula 1 &amp; 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17.</td>
<td>Using the ELD Assessment Portfolio: An investigation of teachers' interpretations of ELD standards and students' language ability</td>
<td>Lorena Llosa</td>
<td>p.21</td>
</tr>
<tr>
<td>18.</td>
<td>Issues in developing a TOEFL validity argument</td>
<td>Carol A. Chapelle, Mary K. Enright, &amp; Joan Jamieson</td>
<td>p.19</td>
</tr>
<tr>
<td>7:00 - 11:00</td>
<td><strong>Banquet &amp; Awards Presentations</strong></td>
<td>Outside, near the Stone House</td>
<td></td>
</tr>
</tbody>
</table>

*Music by the CSULA Brazilian/Latin Jazz Combo*

**Awards**

- IELTS Master's Thesis Award, 2004
- Jaqueline A. Ross Dissertation Award for Outstanding Dissertation Research on Second/Foreign Language Testing, 2004
- Lado Award for Outstanding Student Paper at LTRC 2004
- ILTA Award Outstanding Publication in the Field of Language Testing, 2003
- ILTA-UCLES Lifetime Achievement Award, 2004

**Entertainment**
**SUNDAY, MARCH 28**

**Registration:** 8:00 a.m. - 1:00 p.m., Patio outside Yula 1 & 2

<table>
<thead>
<tr>
<th>8:30 - 10:00</th>
<th><strong>Works in Progress Break-out Session</strong></th>
<th>Kiva 3 &amp; 4</th>
</tr>
</thead>
</table>
| 1.           | Regular expressions in automated scoring: Uses and limitations  
*Shoichi Gregory Kamei & Yuko Haga* | p.43 |
| 2.           | Construct validation of an integrated, portfolio-based, and computerized EAP placement test: A multiple data source approach  
*Young-Ju Lee* | p.44 |
| 3.           | Language learning strategy use and language performance for Chinese learners of English  
*Xiaomei Song* | p.46 |
| 4.           | An investigation into reading test item characteristics: How text difficulties and question types interact  
*Tomoko Wada* | p.47 |
| 5.           | Reflective item types in The Canadian Academic English Language (CAEL) Assessment: An examination of the method effect  
*Wendy Fraser* | p.41 |
| 6.           | Addressing the challenges of assessing ELLs in the K-12 context  
*Shelley Fairbairn* | p.41 |
| 7.           | Responding to non-native writers of English: Grading and feedback in an undergraduate history course  
*M. Usman Erdosy* | p.40 |
| 8.           | Development and validation of a delivery software for the assessment of oral proficiency of adult ESL learners  
*Seyyed Abbas Mousavi* | p.44 |
| 9.           | CELPE-Bras exam and teachers’ perceptions of impact on teaching/assessment of Portuguese as a FL/L2  
*Matilde Virginia Ricardi Scaramucci* | p.46 |
| 10.          | Training the educators of the Federated States of Micronesia in holistic writing assessment, implications for the nation’s multilingual writing goals  
*Samantha Ragan & Mary L. Spencer* | p.44 |
| 11.          | The washback of a new testing system on students preparing for Japanese university entrance examinations  
*Mark Chapman* | p.38 |
| 12.          | Development and progress of the Mt. SAC ESL-CAP Test  
*Margaret Teske & Liza Becker* | p.47 |
| 13.          | Invaluable use of verbal reports for affirmation of items  
*Karen Englander & Leonora Velasco* | p.39 |
| 14.          | The continuum of the listening process from a less skilled to a skilled FL listener  
*Joanna Ankar* | p.38 |
LTRC 2004
Program: Sunday

15. Challenges of World Englishes to language testing in EFL contexts
   Hyun-Ju Kim
   p.43

16. An investigation into student negative reactions to peer assessing EFL individual presentation performance
   Hidetoshi Saito
   p.45

17. Coping with the impossible... - A collaborative approach to national test development
   Gudrun Erickson
   p.40

18. Rating video oral proficiency instruments
   Gene Halleck
   p.42

19. Investigating the relationships among responses to open-ended and multiple-choice questions on an L2 reading test and performance on an L2 writing test
   Glayol Ekbatani
   p.39

20. Measuring English writing achievement across proficiency levels using portfolios
   Diane Strong-Krause
   p.47

21. A Spanish reading screening test for a Master’s program in Mexico
   Diana Jenkins & Harold Ormsby
   p.42

22. Use of cloze tests in teacher certification in Micronesia
   Daniel L. Robertson
   p.45

23. An investigation of the competencies of intermediate proficiency language learners
   Nathan T. Carr
   p.38

10:00 - 10:15 Break

10:15 - 11:15 ILTA Lifetime Achievement Awardee Lecture
   Building and supporting a case for test utilization
   Lyle Bachman, UCLA
   p. 14

11:15 - 1:00 LTRC/ILTA Business Meeting
   Yula 1 & 2

LUNCH BREAK

1:00 - 4:00 Closing Symposium:
   International perspectives on language assessment in schools
   Organizer: Lyle F. Bachman
   p.17

1:00 - 2:10 Presentation: First Session
   Changing tests for teachers: Classroom-based English language assessment in Hong Kong
   Chris Davison, Hong Kong University

   Curriculum standardization, assessment and achievement: estranged bedfellows
   Constant Leung, King’s College, London University

2:10 - 2:25 Break
2:25 - 4:00  
**Presentation: Second Session**

Educational reform and chalk-face change: Classroom language assessment in Australian schools  
Penny McKay, *Queensland University of Technology*

Balthazar's Feast: assessment opportunities and risks  
Pauline Rea-Dickins, *University of Bristol*

**Discussant**  
Geoff Brindley, *Macquarie University*

4:00 - 4:30  
**Closing Remarks**  
Yula 1 & 2

Lyle Bachman, *UCLA*  
Antony Kunnan, *CSULA*  
Nathan Carr, *CSUF*
Language, learning, and assessment: Improving validity

Eva Baker

University of California, Los Angeles
National Center for Research on Evaluation, Standards, and Student Testing (CRESST)

Abstract

What is the theory of action of accountability policy? Will assessment results improve knowledge management? How functional are concepts of alignment, motivation, and the relationship of test performance and learning? In answer, a research-based assessment system will be proposed with a plan for assembling validity evidence of learning, transfer, and language development.

Eva L. Baker is a professor at the University of California, Los Angeles, Graduate School of Education & Information Studies. She is co-director of the National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Her research addresses assessment and accountability models, the design and validation of technology-based learning and assessment systems, and new models to measure complex human performance in large-scale assessments. Professor Baker is involved in international, national, and state policy deliberations on assessment. She was co-chair of the Joint Committee on the Revision of the Standards for Educational and Psychological Testing published in 1999. The recipient of numerous awards, she provides advice and research on educational testing design, impact, and technology to the international community, states, and educational organizations. She is the current chair of the Board on Testing and Assessment of the National Research Council. She has an extensive list of publications.
ILTA-UCLES Lifetime Achievement Award 2004
Awarded to Professor Lyle Bachman

Lyle F. Bachman is Professor and Chair of Applied Linguistics and TESL at the University of California, Los Angeles. After earning an AB in English literature from Indiana University, he served for two years as a Peace Corps Volunteer teaching ESL to high school students in the Philippines. He earned his Ph.D. in English language from Indiana University, where he was a Peace Corps Fellow and a Woodrow Wilson Dissertation Fellow. He has held faculty positions at the University of Hawai'i at Manoa, the University of Illinois at Urbana-Champaign, the Chinese University of Hong Kong, and has taught courses at several TESOL Summer Institutes. His extensive overseas experience includes serving as a Project Specialist for the Ford Foundation in Thailand, as Director of the Illinois-Tehran TEFL Internship Program at Tehran University, and as Chair Professor and Director of the English Language Teaching Unit of the Chinese University of Hong Kong. He continues to serve as a consultant to test development programs in several countries and is an active member of AAAL, ILTA, AERA, NCME, and TESOL. He is a Past President of the American Association for Applied Linguistics and the International Language Testing Association. He was the first recipient of the TESOL Research Interest Section/Newbury House Distinguished Research Award, and has been awarded the Modern Language Association of America's Mildenberger Prize for Outstanding Research Publication in the Field of Teaching Foreign Languages and Literatures twice. He is a past Co-Editor of the journal Language Testing, and currently serves as Co-Editor of the Cambridge Language Assessment Series. His current research interests include validation theory, linking current validity models to test utilization, issues in assessing the academic achievement and academic English of ELLs in schools, the interface between language testing research and second language acquisition research, and the dialectic of constructs and contexts in language testing.

He has contributed numerous publications to the field, including the following books:
An Investigation into the Comparability of Two Tests of English as a Foreign Language: The Cambridge-TOEFL Comparability Study (with Fred Davidson, Katherine Ryan and In-Choel Choi), University of Cambridge Local Examinations Syndicate and Cambridge University Press, 1994.
Language Testing in Practice (with Adrian S. Palmer). Oxford University Press, 1996.
ILTA-UCLES Lifetime Achievement Award Lecture
Mar, 28th, 10:15 - 11:15

Building and supporting a case for test utilization

Lyle F. Bachman
University of California, Los Angeles

Abstract

The fields of language testing and educational and psychological measurement have not, as yet, developed a set of principles and procedures for linking test scores and score-based inferences to test use and the consequences of test use. While Messick (1989), discusses test use and consequences, this framework provides virtually no guidance on how to go about investigating these in the course of practical test development. Argument-based formulations of validity (e.g. Kane, 1992; Kane, Crooks, & Cohen, 1999; R. Mislevy, 2003; R. J. Mislevy, Steinberg, & Almond, 2003) provide a logic and set of procedures for investigating and supporting claims about score-based inferences, but say nothing about test use and the consequences of test use. Recent formulations in language testing (e.g. Bachman & Palmer, 1996; Kunnan, 2003; Lynch, 2001) are essentially lists of more or less independent qualities and questions, with no clear mechanism for integrating these into a set of procedures for test developers and users to follow. Finally, what has been called “critical language testing” (e.g. Shohamy, 1999, 2001) brings us no closer to a theory of test use. This perspective has alerted us to the political uses and abuses of language tests and to the need for test developers and test users alike to be self-critical of the ways in which tests are used. However, it provides little guidance about how to go about either anticipating and avoiding, or redressing, the problems with test use that it discusses, and treats consequences as essentially unrelated to the validity of inferences.

Bachman (2003) has suggested that the argument-based approach to validation should be broadened to include test use, suggesting a two-part assessment use argument. In this presentation I will introduce a general outline for how an argument for test utilization might be structured. I will argue that issues and questions that have been raised by language testers regarding uses, abuses, consequences, and fairness in language testing can provide a basis for articulating claims and counterclaims in an assessment utilization argument, and that these can be linked to claims and counterclaims about the validity of score-based inferences. I will provide examples of utilization arguments for different types of test use. In my view, an argument-based approach to test utilization can provide a substantively-grounded set of procedures for building and supporting an integrated argument that links validity and utilization for a particular assessment.
Recent U.S. legislation, idealistically dubbed ‘No Child Left Behind’ (NCLB), represents the federal government’s most far-reaching attempt to influence educational practices in each state by holding public schools accountable for the test scores of all students and by applying costly sanctions to schools whose students do not exhibit adequate yearly progress. The bill stipulates that all students in grades 3-8 and 11 be tested in both reading and math, and that all students show annual improvement. All students are to be tested, including subgroups of students, such as English Language Learners (ELLs) whom schools have often excluded from testing in the past. Students whose English language proficiency precludes them from obtaining meaningful test scores in reading and math must be administered an English language proficiency test covering the four modalities. The ELLs must demonstrate annual yearly progress on this measure with the objective that they will soon have the requisite language skills to be able to take the state’s reading and math tests in English.

Many language testers in the U.S. are involved in helping state education agencies develop an assessment tool that meets the requirements of NCLB. Most testers agree that no existing tool captures useful information about students’ academic language proficiency skills in all four modalities, can provided information about a student’s progress along a growth scale, and can be used to ‘mainstream’ ELLs into established content area testing practices. The symposium will address how language testers working in three different contexts (a separate paragraph describing each of the three follows) are addressing the needs of states and schools to assess ELLs.

Overview of ITELL, the Iowa Test of English Language Learners

Micheline Chalhoub-Deville, Stephen Dunbar, Minhee Eom, Shelley Fairbairn, Dongmei Li, Lei Wan & Anne Cummings, University of Iowa

The Iowa project started before the NCLB legislation. The genesis of the project was an interest in systematizing the decision of whether ELL students have a ‘threshold’ linguistic ability to enable them to take content area standardized tests and obtain meaningful test scores. Since the beginning, the project design and specifications were firmly grounded in an academic representation of the language construct. Language components such as vocabulary, grammar, and test literacy were the focus of the assessment. For instance, in terms of vocabulary, test design embraced the distinction made between BICS and CALF. Content and academic words are a strong feature of the vocabulary section of the test. After NCLB, the scope of the project had to be revised to accommodate the specifics of the legislation. The test currently measures all four modalities. Current specifications are based on an extensive survey of various state standards, profes-
sional content standards (social studies, math, science, ESL, and language arts), and subject area textbooks. Also, input from ELL teachers has been instrumental in designing these tests. The assessment instruments have been developed for 4 grade spans: K-2, 3-5, 6-8, and 9-12. Pilot evidence of some of these instruments supports the quality of the content and the scores obtained. Specific research findings will be discussed at the symposium.

From needs analysis to test development in a test for ELLs

Sari Luoma & Yeonsuk Cho, Ballard & Tighe Publishers
Gary Buck

Due to the requirements of NCLB, Ballard & Tighe have decided to develop a new testing system to assess the proficiency of K-12 English language learners. In order to determine the proficiency required to function in mainstream classrooms, the assessment team have carried out an extensive needs analysis: examining state standards, text books, and tests, as well as observing classes and conducting focus groups with mainstream teachers. This needs analysis has produced a huge amount of data regarding the language demands of classrooms at all grade levels, and in different subject areas. The problem is to convert this data into a test design that assesses the construct identified by the research. Establishing this connection is at the heart of construct validity, but to the best of our knowledge, there is no established, systematic methodology to convert research results into test design. This presentation discusses how the test development team dealt with this problem. This will address the desirable properties of various analysis categories, which ones were chosen and why, and then how these were used as the basis for test design. Conclusions will be drawn and implications discussed regarding the relationship between needs analyses and construct validity, both from a theoretical and a practical perspective.

Approach to test development for the ACCESS for ELLs test of the WIDA Consortium

Jim Bauman & Dorry Kenyon, Center for Applied Linguistics

The issues involved in assessing English Language Development (ELD) in NCLB are complex and there exist many approaches to address them. One approach, reported on here, involves a consortium of eight states, known as WIDA (www.wida.us), headed by the state of Wisconsin. This approach focused on creating three products: a set of Standards for ELD addressing the common needs of these states, item-level specifications for assessment development based on those Standards,
International perspectives on language assessment in schools

Organizer
Lyle F. Bachman

Presenters
Chris Davison, Hong Kong University
Constant Leung, King's College, London University
Penny McKay, Queensland University of Technology
Pauline Rea-Dickins, University of Bristol

Discussant
Geoff Brindley, Macquarie University

Changing tests for teachers: Classroom-based English language assessment in Hong Kong

Chris Davison, University of Hong Kong

This presentation will report on the changing role of classroom-based assessment in Hong Kong's new standards-referenced outcome-oriented English language curriculum. The traditional reliance on externally set and assessed tests is being overturned, but much work needs to be done to develop the assessment readiness of teachers as well as the wider school community, and to adapt and develop new constructs for old contexts.

Curriculum standardization, assessment and achievement: estranged bedfellows

Constant Leung, King's College London

School education in England has been through unprecedented, large-scale reforms in the past decade. The National Curriculum and its many associated initiatives have introduced increasing centralization and standardization in both teaching content and formal assessment to 'drive up' student achievement. Current national test results suggest that official curriculum achievement targets are over-optimistic. This presentation will trace the ways classroom teacher assessment is being re-cast as a potential saviour of the national project. The assessment of English as an Additional Language (EAL) will be considered within the context of wider national developments.

Educational reform and chalk-face change: classroom language assessment in Australian schools

Penny McKay, Queensland University of Technology

This paper draws on research into classroom language assessment in schools around Australia. In the past teachers in Australia have been given a strong professional role in classroom assessment. The paper addresses the way that classroom assessment has been affected by educational reform measures, and particularly by the introduction of standards and external literacy tests, with some resulting tension and disquiet for teachers. Some (not necessarily successful) solutions proposed by Australian educators to the negative impact of these measures are described briefly.

Balthazar's Feast: assessment opportunities and risks

Pauline Rea-Dickins, University of Bristol

This presentation first describes some of the parameters of school-based assessment, largely through the lens of one local education authority and with specific reference to learners with English as an Additional Language in mainstream schools in England. It then explores roles for learner assessment as part of language programme evaluation before examining key facets of school-based assessment through which some assessment opportunities and risks are raised.
Assessment of what? - Dimensions of language skills at individual and school levels
Lisbeth Åberg-Bengtsson, Gudrun Erickson, & Jan-Eric Gustafsson, Göteborg University

Session 3, 26th Fri., 3:40 - 5:10

The underlying assumption of this paper is that assessment of students' performance is embedded in a social, cultural, economic, and political context. Obviously, assessment entails considerable amounts of power and control at different levels of the educational system for individuals as well as for institutions.

The paper is set in the Swedish school context, in which an extensive program of national assessment with multiple aims is offered in connection with national curricula and syllabuses. The language materials in the program, developed in a collaborative process with different stakeholders, are clearly inspired by the Common European Framework of Reference, with tasks focusing on receptive, productive and interactive skills. As part of the continuous validation process of the program, various analyses of the properties of the different tests have been performed.

The so-called Åp 9 tests, i.e. a set of subject tests for English, Swedish and mathematics designed to be used at the end of compulsory school, were first administered in the spring of 1998. On this occasion individual results from approximately 30 percent of the cohort were reported to the national authorities. The sample was drawn to be representative of the country as a whole.

Educational data typically concern individuals nested within classes that are nested within schools and so forth. The research on the Åp 9 presented in this paper aimed (a) at gaining a deeper knowledge of the internal structure of the tests and (b) at separating individual performance from school performance, in both cases with particular focus on the language tests. Confirmatory factor analysis was used throughout the study.

At the individual level, a baseline model with three major subject-related correlated factors for English, Swedish and mathematics was first identified. This model is not very interesting per se, and the model-fit was not acceptable. Consequently, a number of additional factors were set up and tested. The process led to a theoretically sound, eleven-factor model demonstrating good fit. The same analyses were performed at the school level, using two-level structural equation modeling techniques, resulting in a similar, although somewhat less complex model.

Interesting results to be discussed are, e.g., listening and reading variables in relation to a more general receptive factor, and an obvious relationship between factors of linguistic creativity and mathematical communication. Furthermore, similarities and differences between the one-level and the two-level approach are discussed, i.e. the relationship between the individual and the institutional level. An attempt is also made to place the findings within a broader educational and societal framework, taking into account the crucial aspects of responsibility and consequences of educational assessment.

L2 group oral testing: The influence of shyness/outgoingness, match of interlocutors' proficiency level, and gender on individual scores
William J. Bonk, University of Colorado at Boulder
Alistair Van Moere, Kanda University of International Studies

Session 3, 26th Fri., 3:40 - 5:10

This research investigates the effects of some variables of interest on oral scores in the group oral conversation task. The oral task used is a large-scale (1900+ examinees) end-of-year assessment of individual performance for students with three or more hours per day of English instruction at a private university in Japan. Examinees have a conversation about an assigned topic in a group of three or four peers, and are independently rated by two native speaker ESL instructors on five sub-constructs of English conversational proficiency. This test format is a practicable option for general L2 oral performance testing, such as entrance or placement testing, because it is a relatively quick, easy, and resource-economical task for the kinds of inferences that can be drawn from the resulting performances (Bonk & Ockey, 2003). A combination of Rasch analysis, multiple regression, and hierarchical linear modeling is used to investigate the following variables which have the potential to influence examinees' ratings on these oral performances: rater severity, prompt, gender, shyness/outgoingness, and match (or mismatch) between examinees' oral proficiency and that of their interlocutors. Preliminary analyses revealed the following results:

Even with 2-hour training sessions before each examination for both new and returning raters, rater severity as estimated by Rasch modeling varies greatly and is a likely source of construct-irrelevant variance if it is not controlled for through some sort of modeling. Bias analysis did not reveal systematic rater bias for or against one examinee gender or prompt.

The various prompts used do not appear to differ significantly in the quality of performances they elicit based on Rasch analysis.
In a multiple regression model which predicts oral test scores using shyness/outgoingness scores using gender, and pretest proficiency ratings by instructors (plus their interactions) as predictors, only main effects of holistic ratings and shyness/outgoingness made statistically significant contributions to regression. When these two main effects are controlled for, gender and the interactions it participated in were not significant predictors of performance. The effect of shyness/outgoingness on scores was not different for individuals of differing levels of proficiency going into the test. The inclusion of nonlinear effects of this same set of predictors did not improve model fit.

We are now carrying out the HLM analysis of group proficiency levels and their potential effect on individual scores. The results of these analyses have potentially great consequences for the kinds of inferences program and test administration personnel would like to be able to make about examinees.

**Insights into the construct(s): Paired oral proficiency testing**

Lindsay Brooks, OISE/University of Toronto

**Session 5, 27th Sat., 10:20 - 11:50**

This paper presents some of the results of a larger thesis study comparing paired oral proficiency testing, in which pairs of adult international ESL students interact with each other, to individual proficiency testing, in which each student interacts with an examiner. The data were collected in the context of the exit testing from an Academic Preparation Program at a large Canadian university. Because successful completion of the program with an appropriate grade, together with passing the exit testing, satisfies the language proficiency requirements of the university, the testing is high stakes for the students and for the program. The participants (N=80) each had two speaking tests: one in the individual format and one in the paired format. Comparable speaking prompts were used in both the paired and individual tests. After the exit testing, participants responded to a questionnaire surveying their views on the two testing formats and 12 students participated in semi-structured interviews. The focus of this paper is on the quantitative and qualitative analyses of the differences in interactions between the paired oral proficiency test and the individual test format. In addition to statistical results, analyses of the dialogue of students as they interacted in both formats will be presented. The results suggest that there was a person by format interaction with some students performing differentially depending on whether they were tested in pairs or individually. More varied language functions and patterns of interaction were found in the paired format. Overall, students received higher ratings in the paired proficiency test. The findings from this study have implications for test development and theory as analyzing the language in these test formats provides insight into the construct(s) being measured.

**Issues in developing a TOEFL validity argument**

Carol A. Chapelle, Iowa State University
Mary K. Enright, Educational Testing Service
Joan Jamieson, Northern Arizona University

**Session 6, 27th Sat., 3:45 - 5:15**

Like high-stakes language tests in schools, the Test of English as a Foreign Language (TOEFL) is expected to meet the measurement requirements of the APA/AERA/NCME Standards (1999). However, any attempt to put ideal guidelines into practice reveals areas of tension among the relevant theoretical concepts as well as imprecision in the way that guidelines are expressed. Therefore, it is crucial for language testing research and practice to examine specific cases of validation arguments and the process of their development. This paper offers one such examination by explaining how the framework for the TOEFL validity argument was developed through an attempt to operationalize the approach to validity offered in the Standards, which directs validation researchers to gather “kinds of evidence” that are needed to evaluate the “intended interpretation” of test scores. The identification of the kinds of evidence is to be guided by “developing a set of propositions that support the proposed interpretations” (p. 9).

Our examination of the TOEFL validity argument demonstrates the need for more explicit guidance on how to formulate an “intended interpretation” and the “propositions” that are supposed to point to the “types of evidence” that constitute the validity argument. We will explain how the “interpretive argument” as defined by Kane (1992; 2001) and Kane, Crooks, & Cohen (1999) helps to surmount the challenges presented by the less precise guidelines in the Standards. The “interpretive argument” includes an explicit statement of the inferences and their associated assumptions that underlie score interpretation. We show how the requirement for explicit statement of inferences provides a means for identifying and organizing the types of validity evidence that are relevant to a particular test score interpretation. We will discuss these issues through presentation of the interpretive argument for the new TOEFL, which is comprised of six inferences: conceptualization, evaluation, generalization, extrapolation, explanation, and decision making. We will illustrate how this framework forms the basis of the TOEFL validity argument, which is used to interpret TOEFL research results and identify areas where additional research is needed.
The challenges of the Ontario Secondary School Literacy Test for ESL/ELD students

Living Cheng, Queen's University

Session 4, 26th Fri., 8:30 - 10:00

English as a Second Language (ESL) and English Literacy Development (ELD) students have greatly increased in Ontario secondary schools in recent years. Although immigration to Ontario increased 23% in one year (2001-2002), the number of ESL teachers and programs in Ontario schools has actually declined by 30% in the past five years. In addition, ESL/ELD students were failing the Ontario Secondary School Literacy Test (OSSLT) at a rate higher than any other group of students. The OSSLT, which consists of a reading and a writing component, is a useful quality assurance measure that shows the extent to which Ontario students are meeting a common, basic standard for literacy across the province (EQAO, 2002, p.1), and a pass in this test is required of all students for secondary school graduation in Ontario. In the reading component, students are asked to read a total of 12 short selections in three different text types: information 50% of the component; graphic 25%; and narrative 25%. There are three test formats: multiple-choice (MC) (40 questions), constructed response (CR) (35 questions), and constructed response with explanation (CRE) (25 questions). In the writing component, students are asked to complete four writing tasks: a summary, a series of paragraphs expressing an opinion, a news report, and an information paragraph. Students need to pass both components of the OSSLT. Results of the OSSLT (EQAO, 2002) show that among 1735 ESL/ELD students in the public schools who wrote the test, only 37% of ESL/ELD students successfully passed the OSSLT. Among them, 26% passed the writing only, and 3% passed the reading only. A third of the students, 34%, failed both reading and writing, in spite of the fact that 52% of the total number of ESL/ELD students were allowed to postpone their taking of the test in 2002. This result, therefore, creates a serious concern regarding ESL/ELD students' successful graduation from Ontario secondary schools.

A study was therefore conducted to analyse the February 2002 OSSLT data with the permission of the Educational Quality and Accountability Office (EQAO), a quasi-governmental provincial testing agency. The purpose was to examine closely ESL/ELD students' performance (n=4848) in each aspect of reading and writing on the OSSLT compared with native English-speaking Canadian students (n=103391). Research studies have shown that the test items, the content, types and context of reading passages, and the relationships of these intervening factors can have a significant impact on students' performance (Anderson et al., 1991; Freedly & Kostin, 1993; Kobayashi, 2002, Lee, 2002; Peretz & Shoham, 1996; Perkins, 1992). Research into writing assessment has demonstrated that different types of writing tasks create different challenges for students (e.g., Connor-Linton, 1995a, 1995b, Kobayashi & Rimmert, 1996, Hamp-Lyons, 1996). The results of this study showed the relationship between ESL/ELD students' reading performance on three reading types using three reading skills and four reading strategies tested through MC, CR, and CRE formats. This study also demonstrated the relationship among the ESL/ELD students' writing performance on the four writing tasks. In reading, ESL/ELD students had better performance on the literacy and information than on graphic text types. It was also found that CRE format is the most difficult for ESL/ELD students, although it is the best predictor for both reading and writing tasks. In addition, an examination of overall student performance on the four writing tasks through 30 randomly selected writing samples showed that the news report and the information paragraph were the most challenging. The above results can help us understand why ESL/ELD students failed the OSSLT, and provide valuable feedback to EQAO and help inform ESL/ELD instruction in Ontario.

Modeling development of spoken language ability

John H.A.L. de Jong, Language Testing Services

Session 5, 27th Sat., 10:20 - 11:30

One of the difficulties in teaching the spoken language is a limited understanding of which aspects of language skill are important in achieving efficient communication and also which linguistic variables need special attention at the various stages in the language acquisition process.

This study reports on the relative importance of different linguistic variables (e.g. intelligibility, syntactic awareness, vocabulary range, fluency, pronunciation, and listening comprehension) at a number of different stages in the development of overall spoken language ability. Understanding these relationships may help language teachers address the most relevant aspects of speaking skills depending on the learner's level and native language.

Higgs & Clifford (1982) proposed a relative contribution model (RCM), which posits different relative contributions to spoken language proficiency from structural and channel-control dimensions as a function of overall language ability level. RCM suggests that, at lower levels, learners earn their overall scores mainly from channel control aspects, such as pronunciation and fluency. Lower level learners would typically lose points on grammatical and lexical accuracy or on other structural aspects. RCM suggests that higher level learners gradually reach a status where all aspects contribute more evenly to an overall proficiency. Indeed, Higgs & Clifford suggest that for some learners a barri-
er exists at the 2/2+ level, because these learners seem unable to reach the next required level of structural accuracy. Accordingly, these students continue along communicating fluently but ungrammatically.

The paper reports and analyzes data that test and generally confirm, the RCM hypothesis. During the development of an instrument for assessing spoken Spanish, both general ability (CEF scale) scores and four diagnostic scores (vocabulary, fluency, pronunciation, sentence structure) were assigned to 528 students learning Spanish. For each subject, the average diagnostic scores were divided by the average overall score to obtain their relative contribution. The same procedure was followed for all subjects assigned to levels A1 through C2 on the CEF scale. Simply plotting and inspecting the data suggests that RCM describes a strong pattern in spoken language learning. Statistical analysis of the data confirms that the channel control aspects, pronunciation and fluency, give higher relative contribution to the overall CEF scores than the linguistic aspects for subjects assigned to levels from A1 through B1, with the pronunciation score providing the highest contribution and the vocabulary the lowest. Towards the higher levels of overall spoken language ability, the relative contributions of the four diagnostic scores come closer together, but at the highest levels, C1 and C2 the position of the channel control and linguistic aspects is reversed. The findings were confirmed when sub-samples of the learners were again sorted, but this time according to concurrent certified OPI tests reported on ILR and ACTFL scales.

Using the ELD Assessment Portfolio: An investigation of teachers' interpretations of ELD standards and students' language ability

Lorena Llosa, University of California, Los Angeles

Session 6, 27th Sat., 3:45 - 5:15

The use of standards-based classroom assessments to test English learners' language proficiency is becoming increasingly prevalent in many countries. Yet many questions remain as to the reliability and validity of the use of classroom assessments within a high-stakes accountability system (Brindley, 1998, 2001; Rea-Dickins & Gardner, 2000). Among the main concerns are: 1) variability in teachers' interpretation of language ability; and 2) the use of formative assessments for summative reporting. The present study investigates these issues in the U.S. context by examining teachers' use of the English Language Development (ELD) Assessment Portfolio, a standards-based classroom assessment of English proficiency used in a large urban school district in California.

The ELD Assessment Portfolio documents student progress towards each of the California ELD standards. It is intended to be used for formative purposes, but it is also used to make high-stakes decisions about English learners' progress from one ELD level to the next, and serves as one of the criteria for redesignation as Fluent English Proficient. Each classroom teacher has to determine the extent to which students have made progress toward each of the standards and must include in the portfolio samples of student work that demonstrate student mastery of the standards.

Ten elementary school teachers were interviewed about their use of the ELD Assessment Portfolio and were asked to engage in a think-aloud as they scored their own students' portfolios. These concurrent verbal protocols were audio-recorded and later transcribed. This paper will report on the processes that teachers were engaged in while scoring the assessment portfolio; particularly, the extent to which teachers interpreted the ELD standards consistently and the kinds of evidence they used to determine student mastery or progress towards mastering those standards. Also, teachers' perceptions and use of this assessment as formative versus summative will be presented. The findings from this qualitative study will be discussed in terms of their implications for the use of standards-based classroom assessments within a high-stakes accountability system.

Linking the results of teacher-developed tests and action research findings to curriculum development

Lucilla Lopriore, University of Cassino
Guido Benvenuto, University of Rome

Session 1, 26th Fri., 10:50 - 12:20

In language testing the issue of content validity has been debated for many years. This paper describes the results of language tests which were developed by teachers themselves and administered at local and national level, and how those results -- together with the findings of a two-year large-scale action research project -- have provided the support for curriculum renewal.

The language tests were only a part of the research work carried out by 70 teacher trainers, over 700 teachers, and more than 8000 students over almost two years in Italian foreign language classrooms at all school levels within the framework of the Italian National Teacher Research Project which was based upon Action Research. The Project which was developed and organised jointly by the Italian Ministry of Education (MIUR) and five Italian professional associations of foreign language teachers and co-ordinated by six experts was carried out between 2001 and 2002. Its main aims were to

- foster the development of a national foreign language curriculum for all school levels resulting from:
- the adjustment of the descriptors of the Common European Framework to the particular needs of the Italian school system,
- the findings of teachers' inquiry into their own classroom practice through a model of collaborative research,
- the results of language tests administered at each of the main school exit levels (primary, middle, and two levels of high school)
  • provide the basis for an active and long-term collaboration between the Italian Ministry of Education and the Italian professional associations of foreign language teachers
  • develop, and establish within the school system, new forms of collaborative inquiry among teachers.

The Project had two main phases of implementation:
  • the actual research, entailing the investigation of each individual teacher's real pedagogical approaches through the use of specific tools (such as questionnaires, worksheets, observation charts) designed by experts at national level;
  • the main study, entailing the administration and marking of a set of tests specifically designed by a group of teacher-trainers, who also work as teachers at school, under the supervision of experts. Such tests were calibrated according to the CEF levels A2, B1 and B2, and aimed respectively at students in their final year of primary, middle, second and fifth year of upper secondary school.

This specific research framework allowed the collection of data related to the pedagogical approaches of a quantitatively significant number of classes and teachers throughout the Italian territory. Such data, gathered and immediately examined by the teachers themselves on a qualitative basis, could subsequently be analysed not only to review the processes of teaching and learning taking place in various Italian contexts, but also to measure the actual proficiency of students in relation to the Common European Framework levels.

A comparison of still photo and video-mediated web-based listening tests

Gary J. Ockey, University of California, Los Angeles

Session 6, 27th Sat., 3:45 - 5:15

As web-based assessments evolve into the standard mode of delivery in high stakes testing situations, unanswered questions have emerged regarding the role of visuals when presenting texts for listening comprehension tasks. While researchers (e.g., Ginther, 2002; Gruba, 1999; Kasten, 1995; and Chung, 1994) have convincingly argued for the importance of visual stimuli in listening tasks, little research has been conducted to suggest how different types of visuals might affect the way a text is processed by test-takers. The aim of this exploratory study was to compare two commonly utilized forms of visual stimuli in web-based listening tests, namely video and a series of still photos.

Participants in the study were six ESL students who were attending a large university in the United States. Two parallel web-based tests of listening comprehension were utilized, one video-mediated and the other with a series of five still photos of the lecturer. The study employed a retrospective verbal protocol procedure during the text input phase; the input source was repeatedly paused and participants were asked to provide immediate retrospective reports on the segment that they had just heard. After completing each test, the participants were asked to compare the two test formats. The extent to which participants engaged with the different types of visuals along with possible construct implications will be discussed.

Portfolio assessment in practice

Nishita Prasad, St. Francis’ Canossian College
Jo Lewkowicz, English Centre, University of Hong Kong

Session 4, 27th Sat., 8:30 - 10:00

Hong Kong is undergoing considerable change in its education system. At the high school level it is moving towards school-based assessment, and in terms of language education, task-based assessment is being promoted. Among the practices advocated by the Education and Manpower Bureau is portfolio assessment. Although this form of assessment has received considerable attention within contexts such as the European Union, to date such assessment has been under-utilized and under-researched in Hong Kong. A few studies have been reported on portfolio assessment here, but generally these have been at the tertiary level (e.g. Bryant & Timmins, 2002; Hamp-Lyons, 2002).

This paper focuses on a study conducted at senior high school level. Its primary aims were to determine whether school students within the Hong Kong (Confucian) context could take greater responsibility for their own learning and benefit from compiling a language portfolio. The first step in the study was to negotiate with the group of students the content of their portfolio. This stage, which was crucial to the success of the project, took a considerable amount of class time. In the second stage, as the students went about compiling their portfolio over the course of the school year, they were monitored by their English teacher. Feedback was solicited from them on how they perceived their portfolio journey. Two focus groups were identified, one made up of linguistically more proficient students and the other of less proficient students. The focus groups were interviewed at various stages to provide a more in-depth understanding of the students’ perceptions of the portfolio process. At the end of the school year, all
the students completed a self-reflection in which they assessed their work, commented on their own achievements, and on their perceptions of the whole process. These self-reflections were analyzed and compared with the teacher's perceptions of the students' work and progress.

The paper will report on the findings of this study which suggest that through the process of negotiating and compiling the portfolio, the students experienced enhanced motivation in terms of the effort they made to improve their English. Most importantly, they suggest that even the weaker students saw the benefit of the portfolio and improved as a result of going through the process of negotiating its content and compiling the portfolio. The paper will end with suggestions for making portfolio assessment manageable should it be introduced on a larger scale.

The lexical component of speaking test performance

John Read, Victoria University of Wellington

Session 2, 26th Fri., 1:50 - 3:20

An aspect of speaking test performance which is often taken for granted is the adequacy of the learners' vocabulary knowledge to express the meanings and functions necessary to perform the test tasks. Conventionally in lexical studies productive vocabulary knowledge is defined in terms of the range of either words or idiomatic expressions that the learner can use in speech and writing. However, current work on formulaic language is highlighting the extent to which lexicalized multi-word units of various kinds play a role in fluent, native-like performance of production tasks.

This paper is based on analyses of a small corpus comprising transcripts of 100 candidates taking the Speaking module of the International English Language Testing System (IELTS). The speaking test, which takes 11-14 minutes, consists of three main parts: an interview and a prepared talk, which are carefully structured, followed by a more open and probing discussion between the examiner and candidate. The candidate's performance is rated according to four analytic criteria, one of which is Lexical Resource. The sample tests were taken from a large pool of tapes from test centres worldwide, with the main selection criteria being the topic of Parts 2 and 3 of the test and the candidate's final speaking score.

The analyses include the calculation of lexical statistics to investigate the frequency, variation and sophistication of words used in the whole corpus and in the speech of individual candidates. However, the main focus of this presentation is a more qualitative investigation of the lexical characteristics which distinguish performance at Bands 8, 6 and 4 on the nine-band IELTS proficiency scale. Particular attention is paid to criteria for identifying formulaic sequences in the candidates' speech and how these features may influence both the rating for Lexical Resource and the final speaking test score. Apart from contributing to the validation of the IELTS Speaking module itself, the research is intended to develop our understanding of the relatively neglected area of oral vocabulary use in a testing context.

Investigating the relative significance of vocabulary breadth and syntactic knowledge in the prediction of reading test performance: A structural equation modeling approach

Toshikiko Shiotzu, Kurume University

Session 2, 26th Fri., 1:50 - 3:20

Sources of difficulties in L2 reading and predictors of individual differences in L2 reading test performance have constituted a major area of research (Alderson, 1984), and the available evidence generally supports the view that one's "L2 linguistic knowledge" (or "L2 proficiency"), as opposed to one's L1 literacy, accounts for the individual differences in L2 reading test performance better (Bernhardt, 1999). In studies that investigated this issue, "L2 linguistic knowledge" was operationalized through several measures including tests of vocabulary and grammatical knowledge, each of which has been empirically shown to be linked to reading test results (e.g., Laufer, 1992; Alderson, 1993, respectively). Their relative significance in L2 reading, however, deserves more focused research. The study to be reported aimed to explore the degree to which L2 vocabulary and L2 syntactic knowledge account for the individual differences in L2 reading test performance.

A total of 624 EFL students took tests of reading comprehension, vocabulary breadth, and syntactic knowledge, and the data were submitted to structural equation modeling for analysis. The fit data indicated that a model which hypothesizes intercorrelated vocabulary and syntax factors both accounting for the reading comprehension factor is relatively more acceptable than the other model variants. Through a model comparison, it was also confirmed that syntax had a stronger link to reading than did vocabulary. A test of equivalence across sub-groups indicated that the higher achievers and the lower achievers exhibit slightly different patterns of abilities although the prevalence of syntactic knowledge over vocabulary breadth was clearly observed regardless of the group division. The implications of the findings will be discussed and some recommendations for future research suggested in the presentation.
English language assessment and its relative importance in school accountability in California

Simeon Slovacek, Laura Pantoja, & Wing Chi Lee, California State University, Los Angeles

Session 1, 26th Fri., 10:50 - 12:20

Abstract: The purpose of this research was to examine the relative importance of English Language Testing in California’s K-12 public schools. There are two recent legislative mandates that have shaped school accountability in this state. The first is the California Public Schools System of Accountability Act of 1999 (PSSA) and the second is the Federal “No Child Left Behind Act”. The California legislation spells out an index called the Academic Performance Index or API, while the Federal legislation spells out the conditions for determining Annual Yearly Progress or AYP. This research analyzes the relative contribution English language testing and assessment plays in determining school effectiveness.

From the charts below we see that language assessment accounts for 60% of the components or weighting in the Academic Performance Index. We conducted a small study of elementary teacher instructional time spent on each subject and compared that with the API’s emphasis. Surprisingly, while the state of California weights language assessment (through multiple instruments) 60% of the API index, teachers in our study and other studies spent much less time on teaching language arts (39%) and more time on teaching other subjects. In fact, the API for elementary schools only measures language arts and math yet the state standards require elementary teachers to teach all of the core subjects including science, social studies, and art in addition to English and math. The Federal No Child Left Behind definition of “Adequate Yearly Progress” is a bit more complicated but again focuses on students’ English and math performance.

Apparently Language testing is more important in school accountability than any other subject (or all other subjects combined!) This raises some interesting issues for the California system of school accountability, which we explore in this paper. Such issues include the mismatch between the state grade level subject matter content standards on one hand (that which teachers are supposed to teach) and what teachers and schools are held accountable for on the other hand. Issues of test use for accountability and fairness are addressed as well.

Revising assessment criteria and rating scales for assessing L2 writing: The contribution of quantitative analyses in the validation phase

Lynda Taylor, Stuart Shaw, & Graeme Bridges, University of Cambridge ESOL Examinations

Session 4, 27th Sat., 8:30 - 10:00

Routine monitoring of operational test performance together with ongoing validation studies sometimes lead to a decision to revise certain features of the test in question, e.g. test format, task design, assessment criteria or rating scales; revision of any of these features requires that appropriate validation studies are conducted before the revised test is implemented in the live operational context.

This paper reports on validation studies conducted following revision of the writing assessment criteria and rating scales of a large-scale, international, high-stakes test of English language proficiency. The presentation will briefly outline the nature of the changes made to the existing writing criteria and scales. It will then go on to describe the quantitative analyses carried out during the validation phase of the revision project, and in particular the use of Multi-Faceted Rasch and Generalizability Theory to answer a number of different questions relating to the measurement qualities of the revised approach to writing assessment.

The research involved a multiple rating exercise: 12 raters - identified as representative of the ‘universe’ or worldwide population of raters for this test - rated a total of 58 writing performances. All raters in the study were experienced raters of this particular writing test and had been retrained and standardised to use the revised assessment criteria and rating scales. The 58 writing performances were identified as benchmark scripts representing the full proficiency continuum for the test. The trial rating activity was staged over a period of 3 weeks to allow for 4 marking phases, and raters received feedback following each phase.

Observations resulting from the multiple rating exercise were analysed in a variety of ways: inter-rater reliability coefficients, ANOVA and Generalizability Theory were used to estimate examiner reliability and to investigate scale functioning; correlational and partial credit analyses were also used to explore between-rater relationships and the integrity of the scales; finally, Multi-Faceted Rasch was used to explore issues of construct validity and aspects of rater behaviour (rater consistency and severity, misfitting responses). Findings from all these analyses will be presented and discussed.

Results of the quantitative analyses provide evidence of the validity and reliability of the revised assessment criteria and rating scales; supplementary findings from complementary qualitative studies carried out during the validation phase point to the practicality and posi-
The prescriptive impact of the criteria and scales from the raters' perspective. The presentation will discuss how information from both quantitative and qualitative analyses have fed directly into the rater retraining and standardisation programme in preparation for introduction of the revised writing assessment criteria and scales for the test from the beginning of 2005.

What does it take to get a high score on a paired task? Examining discourse, scores and the construct of L2 proficiency

Carolyn E. Turner, McGill University
Joanna White, Concordia University

Session 5, 27th Sat., 10:20 - 11:50

The communicative movement in L2 education has brought with it several challenges in terms of assessment, score inference and generalization. Performance assessment has evolved to a place of importance and with it has come the interest in evaluating oral language through tasks that require extended discourse (McNamara et al., 2000; Turner & Upshur, 2002). In educational contexts, this includes a growing use of paired and group oral tasks. Educators are presently faced with the dilemma of determining criteria for scoring. One of the main reasons for this is that the discourse generated in such tasks is increasingly viewed as being co-constructed by participants in a specific local situation (Swain, 2001; Young, 2000). As Chalhoub-Deville (2003) discusses, traditionally, the field of language testing has focused on the construct of language proficiency by examining individual ability (i.e., the cognitive perspective). The characteristics of paired/group tasks lead us to the examination of a joint performance from the social interactional perspective. So where does this leave us in terms of interpreting individual scores as is the need in L2 educational contexts? What does a score on a paired task tell us? These questions contribute to an already complex situation in terms of defining the construct of speaking in oral tasks and are addressed in this presentation.

This paper will describe a study that examined the co-constructed performance of paired testees in a timed information gap task. The participants were 74 grade six students (age 11-12) in a communicatively-oriented intensive ESL program. The data come from a recent study involving three elementary schools in Quebec where we investigated the gains in oral proficiency across two ESL programs, regular and intensive. The paired task performances were audio taped and transcribed.

Through a discourse analysis of the joint performances, variation was observed in pair interaction, but in general a pattern emerged demonstrating a "game" strategy. Salient language requirements for task "success" were identified. Although speed and strategies used in each pair appeared to vary according to proficiency levels of the participants, this did not always have the relation to scores as predicted. These results will be discussed within the framework above.

Paired tasks require examination of jointly constructed discourse. This appears to be an "instance" of language performance that needs attention in current assessment contexts. The results of this study indicate that learning more about such "instances" can help us better understand the L2 construct as it is manifested in paired interactional discourse for evaluation purposes. This may help address the issue concerning the "tension between the needs for accuracy in assessing a particular performance and generalization to broader domains of language use." (Upshur & Turner, 1999, p. 107).

Measuring the knowledge of text structure in academic English as a Second Language (ESL)

Viphavee Vongpumivitth, University of California, Los Angeles

Session 2, 26th Fri., 1:50 - 3:20

Much research in both first and second language reading shows that knowledge of text structure, i.e., the ability to locate main ideas and supporting details and understand the hierarchy of ideas in the text, is an important component of academic reading ability. However, the nature of this knowledge, especially as it is operationalized as a construct in an ESL academic reading test, is not yet clear.

The current study is an attempt to investigate how the knowledge of text structure can be assessed through the use of performance-based test tasks. Four performance-based tasks have been developed based on suggestions from reading research and textbooks on second language reading. The four tasks are: (1) a set of open-ended questions, (2) an incomplete outline, (3) a graphic organizer, and (4) a summary. Qualitative and quantitative analyses were performed with data collected from approximately 145 undergraduate and graduate English as a Second Language (ESL) students at a large public university in the US.

The focus of this presentation will be on the relationship between the test takers' knowledge of text structure as measured by the four test tasks and their academic reading and writing ability as measured by other ESL tests. First, multivariate generalizability analysis indicates whether the knowledge of text structure, as measured by the four test tasks, is a single construct, i.e., the ability to understand hierarchy of ideas in the text, or two constructs, i.e., the ability to locate main ideas and the ability to locate supporting details. Multivariate generalizability analysis also reveals the extent to which test takers' performance vary among...
The effects of visual chunks and planning time on the language quality of verbal descriptions of graphs

Xiaoming Xi, Educational Testing Service

Session 3, 26th Fri., 3:40 - 5:30

This study examined the extent to which task characteristics (the number of visual chunks and the amount of planning time) and test taker characteristics (graph familiarity) influenced the perceptual and cognitive processes involved in graph comprehension, the strategies used in describing graphs, and the scores obtained on the graph description task in the Speaking Proficiency English Assessment Kit (SPEAK), a semi-direct oral test. Specifically, this study investigated whether providing planning time and reducing the number of visual chunks in bar and line graphs influenced the perceptual and cognitive processes involved in graph comprehension and helped mitigate the influence of graph familiarity on language performance on an oral graph description task. These research questions were explored using a structural equation modeling approach in a framework of the influence of the interaction between task characteristics and test taker characteristics on task performance.

It was found that the participants' graph familiarity affected the overall communicative quality and some specific components of their performance on these graph description tasks. Test takers' graph familiarity thus represented a potential source of construct-irrelevant variance. However, reducing the number of visual chunks in a graph and providing planning time positively impacted the perceptual and cognitive processes involved in graph comprehension and helped mitigate the influence of graph familiarity on performance on the graph description task. This could be seen from improved performance in two ways: under the planning time condition, both holistic and analytic scores were higher, and responses used more sophisticated organization patterns and descriptive strategies to integrate the information in the graphs. With the few-chunks line graph tasks, holistic scores were higher.

Theoretical implications of this study will be discussed in terms of the effect of planning on oral output and the effect of visual patterns and planning on perceptual and cognitive processes in graph comprehension. Practical implications for designing graph description tasks, for tightening up the task specifications, and for improving rater training will also be discussed.
Technical innovations in the assessment of oral reading fluency in 4th graders

Isabella Barbier, Ordinate Corporation

As part of the 2002 NAEP reading assessment, an Oral Reading Study was conducted in order to assess the oral reading accuracy and fluency of a sample of approximately 2000 fourth-grade students. Each selected student read aloud a grade-appropriate story.

Technological innovations of the assessment included a computer-mediated administration mode, a web-based rating system which allowed raters to listen to responses online and enter their fluency ratings and accuracy miscue codings using custom-designed web-pages and also included automatic measures related to reading rate and fluency.

The computer-mediated administration mode involved the use of laptop computers. The laptop computers ensured that the data-collection procedure was highly standardized: administrators followed directions printed on the screen. Students followed prerecorded instructions heard through the earphones of their headset and read the stories aloud into a microphone so that their oral readings were captured on the laptop.

Each oral reading was scored for both accuracy and fluency by human raters. A specially designed web-based interface allowed fluency raters to listen to randomly presented responses online and enter their fluency score. Accuracy raters had access to custom-designed web-pages which allowed them to code miscues by selecting appropriate miscue codes from drop down boxes. The web-interface also allowed master raters to monitor the performance of other raters with regard to intra-rater and inter-rater reliability.

In addition to the human ratings, several automatic measures related to reading rate and fluency were calculated for each oral reading. These included narrow time (passage reading time without leading and trailing signal), initial words per minutes (the number of words read in the first 60 seconds of speech after onset of passage reading) and run rate (average run rate for the passage read, defined as the average number of words read continuously between adjacent inter-word silences).

Assessment of the initial stages of linguistic competence among children of immigrant families attending public schools in Italy

Monica Barni & Sabrina Machetti, Università per Stranieri di Siena

The objective of this paper is to present the results of a series of experiments carried out in various schools and which sought to outline and assess the initial stages of the development of linguistic competence among children of immigrant families attending public schools in Italy. Over the past ten years or so the presence of immigrant children in the Italian school system has grown exponentially, whether such children are themselves immigrants to Italy or whether they are born in Italy to immigrant parents (Caritas di Roma, 2003). Italian schools, therefore, are faced with daily challenges regarding the linguistic, socio-cultural and individual needs of these children. Such needs clearly include the development of competence in Italian and, as a consequence, needs for learning contexts which give voice and meaning to the peculiar semiotics of immigrant children (Barni, 2001; Vedovelli, 2001). In turn, this situation has had a profound impact on the teaching of Italian as a second language. Questions have been raised regarding the relevance of traditional teaching methods, methods which presented Italian as L1, not L2, and, ultimately, although having passed through an initial 'emergency phase', didactic tools, methods, assessment systems and general approaches have been found to be ill-adapted to the needs of immigrant learners of Italian.

Aware of the need to address this situation, the Centro CILS - Università per Stranieri di Siena - has been developing a system for the assessment of linguistic competence of immigrant children (Machetti, 2002; Barki et al, 2003) which corresponds to the levels A1 and A2 as outlined by the Council of Europe’s Framework (Council of Europe, 2001). In doing this it has been necessary to redefine the descriptors of competence for these levels and reflect upon the actual functions of assessment for this type of learner and at these levels of competence.

The research and experiments carried out in Italian schools has highlighted the importance of an assessment system which reflects a curriculum design specific to the needs of immigrant children and which necessitates the involvement of teachers of that curriculum at every stage from design to assessment (Brindley 1998, 2002, McNamara 2002, Rea Dickins 2003).
Assessment of outcome in an advanced ESL writing course and its admission policy implications

Doreen Bayliss, University of Ottawa

The impetus for this research came from a change in our University's admission policy on ESL language requirements for ESL speakers which took effect in the fall of 2001 and which had a direct impact on who was admitted to our advanced writing course. For some time the University has included writing ability as one of its admission criteria for those tests which measure writing, primarily CanTEST (4.0 required) at this University. As competition for students has increased, the University, seeking ways to be more attractive to international students, began offering conditional admission with this course being made mandatory for those with writing scores just below the acceptable level. The reasoning behind this policy was based on what theoretically might be expected to change during a 13 week, 3 hours per week ESL writing course which focused on the organization and development of expository prose. The University began offering the same conditional admission to students with passing TOEFL scores and a TWE of 4.0 effective last fall.

In the past most students were admitted on the basis of the results of an in-house placement test, measuring listening and reading comprehension, and a grammar test (a guided productive measure of grammar knowledge). Last fall, however, many students were admitted to the class based on the conditional admission requirements outlined above. It was soon apparent that many of the students admitted were weaker than expected as determined from the grammar test and a writing sample. Thus, the decision was made to track all students (56 in 2 classes) for the semester in order to 1) track progress as it related to organizational ability, language use and use of vocabulary and expressions, 2) determine success both in terms of meeting the goals set for the course and in terms of meeting the minimum writing standard expected by the University (in this case, a CanTEST of 4.0) and 3) try to determine whether there were any entry criteria which could be used to predict success in the course as outlined in 2).

Data gathered during the semester included the following: personal data including L1, field and level of study and incoming ESL admission test scores; marks awarded on a grid evaluating seven aspects of organizational ability, overall organizational ability, language use and use of vocabulary and expressions on an expository writing task given at the beginning of the semester and again about two-thirds of the way through the semester; incoming grammar test mark and writing mark; marks on the three in-class writing assignments and on the in-class final essay (success in the course); the final essay was also marked according to the criteria used on the CanTEST (success in meeting the University's minimum writing standard).

The study has already resulted in a change of admission policy with regard to the writing course admission requirement regardless of ESL test source. Any conditional writing requirements must be based on the results of the in-house grammar test (the best predictor in the study for a final CanTEST of 4 or better) and a writing sample. Results from all data sources will be discussed including what is revealed about the progression in ESL writing achievement, possible admission strategies and useful indicators to help determine a successful outcome.

Development and validation of a spoken Spanish test

Jared Bernstein, Elizabeth Rosenfeld, & Isabella Barbier, Ordinate Corporation

This paper describes the development and validation of a computer-based Spoken Spanish Test (SST). The SST measures fluency, accuracy and lexical range in listening and speaking, by directly exercising real-time performance in core language skills. Items are presented orally in Spanish and are administered and scored automatically. The test takes 15 to 20 minutes and has seven sections: readings, repeats, opposites, short answer questions, sentence builds, open questions, and story retellings.

A pool of more than 700 potential items were reviewed for acceptability by native-speaker linguists from several Latin American countries and from Spain and approved items were recorded using speakers of several different Spanish-speaking countries. The items were then assembled into tests, and presented to more than 500 adult non-native Spanish learners, including US military personnel and students from five universities in the United States, and to more than 400 native Spanish speakers from Argentina, Colombia, Mexico, Puerto Rico, and Venezuela. Automatic scoring algorithms were then developed.

The objective of the validation analyses was to research the quality of SST as an instrument to measure "facility in spoken Spanish", i.e., the ability to track what is said, extract meaning in real time, and formulate and produce relevant, intelligible responses, at a conversational pace. Given this definition one would expect native speakers to obtain high scores on SST. On the other hand, for SST to operate as a measurement instrument, speakers of other languages who are learning Spanish would need to be distributed over a wide range of scores. Although their construct definitions may differ from SST, one would also expect some relation between scores from well-conducted OPs and SST scores.

The data analyses therefore investigated the following hypotheses:
1) Native speakers of Spanish consistently obtain high scores on SST;
2) Learners of Spanish as a second or foreign language are distributed over a wide range of SST scores;
3) SST scores align closely with other measures of spoken Spanish including official OPI tests reported on ILR and ACTFL scales and with judgments of proficiency under the spoken interaction rubrics of the Common European Framework.

The paper presents data confirming these three hypotheses. In addition, the paper presents data showing a remarkable correspondence among different OPI scores for the subject population.

**The Iowa Test of English Language Learning: Vocabulary, writing, and structure & expression**

Micheline Chalhoub-Deville, Craig Deville, Anne Cummings, & Shelley Fairbairn, *The University of Iowa*

The No Child Left Behind Act of 2001 mandates that English language learners (ELLs) be assessed annually in the areas of listening, speaking, reading, writing, and comprehension. In response to these mandates, researchers at The University of Iowa are developing the Iowa Test of English Language Learning. This battery will be utilized to document growth in the development of K–12 ELL students’ academic language proficiency.

This poster will outline the development of the vocabulary, writing, and structure & expression portions of this assessment battery for the following grade spans: 3–5, 6–8, and 9–12. The vocabulary section consists of everyday, academic, and content-area items, the writing section includes two academic writing tasks, and the structure & expression section addresses grammatical structures and mechanics via multiple choice items.

The vocabulary section was crafted to measure both basic interpersonal skills (BICS) and cognitive academic language proficiency (CALP). Vocabulary lists were compiled based on academic and high frequency word lists, examination of language in standardized tests, and conversations with content experts and ESL teachers. Additionally, social studies, science, and mathematics textbooks were consulted. Everyday, content, and academic words were assessed via three item types: picture-based, context-dependent, and matching.

The writing section was developed following a thorough sampling of state writing standards and a broad survey of the literature addressing both L1 and L2 writing. Based on the findings of this research and the time constraints in the K–12 school context, students in each grade span will complete two tasks representing genres commonly addressed in each specific grade span.

The structure & expression section of the test was constructed after a careful examination of state language arts standards, national and state ESL standards, and ESL textbooks. In addition, literature discussing both order of acquisition and the mismatch of language addressed by common English language proficiency tests and that included in standardized achievement tests was reviewed. All of this information was synthesized in the development of test items based on structures documented in the literature which are found in beginning, intermediate, and advanced texts, as well as various mechanical features. Where possible, item distractors were also based on findings of order of acquisition studies, enabling items to yield as much information as possible regarding a student’s grammatical development.

**Academic language-based assessment: A listening test**

Micheline Chalhoub-Deville & Minhee Eom, *University of Iowa*

Currently, ELL students are required to take standardized content area tests regardless of their English proficiency. In the case of ELL students with a low level of English proficiency, it is difficult to clarify if their poor performance is due to their poor knowledge of the content area or due to their insufficient English abilities. Several studies (August, D. and Hakuta, K., 1994; Bailey, A. L., 2000; Chalhoub-Deville, 2001) have suggested that it is necessary to develop a guideline to determine whether the ELL students are ready to take a standardized content assessment with their native peers. Chalhoub-Deville has led the development of ELL Academic Language Assessment for the following grade spans: 3–5, 6–8 and 9–12. This proficiency assessment battery consists of several modalities including listening.

In the development of the listening test, the theoretical constructs are identified based on the communicative language ability model (Bachman, 1990; Bachman and Palmer, 1996) and the framework of listening ability (Buck, 2001). In order to identify the operational model or assessment framework, an analysis has been conducted of the state standards of six states: California, New Jersey, Florida, Arizona, Utah and Texas. In addition to the analysis of the state listening standards, two sets of language arts textbooks for each target grade have been analyzed. Consequently, the development of the operational construct for the listening module has been grounded in the analyses of three main sources: theoretical models, state standards and textbooks. Based on the operational model, six types of task statements have been developed. The curriculum content in the target level grades has been taken into account in
order to increase authenticity of the texts considered for the test. Thus, the texts in the listening test are designed to look similar to those used in academic discourse in terms of type and content.

Although the test has been well grounded in theories and real world learning activities, additional perspectives are desirable in elaborating the construct of K-12 L2 academic listening. Presenting at LTRC would allow the test development team to obtain valuable input from other language testers regarding various aspects of test development.

The design, development, and validation of a placement test for the Community English Program at Teachers College, Columbia University

Hyunjoo Kim, Jee Wha Kim, Alick Liao, Eun Kyeong Park, Taejoon Park, Michael Perrone, James Purpura, Mary Regan, Yoko Saito, Elvis Wagner, & Cynthia Wiseman, Teachers College, Columbia University

The Community English Program (CEP) at Teachers College, Columbia University is an adult language program administered by the TESOL program at Teachers College. The students in the CEP are Columbia staff and members of the community. Courses are not credit-bearing. All of the teachers in the CEP are also students in the TESOL and Applied Linguistics M.A. programs at Teachers College.

The classes in the CEP are divided into six different language ability levels, and students are placed in a class based on the results of a placement test. Recently, the administrators and teachers in the CEP concluded that the test that was used for placing the students into the six different levels was inadequate. For this reason, it was decided to create a new placement test.

This poster session describes the design, development, and validation of this new placement test. The test was designed and developed by a committee of TESOL and Applied Linguistics graduate students. The design and development process took approximately six months, and the first version of the test was piloted in the summer of 2001 and administered in September of 2001.

The test is theme-based, with "Competition and Cooperation" chosen as the subject. It is composed of five different sections: listening, grammar, reading, writing, and speaking. The listening section contains two video texts (a lecturette text and a two-person dialogue text) with 12 multiple-choice comprehension items. The grammar section consists of three separate tasks, with 40 multiple-choice items. The reading section has three reading texts, with 15 multiple-choice comprehension items. The writing section includes one task, writing an essay on an assigned topic. These four sections of the test are administered on the same day. The speaking section of the test is administered two days later, after the other four sections have been scored. The current CEP speaking test uses a paired interview format. However, due to a number of limitations, the speaking test is being revised. The new speaking test will be devised based on the evidence-centered design principle, and delivered semi-directly via computer. It will consist of 10 items and last about 20 minutes.

This poster session will present the process of designing the test specifications and the development of the test. It will also present the piloting and revision process. Finally, the initial validation studies performed by the test development committee will be presented, including reliability studies and factor analyses of the different components of the test.

Systematic placement testing and its validity for university English language programs

Tomoko Fujita, Tokai University

The purpose of the study is to demonstrate (a) how a university placement test was revised in order to more appropriately measure the proficiency of incoming students, and (b) how the results are used more effectively for making placement decisions. The presenter focuses on the systematic process of revising the old placement test and investigates relevant evidence that comes into play in test validation. The revision procedure was based on ideas expressed in Messick's "progressive matrix for defining facets of validity" and Chapell's "current approaches to validity evidence". Since validity is a unitary but multifaceted concept, evidence and consequences derived from each stage of the placement testing procedure are used to justify the validity of testing outcomes.

The original 100-item multiple-choice placement test was written ten years ago and had been used to place about 6,500 incoming students annually into basic, intermediate, and advanced levels. The test has three subtests: listening, grammar, and reading. The test results were first analyzed in 2002 using a three-parameter IRT model. The information about individual item properties, and the test information curve (TIC) and test characteristics curve (TCC) were estimated with XCalibre(tm) 1.10 (Assessment Systems, 1995). A qualitative analysis examined for the consequential validity of test use.

Results indicated moderate test reliability (K-R21 = 0.78), and an average test information of 8.75. About 45 test items that showed exceptionally high item difficulties of 3.00 or greater needed to be eliminated. Moreover, the test information around the cut-off point score between intermediate and advanced was 25.0 while that between the basic and intermediate levels was only 7.0. The statistical analyses indicated that
the test was too difficult for incoming students so the participants' proficiency level was not being accurately measured in a number of cases, especially those in the basic and lower intermediate classes. The results of a questionnaire about the placement decisions indicated that students' attitudes at the lower proficiency level were less positive than students at the higher level.

A number of changes were implemented in order to improve the validity of test use. For example, items on the old test that were too difficult were removed. A pilot test was conducted and less difficult items were identified that help increase the reliability of scores around the lower cut-off point. In 2003, both quantitative and qualitative analyses demonstrated the improved validity of the revised test in terms of test use and consequences.

Developing a user-friendly item bank for in-house placement and achievement/proficiency tests of university EFL learners

Akiyo Hirai, University of Tsukuba

We have been developing an in-house assessment system for the past several years, administering English language placement and year-end achievement/proficiency tests a total of five times a year to about 2000 university freshmen and some repeaters. All the tests consist of items calibrated using Item Response Theory (IRT) and stored in an item bank being developed. The achievement/proficiency test is particularly important because, without passing it, students below a cutoff point on the test cannot receive English credits completed during that year. In lieu of the demand for many precise test forms each year and the necessity to administer a test in one day to about 2000 students while lacking sufficient budget to hire item writers, all the tests are paper-and-pencil tests and the test sheets are carefully collected after being administered. Some items are repeatedly used on different test forms. Under these circumstances, item management using an item bank is an important part of our test development. In withdrawing items from the item bank to create a new test form, not only the difficulty of the item but also its content and frequency of use should be carefully considered. In our experience, the procedure of selecting and editing items for a new test form should be as easy as possible because the different people may use the item bank. However, there is little information about how to establish an easy-to-use item bank for a university with minimal investment. Thus, a user-friendly item bank is being created as part of the in-house test assessment project. Using Microsoft(r) Access, the procedure of selecting, un-selecting, and re-selecting items from the item bank has been simplified. In the poster session, I will briefly explain the in-house assessment system and then demonstrate the item selection procedure from the item bank.

Empirical Bayes subscore augmentation: A micro and macro look

Paul Jaquith, United Arab Emirates University

That tests may be required to serve multiple purposes is nothing new to the measurement community. Yet with the rapid implementation of standards-based assessment and the passing of accountability legislation across the US, institutions are under increasing pressure to make as much use of the scores on standardized tests as possible. At the same time, advances in technology make it dangerously easy to break a test up and report subscores. This usually makes measurement specialists quite uncomfortable. Tests that were designed to cover a broad range of material, as with many achievement tests and most placement tests, are generally poorly suited to provide scores that are narrow enough and at the same time reliable enough to be of diagnostic value.

Here in the UAE, Princeton Review provides diagnostic feedback to their clients on a mock TOEFL that includes subscale scores on a broad range of sections, some of which may be composed of as few as 3 items. Clearly, a 3-item multiple choice test cannot have much reliability, and any adjustments to a syllabus based on such feedback might be in danger of simply "chasing errors of measurement." In my own area, my colleagues and I have been under tremendous pressure to report subscores on the national English placement test but have so far resisted because of the modest reliabilities of the subsections and the lack of clarity on how the scores will be used.

The problem: How can we obtain scores of sufficient reliability to be useful in providing both diagnostic feedback to teachers and learners, and at the same time assign achievement scores for the purpose of determining summative grades for a course. Two solutions have been proposed. One is computer-based adaptive testing. (A dramatically successful application of this can be found in the ALEKS system for mathematics.) The other is augmented scoring, used to stabilize potentially diagnostic subscores by augmenting subscore data with information obtained from other portions of the test.

The concept of augmented scores has been around for a long time, but has only recently been linked with advances in item response theory. What I propose to explore in my poster session is the application of an empirical Bayesian augmentation procedure to increase the reliability and usefulness of subscores. I will illustrate what an application of this technology might look like at a micro level (as with the minute reporting practiced by Princeton Review) and at a macro level (as with the three large subtests in our national English placement test). I will attempt to grapple with the issues of reliability and validity of the scores, and with some of the thornier ethical issues.
The role of cloze in a model of foreign language proficiency

Jeff S. Johnson & Amy D. Yamashiro, University of Michigan
Jing Yu, Michigan State University

This poster session investigates the question of whether to consider cloze items as an independent construct within a high-stakes EFL certificate examination. This examination includes speaking, writing, listening comprehension, cloze, and grammar/vocabulary/reading sections. By conducting analyses at the item and subsection levels for the multiple-choice sections, the presenters examine the role of the cloze section items in relation to the other subsections. The data for this study come from three years of test administrations in Europe and South America. The participants in this study included all the candidates for the certificate from 2001 through 2003, with group sizes of 19,904, 22,333, and 30,643. This study focuses on the following research questions:

1. Do the cloze passage items measure a unique trait strongly enough to be considered a separate, scorable section of the test, or is the trait measured with these items similar enough to the trait measured in the grammar/vocabulary/reading section to be included with these items?
2. How do these results compare over the three years of test administrations?

To address the first research question, we analyzed our data using structural equation modeling (SEM) to test two hypothesized models: 1) cloze as an independent factor within a language proficiency model; and 2) cloze as part of a combined factor with grammar, vocabulary, and reading items in a language proficiency model. We then compared results over the three years of data.

Tables and charts present relevant descriptive statistics. Fit indices and the structural models with factor loadings will be displayed to answer the research questions. The researchers will also discuss the implications of this study for future research on using cloze within models of language proficiency.

WebLAS Japanese Project

Shoichi Gregory Kamei, Asako Hayashi, Yuko Haga, & Yasuhiro Imao, University of California, Los Angeles

WebLAS-J is the acronym for the Web-based Language Assessment System in Japanese. A joint project of the Department of Applied Linguistics & TESL, the Department of East Asian Languages & Cultures, and the Center for Digital Humanities at UCLA, WebLAS-J features web-database interactivity and consists of modules for task creation, course management, test form creation, automatic scoring, and score reporting. The goal of the WebLAS application in general and the WebLAS-J project in particular is to provide language instructors, testers, and researchers with a flexible, consistent, and easy-to-use platform for creating, delivering, and managing language instructional and test tasks that are more authentic and interactive. These tasks should also serve as valid measures of language ability that provides diagnostic feedback to learners and instructors and information to guide appropriate placement and grading decisions. The computer technology used in this system also allows researchers to investigate various relations between test taker characteristics and language test performance.

Using WebLAS ver. 2, the Japanese test development team created a set of listening, reading, and grammar tasks for a placement test in basic-level Japanese and conducted four test piloting sessions through summer 2003 in preparation for operational use in fall 2003.

In this presentation, we provide an overview of WebLAS, including the rationale for its development and use and descriptions of its component modules with particular emphasis on the automatic scoring of selected and limited response item and the delivery of multimedia files. Test taker feedback to the system, results of item analyses, and how this information was used to revise test tasks and test forms as well as the opportunities and challenges encountered in using and revising WebLAS will also be presented and discussed.

Validation of speaking performance measures: The case of Japanese learners of English with low speaking ability

Rie Koizumi, University of Tsukuba

Speaking performance can be assessed using the following three main methods: direct assessment of task outcomes, discourse analytic measures, and external ratings (Ellis, 2003). Discourse analytic measures have been substantially utilized in assessing fluency, accuracy, and complexity in second language acquisition (SLA) research (e.g., Skehan & Foster, 1997). A particularly complicated issue is that a number of measures exist that are intended to assess one aspect of performance (e.g., to assess accuracy, error-free T-unit ratio, and errors per T-unit; Wolfe-Quintero, Inagaki, & Kim, 1998), and their degree of validity is not always verified. Although Wolfe-Quintero et al. (1998) summarize a wide range of measures used in previous studies, their target is restricted to writing performance measures. Given that there are differences between writing and speaking performance, it is necessary to explore the validity of speaking performance measures.

The purpose of this research is to examine the validity of speaking performance measures designed to assess fluency, accuracy, and complexity. This is done
through consideration of the following three aspects based on Messick (1989) and Chapelle (1999): (a) internal relationships between the same type of measure, and (b) external relationships between three types of measures and speaking ability scores, and (c) how each measure changes as speaking levels increase.

The participants in this study were 172 Japanese secondary school students with low speaking ability who had studied English for two or three years. They took a face-to-face speaking test and a paper-and-pencil test that included vocabulary and grammar sections. The speaking test was designed to estimate students’ speaking ability levels (the Smattering, Tourist, Breakthrough, Waystage, or Waystage Plus or above levels) on North’s (2000) proficiency scale. After a discussion of the validity of inferences and uses of each speaking performance measure, the results of the previous literature are reexamined. By combining the results of the current study and the reexamination of the previous literature, implications for SLA research and classroom assessment are discussed.

**Complementary role of two performance tests in schools**

Yuji Nakamura, Tokyo Keizai University

1. Purpose of the research

Performance tests that require test takers to produce a written or spoken sample of language can be collectively categorized as a single type of performance assessment. However, results produced by test takers in the two different types of tests (writing and speaking) may vary.

The present research attempts to explore the similarities and differences among students as well as among evaluation items in two kinds of performance tests (Writing and Speaking) by using the Many-Faceted Rasch model.

2. Research design and methods

The subjects were 32 Japanese college students majoring in Economics and Business Administration. Their tasks were 1) to write a composition as a Writing Test and 2) to give, in class, an oral presentation as a Speaking Test, which was tape-recorded for later evaluation. Four raters judged the 32 students’ written manuscripts and oral presentation tapes using a 4-point scale. For the rating, 7 items were used for Writing (Grammar, Vocabulary, Fluency, Content, Discourse, Organization and Overall) and 8 items for Speaking (the same seven items as for Writing, plus Pronunciation).

3 Data analysis

The data were analyzed using the Many-Faceted Rasch Measurement Model. The benchmark for the acceptable range of the infit and outfit statistics data was set between 0.6-1.4, as generally accepted, since this was performance test (both speaking and writing) data that involved raters’ judgments.

4 Results and Implications

In terms of items, grammar is additionally necessary for measuring students’ writing ability more precisely, and a pronunciation assessment is also necessary for a more accurate evaluation of speaking ability. In other items, information from the two tests is complementary for the whole performance assessment.

In terms of students, this study showed that students could be divided into three categories: 1) those who were better at speaking, 2) those who were better at writing, and 3) those who showed a positive correlation between writing and speaking ability. The study also showed that in grammar students were weaker in writing than in speaking, which leads us to the conclusion that we need two performance tests (Speaking and Writing).

The present research on the two types of performance tests demonstrates the need for two-dimensional performance assessment involving both speaking and writing. This type of assessment makes it possible for student performance and ability to be measured more accurately and enables raters to make judgments that more precisely reflect students’ performance in tests.

**Assessing communicative competence in adult ESL**

Daniel S. Pittaway & Aaron J. Kelly, California State University, Fullerton

The presenters will discuss the day-to-day assessment of communicative competence in adult ESL classes. One portion of the poster will discuss the evaluative aspects of lesson planning within the context of a pedagogical method the presenters developed called MPMP (Modeling, Pairing, Mingling, Presenting), which is an extension of the Engage-Study-Activate (ESA) (Hamm, 1998) model of lesson plan sequencing.

The other portion of the poster will discuss a general assessment instrument developed for end-of-unit assessment of students who have been instructed via the MPMP method. This part of the presentation will present some guidelines to develop a performance-based classroom test that is compatible with the MPMP method, and will also discuss the development of a scoring rubric that accurately assesses the relevant areas of communicative competence (Canale & Swain, 1980; Bachman, 1990; Bachman & Palmer, 1996) that adult ESL students should demonstrate within a particular target language use (TLU) domain (Bachman & Palmer, 1996).
Developing a rating scale from the testing context

Lia Plakans, The University of Iowa

This poster will narrate the revision of a rating scale for writing assessment in a pre-academic intensive English program. The scale, used to rate a 30-minute writing test, determines course level placement. Before this change, the writing tests were rated using a standard rubric with separate scores in five areas: content, organization, vocabulary, language use, and mechanics. However, the teachers in the program, who were also the raters, felt that this standard rubric scale was not serving its purpose efficiently or effectively.

After this determination was made, the presenter, who also teaches in the ESL program, set out to develop a new rating scale. Upon reading an article by Turner and Upshur (2002) on establishing rating criteria from within the testing context, the program coordinator and teachers agreed to try this alternative method to develop a new rating scale for the writing placement. This process proceeded through several stages.

The first stage sought to draft the rating criteria following the process detailed in Turner and Upshur (1996). A sample of 12 compositions were read and sorted by a group of five raters who had extensive experience teaching writing and rating the placement test. From this session, the first draft of the scale was developed. This scale, similar to a flow chart, asked a series of yes/no questions to determine the final placement score for an essay. This first draft was then used to rate compositions by two other raters to determine reliability.

After sufficient reliability had been determined, the rating scale was implemented for placement in the Spring semester of 2003. While placement decisions based on the scores were satisfactory, raters experienced difficulty in using the scale. Three major issues were using length of the writing as criteria, interpreting subjective questions, and placing writers (a high number of scores fell in the middle range with few at the high and low ends of the scales).

The next stage attempted to deal with these problems. This involved revising the rating scale to include more guidance for each question. In addition, a small study of word counts in past tests was conducted to determine if the question involving length was valid and to detail what approximate word count distinguished the levels. To resolve the placement issue, a second step in rating was added to rank writing within each score group.

After these changes were implemented, the rating scale was used again. As with prior testing, rater reliability was evaluated as well as the validity of final placement decisions. In addition, the writing tests from the previous two semesters were used to develop profiles of writing at each score level. The purpose of this step was to provide the program director with information to refer to when discussing placement with students. Also, it served to supplement the scale during rating. In addition to this narrative, the poster will include a summary of lessons learned from this process as well as an agenda for further research for the ESL program in regard to the scale.

An adaptation of the Modern Language Aptitude Test for Spanish-speaking children

Daniel J. Reed, Indiana University and Second Language Testing, Inc.
Charles W. Stansfield, Second Language Testing, Inc.

This test development project extends to Spanish an evolving framework for creating adapted versions of the Modern Language Aptitude Test (MLAT) for examinees whose first language is not English (Stansfield and Reed, 2003). The primary test-taker population in this case is Spanish-speaking children in the United States. A directly translated version of the elementary version of the MLAT (MLAT-E, formerly EMLAT) is not possible due to the fact that 3 of its 4 parts are English-based. Therefore, the rationale and specifications for the adapted version are presented as a main component of the poster. Special consideration is given to language-specific features of Spanish in the creation of item formats that target the components of language aptitude identified by Carroll (1962, 1981, 1993). The framework presented also includes a plan for field testing, norming and establishing the predictive validity of the new form. Results from preliminary field testing will be displayed and analyzed. In addition, specific uses for the adapted test will be discussed including selection, placement, guidance, and the diagnosis of foreign language learning disability. The discussion of these uses will take into account recent work on language aptitude (cf. Grigorenko, Sternberg & Ehrman, 2000; Robinson, 2001, 2002; Skehan, 1998, 2002).

Training nonnative speakers of English to perform holistic scoring of same language group (Chinese) nonnative speakers speech samples in the LanguEdge™ Instructional Tool

Marcia J. Reeves, Theresa Richter, & Terry Cryan, Educational Testing Service

This past year the presenters faced the challenge of preparing nonnative speakers of English in China to not only conduct holistic scoring of other nonnative speakers of English, but to train "Master Trainees" to train other Chinese teachers to score the English speech samples of their students using the materials in TOEFL's LanguEdge Product. We would like to share the materials and methods we developed, as well as our discoveries in training nonnative speakers to score
English speech samples of speakers in their same language group. We think our findings offer some interesting research pursuits, i.e. Do same language groups evaluate English speech samples differently that native speakers? If so, can raters of same language groups be trained to hear the speech as a speaker of English would? What kinds of materials should be used to facilitate accurate scoring results?

Using item response theory to create a common scale for comparing results from large-scale prefectural English proficiency tests

Chisato Saida, Namiki High School
Tamaki Hattori, University of Tsukuba

English educators in Japan need to develop a method for comparing results obtained from different tests on a common scale. Results from tests administered in Japanese schools are generally analyzed and reported using a classical testing theory model and this means results obtained from different tests cannot easily be compared. For this reason, although concerns are increasingly being raised that revisions to the Courses of Study by the Ministry of Education, Culture, Sports, and Technology are lowering standards, data for investigating changes in scholastic ability are not available. The application of Item Response Theory to test development, however, can make comparisons among students’ achievement levels possible by allowing results from different tests to be equated on a common scale. This would also provide a data-based method of assessing the effectiveness of revisions to the Courses of Study and other educational reforms. To date, however, there has been no research published into applying IRT to large-scale English tests administered in Japanese schools (that is, conducted at the national or prefectural level).

This study describes the use of Item Response Theory to derive ability parameter estimates on a common scale for approximately 140,000 high school students. The research utilized results from prefecture-wide English tests administered in Ibaraki prefecture, Japan, over an 8-year period. As the tests did not include any common items, 7 additional tests were specially produced to enable the equating of ability parameter estimates through a common test-taker design. Changes in the patterns of ability scores from year to year were identified and the results suggest that the English ability of students entering high school gradually decreased over the 8-year period covered in the study. The paper discusses the possibility of a relationship between the changes in average annual ability scores and the introduction of the revised Course of Study guidelines for junior high schools in 1989. Finally, the paper argues that item response theory needs to be applied systematically to the development of large-scale English tests in Japan.

Eigoquest: A game/test providing proficiency measurement and, hopefully, motivation

Randy Thrasher, Okinawa Christian Junior College
Katsumi Yuasa, International Language Centre

This poster reports on the development of a test presented in a computer game format. This game/test was created in response to a request for a low stakes measure of the English proficiency of Japanese technical college students. The purpose of the test is to give these students, who are both weak in English and lacking in motivation to study that language, an indication of their proficiency and hopefully, to encourage them to raise their proficiency by further study either in the English classes offered by the technical colleges or on their own.

We believe that this project and the resultant product will be of interest to LTRC participants because it has given us the opportunity to experiment with unusual ways of informing the test takers of their success or failure on individual items, to use hints or other information before giving a second attempt to answer a question initially mistakenly answered, to utilize computer adapted testing techniques to assure that all test takers finish the test within certain time constraints and to report overall test results to test takers graphically.

We are presently beginning the piloting of the test and expect to be able to report on how well the various item formats function, the degree to which the test yields useful information to determine proficiency, and test taker feedback on questionnaires and in post test interviews.

Language test preparation over the Internet - High stakes interpreter testing de-mystified

Kathy Tinnios-Rose & Christian N. Jones, CPS Human Resource Services

CPS Human Resources Services develops and administers the Court Interpreter, Administrative Hearing Interpreter, and Medical Interpreter (Worker’s Compensation purposes only) certification tests for the State of California. In an effort to make more qualified, certified interpreters available to the State, CPS has embarked on a long range action plan to assist candidates in preparing for these difficult assessments. Due to the consequence of error in interpreting in these high-stakes fields where mistranslations may cause a mono-lingual speaker medical harm, imprisonment, deportation and in severe cases, death, it is imperative that interpreters possess high levels of language proficiency. Because many people believe they are “bilingual enough,” candidates in these fields fail the certification test in large numbers.

In an effort to assist candidates in understanding the requirements of the interpreter profession for these...
occupations, and provide candidates with a forum for self-assessment, CPS staff worked extensively with language experts, information technology experts, and our Webmaster to create comprehensive online candidate preparation materials. This website provides candidates with a unique opportunity to compare their language proficiency and interpreting skills with those required of the profession. One of the more innovative features of the website is the delivery format. Candidates who access the multiple-choice sample tests online will take a computer-based written test that will score automatically and provide instant feedback on their performance. For certain components (e.g., "Reading Comprehension" or "Spanish Grammar") detailed grammar rules and acceptable translations are included for each question to provide additional instruction and feedback. The website also contains practice oral test "scripts" that allow the candidate to play and replay oral components of the sample tests. Multiple renderings of the most difficult exam components are included to further assist candidates in gauging the required standard of performance.

By providing candidates with examples of language interpretation for several test components, it is anticipated that borderline candidates will gain the knowledge and skills needed to be successful and become certified. Conversely, it is hoped that candidates who do not possess the needed skills will seek additional training by qualified educational providers prior to taking the test.

A comprehensive demonstration of all aspects of the website will be presented and time for questions and answers will be included.

TOEFL influence in the ESL Placement Test (EPT) at UIUC

Hui-jeong Woo, University of Illinois at Urbana-Champaign

Officially, the first administration of Test of English as a Foreign Language (TOEFL) was established in thirty-four countries to 592 examinees in early 1964. But since the number of foreign students already began increasing after World War II, some universities and institutions had created and operated their own local tests to assess foreign students’ English proficiency and provide appropriate level of English as a Second Language (ESL) classes. Many educators and school administrators also began to evaluate the adequacy of existing English language tests for foreigners. They thought there could be a problem of objectivity in local tests including test security and scoring problems. In addition to a problem of objectivity, they felt a need of a test that foreign students even can take in their home country before they come to the US. Indeed, TOEFL was born as need of greater standardization and objectivity in 1964.

The university of Illinois at Urbana-Champaign established its own ESL Placement Test (EPT) during the spring of 1947 and ESL instruction began in the fall of 1947 (Davidson & Cho, 2004). From 1947 to 1978, the UIUC ESL Placement Test (EPT) consists of structure, aural comprehension, composition, and pronunciation. TOEFL records are found from 1974 in students’ records. If TOEFL is a rough screen of students’ English fluency and provides confidence to school about students’ English skills, EPT has been argued to meet the needs of the school. Since many US universities and colleges started to have cut-scores of TOEFL as one of the primary criteria for foreign students’ admission, attention on TOEFL has increased amongst students desiring to attend US universities. Consequently they have had to study much more to get a TOEFL score worthy of admission. This may have also influenced UIUC EPT and students’ performance in their ESL classes.

This study would examine the extent to which TOEFL affected test performance on the UIUC EPT and ESL classes using the students who were admitted at UIUC from the very beginning of EPT in 1947 until the present.

Three leaves of analyses will be reported.
1. Comparison of ESL class performance between student of pre-TOEFL era (1947-1973) and post-TOEFL era (1974-present).
2. Comparison of EPT performance across groups of pre-TOEFL era and post-TOEFL era.
3. Comparison of EPT performance across native language groups in each era.

The relationship of performance on the UIUC English as a Second Language Placement Test (EPT) to scores on the Test of English as a Foreign Language (TOEFL)

Youngmi Yun, University of Illinois at Urbana-Champaign

The purpose of this study is to determine the relationship of scores on the Test of English as a Foreign Language (TOEFL) to placement decisions made by performance on the academic writing test used as a part of the University of Illinois at Urbana-Champaign (UIUC) English as a Second Language Placement Test (EPT). This study provides information regarding the concurrent validity of the EPT, which should be useful for test interpretation and test use.

There are three main research questions to be answered in this study: (1) To what extent does the relationship between TOEFL and performance on the EPT provide evidence for the concurrent validity of the EPT? (2) Is the UIUC TOEFL cut-off score of 610/253 (paper-based/computer-based version) justified? Which score, either TOEFL total or Test of Written English
(TWE) score, has more predictive value in the UIUC context?; (3) What is the relationship between student characteristics and their performance on the TOEFL and the EPT? To what degree do students' gender, language backgrounds, or majors affect the performance on the tests?

The validation study was executed using TOEFL total score and TWE score as criteria. The validity coefficients failed to provide positive evidence for the validity of the EPT. Nevertheless, due to several methodological limitations inherent in this study, it was difficult to arrive at a conclusive statement about the concurrent validity of the EPT solely based on this correlation coefficient.

Logistic regression analyses were performed through SPSS NOMREG to assess prediction of membership in one of three categories of outcome (EPT scores). The result showed that TOEFL total score significantly predicted the EPT score but explained only minimal variation in the EPT score, according to McFadden’s $\rho^2 = 0.045$. TWE score did not even predict the EPT score significantly. Moreover, it was shown that an estimated TOEFL score of 576 (paper-based) was required to be "Exempted" (EPT score 4) as a result of the performance on the EPT. A student with the current cut-off score of 610 was 3 times more likely to receive the EPT score 3 than score 4 (Exempted) on the basis of the analyses. Therefore, it was concluded that the current cut-off score was not empirically justified.

In order to examine the relationship between student characteristics and their performance on the TOEFL and the EPT, one-way analyses of variance (ANOVA) followed by a series of post-hoc tests were carried out. It turned out that language and majors significantly affected students' performance on the tests, while gender did not. It was, therefore, concluded that there exists significant differential performance on the tests that are associated with language and major fields of study.
Works in Progress Abstracts
(in alphabetical order, by first presenter's last name)

The continuum of the listening process from a less skilled to a skilled FL listener
Joanna Ancker, Åbo Academy University

It seems to be established that the main difference between FL listeners on different levels is that advanced or skilled FL listeners have the ability to use a combination of bottom-up and top-down strategies in order to grasp the main points in a text, and reconstruct the development of these ideas. In the case they miss something, they don’t panic, but go on listening, searching for possible hints on the parts they didn’t catch right away. The less successful listeners, on the other hand, get stuck on words they don’t understand or facts they want to remember, consequently often missing the information that follows. (See for example Yepes, 2001). There are also differences in the listeners' ability to make use of various comprehension strategies or tactics (Goh, 2002). The higher ability listeners tend to use both cognitive and metacognitive tactics in a more efficient way compared to lower ability listeners. Vandergrift mentions the metaphor of “orchestrated strategy use” associated with the skilled listener, where “the metacognitive strategies oversee the process, directing the deployment of appropriate cognitive strategies to interact with the input and achieve the final goal of comprehension” (Vandergrift, 2003, 485). The less skilled listener sometimes tends to translate on-line, thus being incapable of keeping up with the incoming input, experiencing greater difficulty holding meaning in memory.

This describes the two extremes of the continuum of listening ability: the processes of skilled and unskilled listeners. What is interesting is the long stretch of the continuum between these two far ends. What characteristics are found in the listening processes of individuals placed somewhere in between the extremes? Is it possible to describe or determine the small or greater steps leading from the unskilled end to the skilled end? Is it possible to place an individual on the scale, at a given moment, independently of the type of listening test item given? How do the processes of individuals with varying abilities differ according to the item? In other words: is there a difference between the listening processes of a learner activated during a multiple choice- item and the processes during an open-ended task? Are the variables associated with a task a function of the characteristics of the L2 listener? Can all learners advance on the continuum, provided that they get sufficient and appropriate teaching or input and strategy training? How are these different listening processes reflected in the results of a learner in a high-stake language test situation like the Foreign Language Test of the Finnish Matriculation Examination?

In order to find answers to these questions, I will study a listening comprehension test of French as a foreign language of the Matriculation Examination administered in spring 2002. Statistical and qualitative analyses of this test and the test results will be completed with applied tests given to subjects with the same background language education, with item type variables: the formats used will be multiple choice -, open-ended, partial dictation and r_sum_. Different retrospective and introspective methods of describing the listening comprehension processes on different language competence levels will be studied and assessed.

An investigation of the competencies of intermediate proficiency language learners
Nathan T. Carr, California State University, Fullerton

This presentation relates to the new university-wide second language graduation requirement at California State University, Fullerton. This requirement states that as of 2006, all incoming undergraduate students must complete three semesters of foreign language instruction, or demonstrate an equivalent level of competence. Many students are expected to attempt to test out of this requirement; the University must therefore select or create appropriate tests of reading, writing, listening, and speaking in a number of languages in the relatively near future. Doing this requires the creation of test specifications that are appropriate for our particular student population, and this in turn requires the identification and description of common language use tasks that learners with the desired level of ability can perform. This project will identify these tasks in the language classes taught in the CSUF Department of Modern Languages and Literatures.

The washback of a new testing system on students preparing for Japanese university entrance examinations.
Mark Chapman, Hokkaido University

The Japanese Ministry of Education has recently announced that it wishes university entrance examinations in Japan to more closely resemble the TOEIC and TOEFL tests. The purpose of this research is to investigate what impact such a change would have on school students who are preparing to sit entrance exams.

The author intends to investigate how the TOEIC (Test of English for International Communication) influences test takers. The main reason for selecting the TOEIC over the TOEFL is the relative lack of existing research
into the TOEIC. This lack of independent investigation coupled with the importance of the TOEIC (over two million test takers per year), makes the test of prime importance to language testing researchers in Japan. Through a qualitative study the author intends to explore how the TOEIC motivates test takers. Initially the study will be conducted by questionnaire and later by interview, in an attempt to gauge how the TOEIC has influenced previous candidates. Of specific interest is the extent to which candidates are motivated to develop productive linguistic skills given that the TOEIC is a discrete item test of listening and reading.

Several years ago the Japanese Ministry of Education stated its hope that university entrance examinations would be constructed to reflect the need for improved communicative competence in Japan. As a result, some Japanese universities have introduced writing and listening tests to complement the more traditional testing of grammar and translation. It is unclear how this latest shift in testing policy will impact school students. The author believes that this study can shed light on how the TOEIC motivates test takers and hence, how the new policy will influence school students. The chief aim is to compare the washback of the TOEIC with the government’s stated aim of fostering higher levels of communicative competence in students within the compulsory education system. It is hoped that through presentation at LTRC, the author can gain advice from more experienced language testing researchers on how to proceed with the qualitative study.

Investigating the relationships among responses to open-ended and multiple-choice questions on an L2 reading test and performance on an L2 writing test

Glayol Ekbatani, St. Johns University

This study explores the relationship between performance on open-ended (OE) and multiple choice (MC) questions in L2 college readers and the relationship between this and their performance on writing tests. The study was prompted by a significant disparity that we have often observed between the scores of some test takers in MC reading tests and their scores in other skills, such as writing, leading us to hypothesize that these test takers might have a problem with the format of the test and not the skill. L2 researchers have long noted the limitations of the MC format of testing in reflecting the actual skills of L2 readers Gordon (1987). We developed an open-ended form of a reading comprehension test to be administered in conjunction with MC questions and a writing test. The participants were 87 international students taking the placement test at a large private university. The instruments were developed in-house and were refined through repeated piloting. Pearson correlations will be used to examine the relationships among the three variables.

Invaluable use of verbal reports for affirmation of items

Karen Englander & Leonora Velasco, Universidad Autónoma de Baja California

Multiple choice tests continue to be used for assessing student performance so their manner of construction is critical to both test creators and, especially, test takers (Thorndike 1971). At a Mexican university, every graduating student must pass a multiple-choice exam of English competence which has been created in a partnership between the university’s school of languages and the faculty of educational measurement. The exam items were initially subjected to statistical measures of difficulty and discrimination and adjusted until they fell within the accepted range. However, as teachers of language and not testers per se, we at the language school were particularly interested in how students were making their answer choices and whether the specified objective of each test item was being met.

Several practitioners (Cronbach 1984, Alderson 1990, Baxter & Glaser 1993, Paxton 2000) have stated that verbal reports (Ericsson & Simon 1984) might help test creators understand why test takers choose the answers they do. In our study, we were interested in verbal reports not as windows to the cognitive processes of the mind itself (see critique by Smagorinsky 1998), but rather as a report of activity from short-term memory.

Bearing in mind the limitations of verbal reports, we set the following goal: to elicit what students said about how they chose a particular answer from among the four multiple choice options. We adopted Ericsson & Simon’s (1984) assertion that short-term memory is accessible. We did not presume that we could uncover cognitive processes. The verbal reports were elicited to reveal what data test-takers used to solve the task. It is on this project we will report.

To date we have found that verbal reports confirmed that some test items operated as intended, but they also revealed problems with items that otherwise had been undetected. In fact, our initial findings revealed that one third of the items that had been statistically acceptable became invalid when subjected to a verbal report. Alterations are now being made to the exam and further statistical evidence will be sought. It seems clear, too, that test items which seem well constructed both by observation (Moss 1992) and statistical analysis still need to be subjected to the evidentiary procedure of verbal reports.
Coping with the impossible... - A collaborative approach to national test development

Gudrun Erickson, Göteborg University

The relationship between assessment and learning is far from always a positive one. Summative tests in particular are often regarded as something that hampers rather than enhances learning. Formative assessment, on the other hand, is commonly believed to have a larger potential both for the individual and for the educational context.

The project in focus is set in the Swedish school system, in which an extensive program of national assessment is offered in connection with national curricula and syllabuses. The tests are of different kinds-low-stakes diagnostic materials, aimed to be used flexibly in the classroom, exist parallel to high-stakes summative tests at the end of compulsory school and in upper secondary education. All materials have several aims, something that is often considered theoretically less recommendable, or even impossible. The test of EFL at the end of compulsory school, for example, aims (a) to support teachers in their decisions regarding students' competencies in relation to the national goals and grading criteria; (b) to enhance comparability and fairness within the school system by facilitating equal interpretations of the nationally decided goals and criteria; (c) to clarify the communicative and functional view of knowledge and language as expressed in the national curricula and syllabuses.

To cope with the situation, a collaborative approach is used in the process of developing and evaluating the assessment materials. This entails, e.g. cooperation with large and diverse groups of researchers, teacher educators and experienced teachers throughout the process. Moreover, in the different piloting and pretesting phases, all teachers and students taking part are encouraged to give their comments on various aspects of the materials. Different methods are used. Feedback from the test-takers, for example, is collected orally and informally, as well as in large-scale questionnaires comprising Likert scales and open questions.

Qualitative as well as quantitative methods are used to analyze the data. The results are used internally, in the continuous developmental and validation work, as well as externally, in teacher guidelines, information materials, etc.

In the presentation the collaborative approach as such is discussed, and some examples of findings concerning both teachers' and students' opinions and attitudes are presented. This may involve, e.g., teachers' conceptions of difficulty, or what seems to constitute students' positive as well as negative opinions about tasks of different kinds. Also, analyses have been made of the relationship between students' attitude and performance, attitudes specific to different sub-groups, and gender-based differences in self-assessment.
It is believed that an open, collaborative and dynamic approach, in which different stakeholders' voices are heard, contributes positively to the validity of the tests. The ultimate hope is that even large-scale, summative, national tests may, at least to some extent, be tests not just of learning but also for learning.

**Addressing the challenges of assessing ELLs in the K-12 context**

Shelley Fairbairn, *The University of Iowa*

As a veteran ESL teacher, I am well aware of the challenges of teaching and assessing English language learners (ELLs). In fact, watching my colleagues struggle to prepare high-stakes tests mandated at the district level in Iowa played a role my return to "the other side of the desk" as a doctoral student; I felt the need to seek out answers to the challenge of effective assessment, particularly that of ELLs. Thus far, I have addressed four facets of this challenge:

1. the psychometric facet (how to obtain reliable scores that lead to valid inferences regarding the content-area knowledge of ELLs),
2. the policy facet (how to fulfill both the letter and the spirit of No Child Left Behind [NCLB] requirements),
3. the teacher education facet (how to ensure that teachers are well-educated and well-supported in the area of test development), and
4. the collaboration facet (how schools and universities can partner to ensure improved student assessment).

In my work as a graduate student, based on the recommendations of LaCelle-Peterson & Rivera (1994), I initiated a school-university partnership in test development as a means of addressing the psychometric and collaboration facets of the ELL assessment challenge. A fellow graduate student, Maria Martin, and I evaluated an existing high-stakes reading test for 7th grade students with an eye toward making it more accessible for ELLs. Though the focus of our project was linguistic simplification of the test items, building on the work of Abedi, Lord, Hofstetter, & Baker (2000), we found that the assessment was in need of a number of technical modifications. This, coupled with the teachers' initial resistance to linguistic simplification, made clear the need for increased collaboration and for professional development opportunities for teachers in the area of test development (the teacher education facet of the ELL assessment challenge).

I have begun to address the policy issue through an in-depth study of NCLB mandates regarding the assessment of ELLs, done with Ms. Martin. In addition, I have participated in discussions regarding the Wisconsin Alternate Assessment for English Language Learners (Wisconsin Department of Public Instruction, 2003) and the WIDA Consortium's alternate content assessment for ELLs, both designed to fulfill policy mandates. The enormous complexity of the NCLB legislation and its hurried implementation timeline seem difficult to deal with, at best. However, the WIDA Consortium states have taken advantage of the opportunity for ELLs to participate in alternate assessment and are developing those systems. My knowledge of these types of assessments is enabling me to better understand ways in which to address both the policy and psychometric facets of the ELL assessment challenge.

In answer to the need for collaboration, I have also conducted a qualitative study investigating the specific challenges that middle school teachers perceive in the assessment of ELLs. Two content-area teachers and an ESL teacher from a suburban school district were interviewed in order to ascertain their views on these matters. Their perspectives and needs will play a role in the development of my research agenda.

In the "Work in Progress" session, I will be seeking input regarding how to focus my work in order to make substantial contributions to the field of language testing while simultaneously serving the needs of K-12 ELLs and those involved in their education. Specifically, I hope to gain direction for my dissertation research.

**Reflective item types in The Canadian Academic English Language (CAEL) Assessment: An examination of the method effect**

Wendy Fraser, *Carleton University*

Recent studies on the effect of test method (Bachman, Lynch, and Mason, 1995; Corrigan and Upshur, 1982; Lazaraton, 1996; Lumley and McNamara, 1995) have shown that various facets of the test itself (Bachman, 1990), such as item type (Upshur and Turner, 1999), genre (Fox, 2003), and task type (Kenyan, 1995) can have a systematic effect on scores. For example, several studies have documented task effect in oral proficiency tests. Kenyan (1995) explored the effect of difficulty in performance-based oral proficiency tests. Shohamy and Stansfield (1991) and O’Loughlin (1995) examined the scores derived from direct and semi-direct interviews, finding a high correlation between them, but noting that the type of language elicited by each test was different, as a result of the interaction between the test taker and the interlocutor. Other studies have shown that item types can affect test takers' scores. For example, Shohamy (1984) and Gordon (1987) examined the effect of multiple choice and open ended item types on test taker text processing and recorded significant differences in the scores. Ellis (1994) has suggested that such test method facets as task formality and planning time can affect the discourse produced.

The purpose of this study is to explore the effect of re-
Reflective test items on the scores of writing produced in the context of an academic writing test, specifically the Canadian Academic English Language (CAEL) Assessment (Fox, 2000). Over the past two decades, many researchers (for example, Elbow, 1991; Emerson, 1983; Faigley, 1994; Fulwiler, 1988; Moon, 1999, to name a few) have examined the role of reflection in developing writing ability. It is generally agreed that reflecting on the meaning of information and articulating this meaning in writing allows writers to "discover" (Britton, 1975; Engin, 1983; Freedman, 1987) and expand on their thoughts, ultimately writing with more clarity and richness. However, there is no research to date examining the effect of reflective test items on the discourse produced in a test. This present study explores the method effect, if any, of the three reflective items at the end of the first reading on test taker essay response.

Rating video oral proficiency instruments

Gene Halleck, Oklahoma State University

The purpose of this study is to investigate whether the ACTFL Guidelines could be used successfully to rate Video Oral Communication Instruments (VOClS). According to Higgs (n.d.), VOClS could be rated using the Guidelines even though these guidelines were created for another instrument (the Oral Proficiency Interview - OPI).

Both Barnwell (1985) and Halleck (1996) found that interrater reliability was not sacrificed by using appropriately trained student raters to make proficiency judgments on OPIS. In this study graduate students in an oral testing seminar were trained to use the Guidelines. Trainee's ratings were compared with the judgments of a certified ACTFL tester. The entire corpus consists of 90 VOClS, 30 at each of three levels. The subjects were all International Teaching Assistants who participated in the testing as part of an orientation program for prospective classroom teachers. At this point in the analysis 60 tapes have been rated and analyzed - 20 Superior, 20 Advanced, and 20 Intermediate.

Each VOCl was rated by the certified tester and five (5) trainee raters. The data indicate that Intermediate interviews resulted in the highest level of interrater reliability. For 81% of paired ratings concerning Intermediate VOClS there was perfect agreement; 17% of the ratings were one level higher (Advanced), and only two percent of the ratings for the Intermediate VOClS were rated two levels higher (Superior). The next highest interrater reliability occurred at the Superior level, where 69% of the paired ratings resulted in perfect agreement; 31% of these VOClS were rated one level lower. The level that caused the most difficulty for the trainee raters was the Advanced level where only 66% of the ratings agreed with the trained rater’s judgment and 34% were off by one level (22% of the ratings were one level lower and 12% of these VOClS were rated one level higher). Thus this study does confirm Halleck's (1996) finding that Superior and Intermediate samples resulted in the highest level of interrater reliability, and that Advanced level samples were more difficult for the trainee raters to judge.

Although the results do not match the extremely high interrater reliability of OPI ratings found by Clark (1978), Adams (1978), Bachman & Palmer (1983), or Shohamy (1983) on OPI data, they do indicate that with training, graduate student raters could achieve perfect agreement with a trained rater in 66% to 81% of the cases. The question remaining is whether this level of agreement is the best that can be expected with such data and whether this level is acceptable for low-stakes testing situations.

A Spanish reading screening test for a Master's program in Mexico

Diana Jenkins & Harold Ormsby, Universidad Nacional Autónoma de México

This is a gatekeeping test for the Masters in Applied Linguistics program of the Linguistics Graduate Program of the National Autonomous University of Mexico. The test was administered for the first time in 2003 to 37 applicants. It will become a permanent fixture in the program's admissions process, which also involves tests of English reading comprehension and Spanish writing, as well as an in-depth interview. Most, but never all, test-takers are native speakers of Spanish.

The need for the test was determined by the facts that (1) all of the program's courses require some, if not a great deal of reading in Spanish, (2) all teaching, including teacher- and student-made materials, as well as any tests, are in Spanish, and (3) all papers and the thesis must be in Spanish, as well. Over the years, it has been seen that any weakness in Spanish reading ability diminishes a student's chances for success. Weaknesses can have any of a number of causes; an interesting one is that many applicants are trained as foreign language teachers, and may well have had very little college-level experience and training in Spanish, even though it is their native language.

While there is general agreement that a Spanish reading test is desirable and that the 2003 test strengthened admissions decisions, virtually all aspects of the test are very much open to reflection, research and development. The difficulties of researching and developing a screening test that is in use still have to be faced. And in fact, whether the test is deemed to being assessed as "quick and dirty" or can deserve a more dignified description is also an unanswered question, although of course, in view of the stakes involved, the former is very unattractive.
The text for the 2003 test consisted of a 4750-word article on implicit attention to form in language teaching, which was slightly edited for this purpose. Two tasks were designed. (1) An outline of the edited text was prepared and about one-third of the “points” (major, intermediate and minor) were deleted and turned into blanks. The instruction was to compare the text and the outline, and to fill in the blanks in about 60 minutes. (2) An experiment is reported in the article. The instruction was to locate the report in the text, and to fill in a form with “hypotheses,” “subjects,” “the grammatical point researched,” “condition 1,” “condition 2,” “condition 3,” “results” and “conclusions.”

Final scoring keys were set only after reading all protocols, so as to take all possible correct answers into account. The range of variance found was remarkable and, it would seem, very informative. In fact, some admissions committee members went away with the impression that this test was more telling than the composition.

We will make available the 2003 test, its results, and some results of on-going research, particular error-analysis. We will also share the most recent draft of the 2004 test. Our purpose for presenting this work in progress is to get colleagues’ overall comments, criticism and suggestions. We would also be very happy for a colleague to use the test for any appropriate purpose.

Regular expressions in automated scoring: Uses and limitations
Shoichi Gregory Kamei & Yuko Haga, University of California, Los Angeles

One of the promises of technological solutions to language testing is the reduction of time associated with scoring responses. Matching is the basis for most automated scoring systems. However, most users think of this as matching key words or phrases. This is straightforward for selected response items, where there is only one correct response specified by the item writer. For open-ended response items, even limited response questions, the variability of acceptable and actual responses is quite wide. In such situations, one-to-one key word matching is found wanting.

This work-in-progress session describes a research project on the use and limitations of regular expressions (REGEX). Regular expressions are not a programming or scripting language, but rather a way of prescribing character strings that can be interpreted by many programming languages, such as Perl and Java. The focus of this study is the utility of REGEX and how automated scoring using it differs from human scoring. In this session, an overview of REGEX and examples of regular expressions will be provided.

Challenges of World Englishes to language testing in EFL contexts
Hyun-Ju Kim, University of Iowa

The fact that English is widely used as an international language makes us revisit the testing of English as a foreign language in international contexts. EFL learners generally do not know that many varieties of English exist worldwide, including outer and expanding circles as well as inner circles in Kachru’s (1986) terms. English learners and educators should be aware of the world English perspectives and shape new communicative curriculums and evaluations that are more sensitive to their own contexts.

The population of English speakers is increasing, not only in native-English speaking nations, but also in foreign language contexts (Crystal, 1997). Therefore, English spread is happening widely, and English change is an unavoidable phenomenon. Thus, we need to rethink the concept of a variety of Englishes, which are nativized or localized. It may be a big fallacy to think that the variety of English is “interlanguage” or “learner language” because it is used in its own localized way. We also need to consider that increasingly in EFL contexts, people have many more opportunities to speak English with ESL and EFL speakers than with ENL speakers. So, we have to reexamine language testing for English in EFL contexts. The present paper addresses why the world Englishes perspective is needed and how it can be implemented in language testing, especially for Korean students.

Korea is a prime example of English as Foreign Language (EFL) contexts. Over the 120 years of English Education in Korea, American English as a Standard English, has changed in many aspects: lexicosemantic, morpho-syntactic and pragmatic (Shim, 1999).

However, the variety of Englishes in teaching and testing are still restricted to two native models: American and British English (Friedrich, 2002). Thus, students may feel confused by the gap between the real use of English in their world and the use of English in language tests. For example, Korean students learn the codified Korean English until high school and then start to study the English for Test of English as a Foreign Language (TOEFL) or The Test of English for International Communication (TOEIC), which are based on American English. It makes them feel frustrated when they are confronted with the discrepancy between those two varieties of English. Therefore, English researchers in EFL contexts should look more closely at the evolution of the codified varieties of English. Also test developers of internationally used English proficiency tests should consider a variety of Englishes in the world.

This paper explores how a world Englishes perspective can be implemented in testing. This perspective is worthwhile for the purpose of interpreting and using test scores in EFL contexts.
Construct validation of an integrated, portfolio-based, and computerized EAP placement test: A multiple data source approach

Young-Ju Lee, University of Illinois at Urbana-Champaign

The Enhanced ESL Placement Test (EEPT) at the University of Illinois at Urbana-Champaign (UIUC) is in need of a comprehensive validation study since it was developed two years ago and is administered via a new test-taking model. This study is designed to translate Messick's (1989) validity framework into such a comprehensive argumentative model based on multiple sources of evidence. Several components of the EEPT (i.e., extended time for writing, facilitative activities, and new test-taking modality) will help illuminate its practical importance in a specific academic setting.

The present study will investigate five research questions. The first research question addresses the predictive validity of the computerized EEPT (CEEPT) based on qualitative and quantitative data. Specifically, I will investigate the extent to which CEEP scores predict international graduate students' English proficiency and ability to cope with the academic demands of their coursework. This will be investigated using three measures of academic performance – GPA, faculty evaluations, and students' self-assessments.

The second research question will investigate convergent validity (i.e., high correlations between tests of similar constructs) and discriminant validity (i.e., low correlations between tests of distinct constructs). This will be addressed by comparing patterns of correlations among various measures such as each subsection of the TOEFL and the GRE.

The third research question will investigate the effect of revision process on the quality of the second draft. Three kinds of evidence will be used to compare the quality of the two drafts. First, the number of words, T-units, and the syntactic complexity will be compared between the first and the second draft. Second, the first and the second drafts will be scored, using a holistic rating scale. Third, features in first drafts, second drafts, and content-field writing samples will be analyzed to investigate which draft of each essay has more similarities with the content-field writing samples in terms of textual quality.

The fourth research question will examine consequences of the decisions made on the basis of CEEP scores, based on interviews from selected students and faculty. Students will be asked if they perceive they are misclassified as masters or non-masters. I will also refer to faculty perceptions to see if there is any mismatch between CEEP results and language needs or deficiencies. Narratives from the CEEP test-taking experience to the first semester of graduate program will be provided.

The fifth research question addresses test fairness, lack of bias in prediction. I will investigate the relationship between CEEP scores and academic performance for males and females, by comparing two regression equations.

Construct validity evidence in favor of the CEEP will make testing practitioners consider implementing a multiple-draft essay test and computer mode in their assessment contexts at the sacrifice of logistical constraints such as time, money, and labor. It is hoped that the findings in this study will advance our understanding of ESL writing assessments and will be generalizable to other similar academic contexts.

Development and validation of a delivery software for the assessment of oral proficiency of adult ESL learners

Seyyed Abbas Mousavi, Griffith University

Oral proficiency of ESL learners can be measured through a variety of testing procedures; among them is a tape-mediated test namely Simulated Oral Proficiency Interview (SOPI) developed at Centre for Applied Linguistics (www.cal.org). SOPI was later upgraded to what is known as Computerized Oral Proficiency Instrument (Malabonga & Kenyon 1998, 2001). Following the emergence of concerns with SOPI (Norris 2001), in relation to a) the limited number of items which would potentially reduce reliability; b) the nature of the questionnaire used to determine the affective variables (attitudes) of testees; c) the "shortcoming of the research design" (Norris 2001, p 100), the current project is aimed at developing and validating a multimedia package which accounts for the inadequacies found in previous similar software. Test specifications have been developed and the project is at the stage of interface layout (Fulcher 2003) and Visual Basic (VB) coding. The current project will allow researchers to manipulate VB algorithms, delivery conditions and task types which in turn would influence both the quality of speaking performances and the adequacy of proficiency ratings. It also has the potential in doing research in relation to interface design and human-computer interaction.

Training the educators of the Federated States of Micronesia in holistic writing assessment, implications for the nation's multilingual writing goals

Samantha Ragan & Mary L. Spencer, University of Guam

In 1986 the Federated States of Micronesia (FSM) emerged as a new nation, with United States compact affiliations, following its United Nations status after World War II as part of the Trust Territory of the Pa-
specific Islands. With a population of citizens embracing multiple indigenous languages which are still predominantly oral in nature, and with national language policy delineating bilingual English and local language development in the education systems, FSM has embarked on the enhancement of its student writing curriculum. Part of this effort involved the development of National Writing Standards for Students at Grades 6, 8, and 10, writing assessment training of educators, and the assessment of teacher writing throughout the nation. Assessing teachers’ writing skills and providing training to improve their writing are part of a new national teacher certification policy.

During the summer of 2003, educator training was provided by a University of Guam Team (Robertson, Spencer and Ragan) on how to holistically assess the English writing of students at both the national and state levels in Pohnpei, and at the state level in Kosrae, Chuuk, and Yap. On the basis of this training, and standards resulting from it, the assessment of teacher writing in English and Chuukese was conducted by FSM National and Chuuk State Department of Education personnel. Other teacher and student writing assessment activities are underway now in Kosrae, Pohnpei, and Yap.

In planning the training of educators in the assessment of student English writing, questions arose about the development or selection of scoring rubrics. Rather than select existing rubrics for the defined grade levels (Grades 6, 8, and 10), the educators were guided through the establishment of national standards as the result of group readings and analysis of authentic FSM student essays from each grade level. Comparison of local versus other U.S. or international standards may need to be addressed later. Questions remain about how to facilitate the parallel development of writing assessment standards for grades 6, 8, and 10, and for teachers, in four national FSM languages (Kosraean, Pohnpeian, Chuukese, and Yapese). Added significance is attached to these writing assessment activities as they are included in the accountability requirements of the United States in the new Compact of Free Association for the FSM.

Use of cloze tests in teacher certification in Micronesia

Daniel L. Robertson, University of Guam

The Federated States of Micronesia is an island nation comprising four states, each of which is a concentration of islands and atolls in the western Pacific. Its geographic area is near that of the continental U.S., but its land area is less than that of Rhode Island. The FSM has recently renegotiated its Compact of Free Association with the U.S. Among the requirements is increased accountability for education related expenses. The FSM Department of Education is introducing certification requirements for its K-12 teachers.

It was determined that a cloze test would be used for the English language measure. The test was developed in two parts, intended to be used alternately as parallel forms. Pretesting on University of Guam undergraduates yielded item statistics which were used to reduce the number of items in each form to thirty. The two forms have now been administered to one state (Chuuk), and the data entry for those results is underway. These data and those from the other three states (Kosrae, Pohnpei, and Yap) will be available within the next two months.

Aside from the cloze test item data, the respondents’ gender, school, and teaching level were recorded.

At present, the plan is to compute reliability estimates on the responses, considering both internal consistency and the relationship of the two forms. These results, in addition to the item analysis, will help guide test revision.

It is hoped, of course, that the results will show acceptable estimates of reliability and that test revision will be minimal.

If that is the case, the database, which will likely include nearly 2000 respondents, will provide the first baseline averages for school, state, and national levels. It is expected that after this initial administration, the tests will be given alternately each year.

The purpose of this presentation is to make colloquium participants aware of this effort and to seek guidance and suggestions on how best to serve the needs of the Micronesian educators over the next few years.

An investigation into student negative reactions to peer assessing EFL individual presentation performance

Hitodoshi Saito, Hokusei Gakuen University

Rising interest in alternative assessment in the language classroom has prompted recent publications of resource books and research studies in this area (Brown, 1998; Ekhatani & Pierson, 2000; O’Malley & Valdez Pierce, 1996). Involvement of peers in the assessment process has drawn research interest as well (Patri, 2002; Saito & Fujita, 2004). A missing piece in the research endeavor on peer assessment, however, is a probe into student reactions to it. Although Cheng and Warren’s (1997) interview study which examined changes in student attitudes toward peer assessment, shed light on student reactions, detailed analyses were not presented concerning factors that influence student attitudes. Hardly known is what leads students to feeling negative about peer assessment.

The present study seeks to explore factors that influence student negative reactions to peer assessment of
EFL individual presentations. One hundred one university students in Japan gave presentations and assessed peer performance in EFL communication skill classes. After the evaluation sessions, students filled in a questionnaire on which they indicated their attitude towards peer assessment. Based on the results of the questionnaire, sixteen students who reacted negatively to the questionnaire were interviewed by two trained graduate students.

The results of interviews were all transcribed and qualitatively analyzed. Two critical concepts emerged from the analysis: students as poor disciples, who committed rating errors, knowing their own biases and looseness in rating, and the teacher as the master, who are knowledgeable and only ones whom they can trust as far as grading is concerned. Although many interviewees also recognized the benefits of assessing peer presentations, their strong fixed belief in the traditional roles they have been playing and the degraded level of proficiency they feel they possess seem to preclude their acceptance of the possibility of proficiently assessing their peers. It is reasonable to assume, from the data, that these negative reactions to peer assessment derived from a lack of training and familiarity with the instrument and understanding the purpose of peer assessment. The author recommends that the teacher take instructional action to reinforce the purposes and merits of peer assessment and negotiate with students to yield mutually satisfactory interpretations and meaningful use of assessment results.

**Celpe-Bras exam and teachers' perceptions of impact on teaching/assessment of Portuguese as a FL/L2**

Matilde Virginia Ricardi Scaramucci, State University of Campinas

The Certificate of Proficiency in Portuguese as a Foreign Language (Celpe-Bras), implemented by Brazilian Ministry of Education and administered twice a year since 1998 in 18 centers in Brazil and 25 abroad assesses the use of written and spoken Brazilian Portuguese through communicative tasks. This paper reports on partial results of a study aimed at investigating the impact of this exam on the teaching of Portuguese as a FL/L2 in some of those contexts. More specifically, it discusses how teachers involved in student preparation perceive the exam and its influence on their teaching/assessing practices. The data comprises answers of 51 Brazilians teaching Portuguese as a FL/L2 in Brazil and abroad to a questionnaire adapted from a previous study. Analysis of past exam papers and the exam manual provide evidence to complement this data. The teachers' answers reveal two main tendencies: a) exam changes teaching/assessment practices, perceived by the teachers as "different" than when the exam was first implemented; b) exam reinforces teaching/assessment practices, already based on a communicative approach. Although both tendencies were observed among teachers in the two contexts (Brazil and abroad), the second one seemed more frequent among teachers working in Brazil. One possible explanation for this could be teacher development. These results are to be taken with caution, however, as they are not evidence of effective changes on teachers' practices. In order to confirm these changes, therefore, there would be the need to observe teachers' actions, which will be done in a second part of this study.

**Language learning strategy use and language performance for Chinese learners of English**

Xiaomei Song, Queen's University

In my experience as a learner of English as a Foreign Language (EFL) and a university EFL instructor for many years, I have observed that some learners are more successful than others in their language learning. One important variable that may account for the difference, according to Dreyer and Oxford (1996), is the use of language learning strategies. The purpose of this study is to investigate, first, learning strategies reported by Chinese EFL learners and, second, the relationships between reported learning strategy use and language test performance in the Mainland Chinese EFL context.

A large amount of research (e.g., Green & Oxford, 1995) conducted in this area indicated that students who had comparatively higher levels of language proficiency generally reported higher levels and greater frequency of use of overall language learning strategies. However, some studies (e.g., Politzer & McGroarty, 1985) concluded that few statistically significant correlations existed between overall learning strategies and language performance although some individual strategy items did correlate with language outcomes. Also, some studies examined whether higher level and frequency of strategy use contributed to better language performance (e.g., Purpura, 1997, 1999). It is not clear whether there is a linear relationship between learning strategy use and language performance. In addition, only very limited studies have been conducted within the Chinese EFL context (e.g., Wen & Johnson, 1997).

In comparison with other research, it has been found that Purpura's classification and measurement techniques of learning strategies (1999) are the most proper for this current study. From a theoretical perspective, Purpura's classification on learning strategies is firmly rooted in a model of human information processing. From a methodological perspective, his studies using the Structural Equation Modeling as a primary analytic tool tested and retested models before generating a model with a fairly good statistical fit and meaningful interpretations. Also, his research focused on charac-
teristics of EFL test-takers, which is the concentration of this current study as well.

Therefore, this study will employ Purpura's questionnaire, which includes a three-factor model of cognitive strategy use and a one-factor model of metacognitive strategy use, to elicit information about learning strategy use. The study will use all 64 items of Purpura's questionnaire, but some items might be slightly adapted to fit the Chinese context. The questionnaire will, however, be translated into Chinese following the back-translation procedure to ensure validity of the instrument. The scores on a national standardized test in China, College English Test Band 4, are to be used as a measure of language performance. The participants of this study will be undergraduate students from a major university in Mainland China.

It is hoped that this study will help to clarify what learning strategies are used by Chinese EFL learners and the relationships between strategy use and language performance. This research may provide some guidance to educators and learners in terms of strategy awareness and language strategy training. Considering individual students' characteristics, researchers and instructors can give learners strategic training, help learners identify their strategies, and address more specifically individual strategy use.

**Measuring English writing achievement across proficiency levels using portfolios**

Diane Strong-Krause, *Brigham Young University*

The purpose of this project is to develop a useful scale to measure English writing achievement across five proficiency levels (novice-mid to advanced-low) at an intensive English program. To demonstrate their writing ability, students turn in a portfolio, which includes two papers (along with at least 3 drafts of each paper) completed during the course, a timed essay written at the end of the semester, and a paper discussing meta-cognitive processes that the student used while drafting/writing one of the two papers included in the portfolio. After training, teachers holistically score the portfolios on a 5-point scale (No Pass, Low Pass, Pass, High Pass, Honors). All portfolios are double rated.

Although teachers rate on similar criteria (e.g., content, organization, vocabulary, grammar, fluency, and understanding of the writing process), standards for passing differ from level to level. For example, students at lower proficiency levels are expected to use more common vocabulary, simple structures, and be less fluent than students at higher levels. The challenge is to create a scale that stretches across proficiency levels.

In an effort to link the proficiency levels onto one scale, we are introducing "linking portfolios." Teachers rate these portfolios using standards set for their level. After analyzing these ratings, we hope to have a better understanding of the relationship between proficiency levels and eventually develop a useful tool to provide feedback about progress to students, teachers, and administrators. Data collected at the end of Winter Semester 2003 will be available for the session.

**Development and progress of the Mt. SAC ESL-CAP Test**

Margaret Teske & Liza Becker, *Mt. San Antonio College*

Mt. San Antonio College is faced with placing 4,500 non-native students annually into one of its seven non-credit levels of English as a Second Language courses. An in-house computer-adaptive placement test is being developed to handle the large volume. Though more complex than using a standardized pencil and paper test, the department chose this form of placement over a standardized tool for the following reasons:

1. The in-house tool better reflects the curriculum and instruction of the program (greater content validity).
2. The adaptive format reduces anxiety of low-level students.
3. The test is quickly administered and scored immediately upon completion.
4. Pilot testing has shown very little fear of using the computer by test takers.

The authors would like to present the development and progress of the Mt. SAC English as a Second Language Computer-Adaptive Placement Test (ESL-CAP) as it stands in March 2004. The test is designed on a branching model or step ladder (Hennig, 1987, pg. 138). Test takers begin at the third level of a seven level continuum. They proceed from item to item going up or down the continuum based on correct/incorrect answers until they have answered a minimum of six items up to a predicted twelve items to satisfy the cutoff for each of two subtests: listening comprehension and reading. Each subtest contains a variety of multiple-choice question types including vocabulary from context, restatement, scanning, and inference.

The authors are seeking input regarding test construct, capturing of statistical analyses within the test programming, and the necessity of using IRT models.

**An investigation into reading test item characteristics: How text difficulties and question types interact**

Tomoko Wada, *Tokyo University of Foreign Studies*

Reading ability is often measured by using a set of reading comprehension test items. A test item is composed
of two parts: text and a question. There are varieties of ways to set these reading questions, but Wada (2001) suggests that these questions could be categorized into four groups according to the amount of information and the amount of inference necessary to answer them: Local-Literal, Local-Inferential, Global-Literal, and Global-Inferential.

A test item characteristic, often calculated by using statistical method such as Item Response Theory to describe its difficulty and discriminating power, is an interaction among the test taker's ability, the text type and the question type. Moreover, in measuring reading ability, what test writer can control before the test takers take the test is the interaction between text type and question type, and this could greatly affect the difficulty and discriminating power.

This research has been examining into the changes in item difficulty and discriminating power when different types of test questions are set with the same text. Texts with variances in text difficulties were coupled with different types of test questions and were given to EFL learners at a Japanese university to see which coupling is best suited to measure reading ability in EFL for academic purposes context. The statistics show some interesting results, and they suggest the importance in considering both text difficulty and question type in making a test item. The author hopes to advance the study to add text types and text organizations as elements that influence test item characteristics. Comments and suggestions from researchers at the conference will be a great encouragement to this study.
<table>
<thead>
<tr>
<th>Last Name</th>
<th>First Name</th>
<th>Affiliation</th>
<th>Country</th>
<th>E-mail address</th>
</tr>
</thead>
<tbody>
<tr>
<td>Åberg-Bergstrand</td>
<td>Lisebeth</td>
<td>Göteborg University</td>
<td>Sweden</td>
<td><a href="mailto:lisebeth.berg@ped.gu.se">lisebeth.berg@ped.gu.se</a></td>
</tr>
<tr>
<td>Abeywickrama</td>
<td>Priya</td>
<td>UCLA</td>
<td>US</td>
<td><a href="mailto:priya@ucla.edu">priya@ucla.edu</a></td>
</tr>
<tr>
<td>Aline</td>
<td>David</td>
<td>Kanagawa University</td>
<td>Japan</td>
<td><a href="mailto:davidadline@hotmail.com">davidadline@hotmail.com</a></td>
</tr>
<tr>
<td>Anekar</td>
<td>Joanna</td>
<td>Åbo Akademi University</td>
<td>Finland</td>
<td><a href="mailto:jaarnekar@abo.fi">jaarnekar@abo.fi</a></td>
</tr>
<tr>
<td>Bachman</td>
<td>Lyle</td>
<td>UCLA</td>
<td>US</td>
<td><a href="mailto:lh@human.nyu.edu">lh@human.nyu.edu</a></td>
</tr>
<tr>
<td>Bailey</td>
<td>Alison</td>
<td>UCLA</td>
<td>US</td>
<td><a href="mailto:abaliley@geis.ucla.edu">abaliley@geis.ucla.edu</a></td>
</tr>
<tr>
<td>Ballaster</td>
<td>Cristina</td>
<td>University of California, Davis</td>
<td>US</td>
<td><a href="mailto:rballaster@ucdavis.edu">rballaster@ucdavis.edu</a></td>
</tr>
<tr>
<td>Barbier</td>
<td>Isabella</td>
<td>Ordinate Corp</td>
<td>Italy</td>
<td><a href="mailto:berni@unimistra.it">berni@unimistra.it</a></td>
</tr>
<tr>
<td>Barni</td>
<td>Monica</td>
<td>Università per Stranieri, Siena</td>
<td>Italy</td>
<td><a href="mailto:cbartel@change-calstate.edu">cbartel@change-calstate.edu</a></td>
</tr>
<tr>
<td>Bartell</td>
<td>Carol</td>
<td>CSULA</td>
<td>US</td>
<td><a href="mailto:darrenbat@yahoo.com">darrenbat@yahoo.com</a></td>
</tr>
<tr>
<td>Battaglia</td>
<td>Darren</td>
<td>CSULA</td>
<td>US</td>
<td><a href="mailto:jimm@cal.org">jimm@cal.org</a></td>
</tr>
<tr>
<td>Bousman</td>
<td>Jim</td>
<td>Center for Applied Linguistics, Washington</td>
<td>Canada</td>
<td><a href="mailto:dbaylis@smith.ubc.ca">dbaylis@smith.ubc.ca</a></td>
</tr>
<tr>
<td>Bayless</td>
<td>Doreen</td>
<td>University of Ottawa</td>
<td>US</td>
<td><a href="mailto:llecher@minh.edu">llecher@minh.edu</a></td>
</tr>
<tr>
<td>Becker</td>
<td>Liz</td>
<td>Mt. San Antonio College</td>
<td>Italy</td>
<td><a href="mailto:guido.benvenuto@uniroma1.it">guido.benvenuto@uniroma1.it</a></td>
</tr>
<tr>
<td>Benvenuto</td>
<td>Guido</td>
<td>Università La Sapienza</td>
<td>US</td>
<td><a href="mailto:jared@ordinate.com">jared@ordinate.com</a></td>
</tr>
<tr>
<td>Bernstein</td>
<td>Jared</td>
<td>Ordinate Corp</td>
<td>US</td>
<td><a href="mailto:bowie@colorado.edu">bowie@colorado.edu</a></td>
</tr>
<tr>
<td>Bonk</td>
<td>William</td>
<td>University of Colorado</td>
<td>Egypt</td>
<td><a href="mailto:borsie@uagegypt.edu">borsie@uagegypt.edu</a></td>
</tr>
<tr>
<td>Borosie</td>
<td>Deera</td>
<td>American University in Cairo</td>
<td>US</td>
<td><a href="mailto:totnapp@yahoo.com">totnapp@yahoo.com</a></td>
</tr>
<tr>
<td>Box</td>
<td>Tom</td>
<td>CSULA</td>
<td>Australia</td>
<td><a href="mailto:glerinzlay@electronic.uw.edu.au">glerinzlay@electronic.uw.edu.au</a></td>
</tr>
<tr>
<td>Brindley</td>
<td>Geoff</td>
<td>Macquarie University</td>
<td>Canada</td>
<td><a href="mailto:lhbrooks@csie.utoronto.ca">lhbrooks@csie.utoronto.ca</a></td>
</tr>
<tr>
<td>Brooks</td>
<td>Lindsey</td>
<td>OSIE/University of Toronto</td>
<td>US</td>
<td><a href="mailto:butter@ece.ucla.edu">butter@ece.ucla.edu</a></td>
</tr>
<tr>
<td>Butler</td>
<td>Frances</td>
<td>UCLA/CRESSY</td>
<td>US</td>
<td><a href="mailto:viktoria@bcglobal.net">viktoria@bcglobal.net</a></td>
</tr>
<tr>
<td>Byczkiewicz</td>
<td>Victoria</td>
<td>CSULA</td>
<td>US</td>
<td><a href="mailto:neare@fullerton.edu">neare@fullerton.edu</a></td>
</tr>
<tr>
<td>Carr</td>
<td>Nathan</td>
<td>CSUF</td>
<td>US</td>
<td><a href="mailto:meryl.cervantes@yahoo.com">meryl.cervantes@yahoo.com</a></td>
</tr>
<tr>
<td>Cervantes</td>
<td>Marely</td>
<td>CSUP</td>
<td>US</td>
<td><a href="mailto:m-chalhoub-deville@uwo.ca">m-chalhoub-deville@uwo.ca</a></td>
</tr>
<tr>
<td>Chalhoub-Deville</td>
<td>Micheline</td>
<td>University of Iowa</td>
<td>Japan</td>
<td><a href="mailto:mymchahmoum@nifty.com">mymchahmoum@nifty.com</a></td>
</tr>
<tr>
<td>Chapman</td>
<td>Mark</td>
<td>Hokkaido University</td>
<td>HK</td>
<td><a href="mailto:michael.eng@polyu.edu.hk">michael.eng@polyu.edu.hk</a></td>
</tr>
<tr>
<td>Chau</td>
<td>Ho Fai</td>
<td>Hong Kong Polytechnic University</td>
<td>US</td>
<td><a href="mailto:chemnings9@126.net">chemnings9@126.net</a></td>
</tr>
<tr>
<td>Chen</td>
<td>Ning</td>
<td>CSULA</td>
<td>Canada</td>
<td><a href="mailto:cheung@edu.queensu.ca">cheung@edu.queensu.ca</a></td>
</tr>
<tr>
<td>Cheng</td>
<td>Liying</td>
<td>Queen's University</td>
<td>US</td>
<td><a href="mailto:yeho@ballard-virga.com">yeho@ballard-virga.com</a></td>
</tr>
<tr>
<td>Cho</td>
<td>Yeonsuk</td>
<td>Ballard &amp; Tighe</td>
<td>US</td>
<td><a href="mailto:mac@linguistics.utoronto.ca">mac@linguistics.utoronto.ca</a></td>
</tr>
<tr>
<td>Christensen</td>
<td>Mary Anne</td>
<td>University of Utah</td>
<td>Canada</td>
<td><a href="mailto:coping.ig@forces.gc.ca">coping.ig@forces.gc.ca</a></td>
</tr>
<tr>
<td>Coppin</td>
<td>Irene</td>
<td>DND</td>
<td>Poland</td>
<td><a href="mailto:mark.crusoe@britishcouncil.pl">mark.crusoe@britishcouncil.pl</a></td>
</tr>
<tr>
<td>Crossev</td>
<td>Mark</td>
<td>British Council, Warsaw</td>
<td>The Netherlands</td>
<td><a href="mailto:john.halen@csir.au">john.halen@csir.au</a></td>
</tr>
<tr>
<td>Cummings</td>
<td>Anne</td>
<td>University of Iowa</td>
<td>US</td>
<td><a href="mailto:craig-deville@uwo.ca">craig-deville@uwo.ca</a></td>
</tr>
<tr>
<td>Davison</td>
<td>Chris</td>
<td>University of Hong Kong</td>
<td>US</td>
<td><a href="mailto:bolo@csiro.au">bolo@csiro.au</a></td>
</tr>
<tr>
<td>De Jong</td>
<td>John H. A. L.</td>
<td>Language Testing Services</td>
<td>US</td>
<td><a href="mailto:dundin@asistate.edu">dundin@asistate.edu</a></td>
</tr>
<tr>
<td>Deville</td>
<td>Craig</td>
<td>University of Iowa</td>
<td>US</td>
<td><a href="mailto:fds@csiro.au">fds@csiro.au</a></td>
</tr>
<tr>
<td>Dobson</td>
<td>Barbara</td>
<td>University of Michigan</td>
<td>Canada</td>
<td><a href="mailto:robert.edwards@nexusbrooke.ca">robert.edwards@nexusbrooke.ca</a></td>
</tr>
<tr>
<td>Douglas</td>
<td>Dan</td>
<td>Iowa State University</td>
<td>Mexico</td>
<td><a href="mailto:karnemuglander@hotmail.com">karnemuglander@hotmail.com</a></td>
</tr>
<tr>
<td>Douglas</td>
<td>Felicity</td>
<td>Iowa State University</td>
<td>USA</td>
<td><a href="mailto:mwright@ets.org">mwright@ets.org</a></td>
</tr>
<tr>
<td>Edwards</td>
<td>Robert</td>
<td>University of Sherbrooke</td>
<td>USA</td>
<td><a href="mailto:mieheems@hotmail.com">mieheems@hotmail.com</a></td>
</tr>
<tr>
<td>Englisher</td>
<td>Karen</td>
<td>Universidad Autónoma de Baja California</td>
<td>Canada</td>
<td><a href="mailto:userdosy@cs.uoregon.edu">userdosy@cs.uoregon.edu</a></td>
</tr>
<tr>
<td>Enright</td>
<td>Mary</td>
<td>ETS</td>
<td>Canada</td>
<td><a href="mailto:userdosy@cs.uoregon.edu">userdosy@cs.uoregon.edu</a></td>
</tr>
<tr>
<td>Last Name</td>
<td>First Name</td>
<td>Affiliation</td>
<td>Country</td>
<td>E-mail address</td>
</tr>
<tr>
<td>-----------</td>
<td>------------</td>
<td>--------------------------------------------</td>
<td>-----------</td>
<td>------------------------------------------</td>
</tr>
<tr>
<td>Erickson</td>
<td>Gudrun</td>
<td>Göteborg University</td>
<td>Sweden</td>
<td><a href="mailto:gudrun.erickson@pred.gu.se">gudrun.erickson@pred.gu.se</a></td>
</tr>
<tr>
<td>Fairbairn</td>
<td>Shelley</td>
<td>University of Iowa</td>
<td>US</td>
<td><a href="mailto:shelley-fairbairn@uiowa.edu">shelley-fairbairn@uiowa.edu</a></td>
</tr>
<tr>
<td>Farrarworth</td>
<td>Tim</td>
<td>UCLA</td>
<td>US</td>
<td><a href="mailto:t.farrar@ucla.edu">t.farrar@ucla.edu</a></td>
</tr>
<tr>
<td>Fox</td>
<td>Janna</td>
<td>Carleton University</td>
<td>Canada</td>
<td><a href="mailto:jfox@ccs.carleton.ca">jfox@ccs.carleton.ca</a></td>
</tr>
<tr>
<td>Fraser</td>
<td>Wendy</td>
<td>Carleton University</td>
<td>Canada</td>
<td><a href="mailto:wfraser@ccs.carleton.ca">wfraser@ccs.carleton.ca</a></td>
</tr>
<tr>
<td>Fujita</td>
<td>Tomoko</td>
<td>Tohoi University</td>
<td>Japan</td>
<td><a href="mailto:tfujita@tohoi.ac.jp">tfujita@tohoi.ac.jp</a></td>
</tr>
<tr>
<td>Garanpayeh</td>
<td>Ardeshir</td>
<td>UCLES</td>
<td>UK</td>
<td><a href="mailto:garanpayeh-a@ucla.org.uk">garanpayeh-a@ucla.org.uk</a></td>
</tr>
<tr>
<td>Gokcora</td>
<td>Deniz</td>
<td>CARLA, University of Minnesota</td>
<td>US</td>
<td><a href="mailto:gockcora2@umn.edu">gockcora2@umn.edu</a></td>
</tr>
<tr>
<td>Green</td>
<td>Brent</td>
<td>Brigham Young University Hawai'i</td>
<td>US</td>
<td><a href="mailto:greenb@byu.hawaii">greenb@byu.hawaii</a></td>
</tr>
<tr>
<td>Hackett</td>
<td>Ed</td>
<td>UCLA</td>
<td>US</td>
<td>no e-mail address</td>
</tr>
<tr>
<td>Hafner</td>
<td>Anne</td>
<td>CSULA</td>
<td>Australia</td>
<td><a href="mailto:liud@unimelb.edu.au">liud@unimelb.edu.au</a></td>
</tr>
<tr>
<td>Haga</td>
<td>Yoko</td>
<td>UCLA</td>
<td>US</td>
<td><a href="mailto:yhaga@ucla.edu">yhaga@ucla.edu</a></td>
</tr>
<tr>
<td>Halleck</td>
<td>Gene</td>
<td>Oklahoma State University</td>
<td>US</td>
<td><a href="mailto:halleck@okstate.edu">halleck@okstate.edu</a></td>
</tr>
<tr>
<td>Hamp-Lyons</td>
<td>Liz</td>
<td>University of Melbourne</td>
<td>Australia</td>
<td><a href="mailto:lihd@unimelb.edu.au">lihd@unimelb.edu.au</a></td>
</tr>
<tr>
<td>Harshbarger</td>
<td>Lisa</td>
<td>Dept of State</td>
<td>US</td>
<td><a href="mailto:roloiaa@yahoo.com">roloiaa@yahoo.com</a></td>
</tr>
<tr>
<td>Hernandez</td>
<td>Fernando</td>
<td>CSULA</td>
<td>US</td>
<td><a href="mailto:flemia@exchange.calstatela.edu">flemia@exchange.calstatela.edu</a></td>
</tr>
<tr>
<td>Hiraizai</td>
<td>Akiyo</td>
<td>University of Tsukuba</td>
<td>Japan</td>
<td><a href="mailto:hiraizai@akura.cc.tsukuba.ac.jp">hiraizai@akura.cc.tsukuba.ac.jp</a></td>
</tr>
<tr>
<td>Hoodhaman</td>
<td>Darin</td>
<td>Defence Language Institute</td>
<td>US</td>
<td><a href="mailto:darinath.hoodhaman@monterey.army.mil">darinath.hoodhaman@monterey.army.mil</a></td>
</tr>
<tr>
<td>Horai</td>
<td>Tomoko</td>
<td>University of Surrey Roehampton</td>
<td>UK</td>
<td><a href="mailto:horai@uwa.com">horai@uwa.com</a></td>
</tr>
<tr>
<td>Imao</td>
<td>Yushhiro</td>
<td>UCLA</td>
<td>US</td>
<td><a href="mailto:yimao@bannet.ucla.edu">yimao@bannet.ucla.edu</a></td>
</tr>
<tr>
<td>Inbar</td>
<td>Ofra</td>
<td>Tel-Aviv University</td>
<td>Israel</td>
<td><a href="mailto:inbaro@bezeqint.net">inbaro@bezeqint.net</a></td>
</tr>
<tr>
<td>Inamori</td>
<td>Yo</td>
<td>University of Tsukuba</td>
<td>Japan</td>
<td><a href="mailto:so2505@ips.tsukuba.ac.jp">so2505@ips.tsukuba.ac.jp</a></td>
</tr>
<tr>
<td>Jang</td>
<td>Eunice Runhee</td>
<td>University of Illinois, Urbana-Champaign</td>
<td>US</td>
<td><a href="mailto:einang@uiuc.edu">einang@uiuc.edu</a></td>
</tr>
<tr>
<td>Jaquith</td>
<td>Paul</td>
<td>United Arab Emirates University</td>
<td>UAE</td>
<td><a href="mailto:jjpompa@yahoo.com">jjpompa@yahoo.com</a></td>
</tr>
<tr>
<td>Jenkins</td>
<td>Diane</td>
<td>Universidad Nacional Autónoma de Mexico</td>
<td>Mexico</td>
<td><a href="mailto:jenkins@servidor.unam.mx">jenkins@servidor.unam.mx</a></td>
</tr>
<tr>
<td>Joon</td>
<td>Young-Jun</td>
<td>CSULA</td>
<td>US</td>
<td><a href="mailto:elji@umich.edu">elji@umich.edu</a></td>
</tr>
<tr>
<td>Johnson</td>
<td>Jeff</td>
<td>University of Michigan</td>
<td>US</td>
<td><a href="mailto:chrism@cs.illinois.gov">chrism@cs.illinois.gov</a></td>
</tr>
<tr>
<td>Jones</td>
<td>Christian</td>
<td>CPS Human Resources Services</td>
<td>US</td>
<td><a href="mailto:m.koby@kanda.ac.jp">m.koby@kanda.ac.jp</a></td>
</tr>
<tr>
<td>Kawakami</td>
<td>Satoko</td>
<td>Institute for International Business</td>
<td>Japan</td>
<td><a href="mailto:kawakami@toyo.or.jp">kawakami@toyo.or.jp</a></td>
</tr>
<tr>
<td>Kennedy</td>
<td>Lauren</td>
<td>Carleton University</td>
<td>Canada</td>
<td><a href="mailto:lauren_kennedy@carleton.ca">lauren_kennedy@carleton.ca</a></td>
</tr>
<tr>
<td>Kenyon</td>
<td>Dorry</td>
<td>Center for Applied Linguistics, Washington</td>
<td>US</td>
<td><a href="mailto:dorry@cuu.org">dorry@cuu.org</a></td>
</tr>
<tr>
<td>Keynes</td>
<td>Janet</td>
<td>CSULA</td>
<td>US</td>
<td><a href="mailto:janifern@csol.com">janifern@csol.com</a></td>
</tr>
<tr>
<td>Kim</td>
<td>Jee Wha</td>
<td>Teachers College, Columbia University</td>
<td>US</td>
<td><a href="mailto:jeewha@yahoo.com">jeewha@yahoo.com</a></td>
</tr>
<tr>
<td>Kim</td>
<td>Hyun-Ju</td>
<td>University of Iowa</td>
<td>US</td>
<td><a href="mailto:byunjo-kin@uwo.edu">byunjo-kin@uwo.edu</a></td>
</tr>
<tr>
<td>Kobayashi</td>
<td>Miroko</td>
<td>Kanda University of International Studies</td>
<td>Japan</td>
<td><a href="mailto:m.koby@kanda.ac.jp">m.koby@kanda.ac.jp</a></td>
</tr>
<tr>
<td>Koizumi</td>
<td>Rie</td>
<td>University of Tsukuba</td>
<td>Japan</td>
<td><a href="mailto:kawakami@toyo.or.jp">kawakami@toyo.or.jp</a></td>
</tr>
<tr>
<td>Kumman</td>
<td>Antony</td>
<td>CSULA</td>
<td>US</td>
<td><a href="mailto:akawakami@unimelb.edu">akawakami@unimelb.edu</a></td>
</tr>
<tr>
<td>Lake</td>
<td>J</td>
<td>Fukuoka Jo Gakuen University</td>
<td>Japan</td>
<td><a href="mailto:jlake@jcom.home.ne.jp">jlake@jcom.home.ne.jp</a></td>
</tr>
<tr>
<td>Lazaraton</td>
<td>Anne</td>
<td>University of Minnesota</td>
<td>US</td>
<td><a href="mailto:laszrato@umn.edu">laszrato@umn.edu</a></td>
</tr>
<tr>
<td>Lee</td>
<td>Young-Ju</td>
<td>University of Illinois, Urbana-Champaign</td>
<td>US</td>
<td><a href="mailto:ylee@umn.edu">ylee@umn.edu</a></td>
</tr>
<tr>
<td>Lee</td>
<td>Jyoung</td>
<td>CSULA</td>
<td>US</td>
<td><a href="mailto:jean28@umnmail.net">jean28@umnmail.net</a></td>
</tr>
<tr>
<td>Lee</td>
<td>Jeong-Won</td>
<td>Chungam National University</td>
<td>South Korea</td>
<td><a href="mailto:jeongwon@enu.ac.kr">jeongwon@enu.ac.kr</a></td>
</tr>
<tr>
<td>Lee</td>
<td>Sue</td>
<td>CSULA</td>
<td>US</td>
<td><a href="mailto:sue2995@hotmail.com">sue2995@hotmail.com</a></td>
</tr>
<tr>
<td>Lennig</td>
<td>Matt</td>
<td>Ordinate Corp</td>
<td>US</td>
<td><a href="mailto:lennig@ordinate.com">lennig@ordinate.com</a></td>
</tr>
<tr>
<td>Leung</td>
<td>Constant</td>
<td>King's College, London University</td>
<td>UK</td>
<td><a href="mailto:constant.leung@kcl.ac.uk">constant.leung@kcl.ac.uk</a></td>
</tr>
<tr>
<td>Lewkowicz</td>
<td>Jo</td>
<td>University of Hong Kong</td>
<td>HK</td>
<td><a href="mailto:joelbow@hkoum.bus.br.hk">joelbow@hkoum.bus.br.hk</a></td>
</tr>
<tr>
<td>Liao</td>
<td>Yen-Fen</td>
<td>Teachers College, Columbia University</td>
<td>US</td>
<td><a href="mailto:yz2057@e-mail.com">yz2057@e-mail.com</a></td>
</tr>
<tr>
<td>Liu</td>
<td>Hsin-Min</td>
<td>Iowa State University</td>
<td>US</td>
<td><a href="mailto:samui_lu@live.com">samui_lu@live.com</a></td>
</tr>
</tbody>
</table>
# LTRC 2004 Participants List

<table>
<thead>
<tr>
<th>Last Name</th>
<th>First Name</th>
<th>Affiliation</th>
<th>Country</th>
<th>E-mail address</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iona</td>
<td>Lorena</td>
<td>UCLA/LAUSD</td>
<td>US</td>
<td><a href="mailto:lorenaa@ucla.edu">lorenaa@ucla.edu</a></td>
</tr>
<tr>
<td>Lendke</td>
<td>Zsuzsa</td>
<td>UCLA</td>
<td>US</td>
<td><a href="mailto:zsuzsa@ucla.edu">zsuzsa@ucla.edu</a></td>
</tr>
<tr>
<td>Lopriere</td>
<td>Lucilla</td>
<td>University of Gassino</td>
<td>Italy</td>
<td><a href="mailto:lucilla.lopriere@famweb.net.it">lucilla.lopriere@famweb.net.it</a></td>
</tr>
<tr>
<td>Luce</td>
<td>Meggan</td>
<td>Ballard &amp; Tighe</td>
<td>US</td>
<td><a href="mailto:nluce@ballard-tighe.com">nluce@ballard-tighe.com</a></td>
</tr>
<tr>
<td>Looma</td>
<td>Sari</td>
<td>Ballard &amp; Tighe</td>
<td>US</td>
<td><a href="mailto:sloomas@ballard-tighe.com">sloomas@ballard-tighe.com</a></td>
</tr>
<tr>
<td>Masumujavid</td>
<td>Undrmasa</td>
<td>CSULA</td>
<td>US</td>
<td><a href="mailto:umasumujavid@calstatela.edu">umasumujavid@calstatela.edu</a></td>
</tr>
<tr>
<td>Machtelti</td>
<td>Sabrina</td>
<td>Università per Stranierti, Siena</td>
<td>Italy</td>
<td><a href="mailto:smachtet@email.it">smachtet@email.it</a></td>
</tr>
<tr>
<td>Marsden</td>
<td>Nicola</td>
<td>Higher Colleges of Technology</td>
<td>UAE</td>
<td><a href="mailto:nicola.marsden@hct.ac.ae">nicola.marsden@hct.ac.ae</a></td>
</tr>
<tr>
<td>Mckay</td>
<td>Penny</td>
<td>Queensland University of Technology</td>
<td>Australia</td>
<td><a href="mailto:pa.mckay@qut.edu.au">pa.mckay@qut.edu.au</a></td>
</tr>
<tr>
<td>Meiron</td>
<td>Beryl</td>
<td>IELTS Inc.</td>
<td>US</td>
<td><a href="mailto:bmeiron@cois.org">bmeiron@cois.org</a></td>
</tr>
<tr>
<td>Min</td>
<td>Hoky</td>
<td>UCLA</td>
<td>US</td>
<td><a href="mailto:hoky@ucla.edu">hoky@ucla.edu</a></td>
</tr>
<tr>
<td>Mitsubishi</td>
<td>Mineo</td>
<td>Institute for International Business Communication, Tokyo</td>
<td>Japan</td>
<td><a href="mailto:m-mineo@ies-tky.co.jp">m-mineo@ies-tky.co.jp</a></td>
</tr>
<tr>
<td>Mouna</td>
<td>Nawal</td>
<td>Canadian Foreign Service</td>
<td>Canada</td>
<td><a href="mailto:nawal.mouna@dfnti-macel.gc.ca">nawal.mouna@dfnti-macel.gc.ca</a></td>
</tr>
<tr>
<td>Moya</td>
<td>Robert</td>
<td>CSULA</td>
<td>US</td>
<td><a href="mailto:betorb@angelfire.com">betorb@angelfire.com</a></td>
</tr>
<tr>
<td>Naidu</td>
<td>Shashi</td>
<td>Oklahoma State University</td>
<td>US</td>
<td><a href="mailto:distanceand@hotmail.com">distanceand@hotmail.com</a></td>
</tr>
<tr>
<td>Nakamura</td>
<td>Yuji</td>
<td>Tokyo Keizai University</td>
<td>Japan</td>
<td><a href="mailto:nyky97@keio.ac.jp">nyky97@keio.ac.jp</a></td>
</tr>
<tr>
<td>Okley</td>
<td>Gary</td>
<td>UCLA</td>
<td>US</td>
<td><a href="mailto:gockey@humnet.ucla.edu">gockey@humnet.ucla.edu</a></td>
</tr>
<tr>
<td>Ormsby</td>
<td>Harold</td>
<td>Universidad Nacional Autónoma de Mexico</td>
<td>Mexico</td>
<td><a href="mailto:ormsby@servidor.uxam.mx">ormsby@servidor.uxam.mx</a></td>
</tr>
<tr>
<td>Palazuelos</td>
<td>Carmen Marques</td>
<td>Universidad Autónoma de Baja California</td>
<td>Mexico</td>
<td><a href="mailto:carmen@yahoo.com">carmen@yahoo.com</a></td>
</tr>
<tr>
<td>Palmer</td>
<td>Adrian</td>
<td>University of Utah</td>
<td>US</td>
<td><a href="mailto:apalmer@linguistics.utah.edu">apalmer@linguistics.utah.edu</a></td>
</tr>
<tr>
<td>Pantuja</td>
<td>Laura</td>
<td>CSULA</td>
<td>US</td>
<td><a href="mailto:lara_pantuja@yahoo.com">lara_pantuja@yahoo.com</a></td>
</tr>
<tr>
<td>Park</td>
<td>Taejoon</td>
<td>Teachers College, Columbia University</td>
<td>US</td>
<td><a href="mailto:taejoon@columbia.edu">taejoon@columbia.edu</a></td>
</tr>
<tr>
<td>Park</td>
<td>Eun Kyeeong</td>
<td>Teachers College, Columbia University</td>
<td>US</td>
<td><a href="mailto:ekp27@columbia.edu">ekp27@columbia.edu</a></td>
</tr>
<tr>
<td>Pantvink</td>
<td>Nandini</td>
<td>CSULA</td>
<td>US</td>
<td><a href="mailto:potnainandini@hotmail.com">potnainandini@hotmail.com</a></td>
</tr>
<tr>
<td>Perrone</td>
<td>Michael</td>
<td>Teachers College, Columbia University</td>
<td>US</td>
<td><a href="mailto:mjppp@sc.edu">mjppp@sc.edu</a></td>
</tr>
<tr>
<td>Piskunas</td>
<td>Lina</td>
<td>University of Iowa</td>
<td>US</td>
<td><a href="mailto:lisa_piskunas@edu.wa.edu">lisa_piskunas@edu.wa.edu</a></td>
</tr>
<tr>
<td>Prasad</td>
<td>Nishita</td>
<td>University of Hong Kong</td>
<td>HK</td>
<td>nishita@<a href="mailto:prasad@hotmail.com">prasad@hotmail.com</a></td>
</tr>
<tr>
<td>Purpora</td>
<td>James</td>
<td>Teachers College, Columbia University</td>
<td>US</td>
<td><a href="mailto:jj248@columbia.edu">jj248@columbia.edu</a></td>
</tr>
<tr>
<td>Ragavan</td>
<td>Samanthi</td>
<td>University of Guam</td>
<td>New Zealand</td>
<td><a href="mailto:johns_ragavan@vw.ac.nz">johns_ragavan@vw.ac.nz</a></td>
</tr>
<tr>
<td>Reck</td>
<td>John</td>
<td>Victoria University of Wellington</td>
<td>UK</td>
<td><a href="mailto:language@luva.net">language@luva.net</a></td>
</tr>
<tr>
<td>Rees-Dickins</td>
<td>Pauline</td>
<td>University of Bristol</td>
<td>US</td>
<td><a href="mailto:mreves@ets.org">mreves@ets.org</a></td>
</tr>
<tr>
<td>Reed</td>
<td>Dan</td>
<td>Second Language Testing Inc.</td>
<td>US</td>
<td><a href="mailto:trichter@ets.org">trichter@ets.org</a></td>
</tr>
<tr>
<td>Reesva</td>
<td>Marcia</td>
<td>ETS</td>
<td>US</td>
<td><a href="mailto:dirzx@uog.edu">dirzx@uog.edu</a></td>
</tr>
<tr>
<td>Richter</td>
<td>Theresa</td>
<td>ETS</td>
<td>US</td>
<td><a href="mailto:joanne_rzr@hotmail.com">joanne_rzr@hotmail.com</a></td>
</tr>
<tr>
<td>Robertson</td>
<td>Daniel</td>
<td>University of Guam</td>
<td>US</td>
<td><a href="mailto:karen.sacre@laidland.afa.mil">karen.sacre@laidland.afa.mil</a></td>
</tr>
<tr>
<td>Ryu</td>
<td>Jeong Ran</td>
<td>Brigham Young University</td>
<td>Japan</td>
<td><a href="mailto:c-aiida@mvl.biglobe.ne.jp">c-aiida@mvl.biglobe.ne.jp</a></td>
</tr>
<tr>
<td>Sacre</td>
<td>Karen Ann</td>
<td>Defense Language Institute</td>
<td>US</td>
<td><a href="mailto:yosh84@columbia.edu">yosh84@columbia.edu</a></td>
</tr>
<tr>
<td>Sato</td>
<td>Chisato</td>
<td>Nanaki Senior High School</td>
<td>Japan</td>
<td><a href="mailto:satoh@shokoku.ac.jp">satoh@shokoku.ac.jp</a></td>
</tr>
<tr>
<td>Sato</td>
<td>Hideaki</td>
<td>Teachers College, Columbia University</td>
<td>Japan</td>
<td><a href="mailto:suelipa@nec.ces.s.u.u.k.o.uk">suelipa@nec.ces.s.u.u.k.o.uk</a></td>
</tr>
<tr>
<td>Saville</td>
<td>Nick</td>
<td>UCLIES</td>
<td>UK</td>
<td><a href="mailto:yawatsi@ets.org">yawatsi@ets.org</a></td>
</tr>
<tr>
<td>Sawaki</td>
<td>Yanoa</td>
<td>ETS</td>
<td>Brazil</td>
<td><a href="mailto:matilde@ic.inilisacamp.br">matilde@ic.inilisacamp.br</a></td>
</tr>
<tr>
<td>Scaramucci</td>
<td>Malitde</td>
<td>State University of Campinas</td>
<td>The Netherlands</td>
<td><a href="mailto:rob.schoonenma@uva.nl">rob.schoonenma@uva.nl</a></td>
</tr>
<tr>
<td>Schoonen</td>
<td>Rob</td>
<td>University of Amsterdam</td>
<td>US</td>
<td><a href="mailto:scott@neurol.org">scott@neurol.org</a></td>
</tr>
<tr>
<td>Scott</td>
<td>Gladys</td>
<td>Northwest Regional Educational Lab</td>
<td>US</td>
<td><a href="mailto:gladys@ballard-tighe.com">gladys@ballard-tighe.com</a></td>
</tr>
<tr>
<td>Seresia</td>
<td>George</td>
<td>Ballard &amp; Tighe</td>
<td>US</td>
<td><a href="mailto:annasbhin@ucla.edu">annasbhin@ucla.edu</a></td>
</tr>
<tr>
<td>Shin</td>
<td>Annalise</td>
<td>UCLA</td>
<td>US</td>
<td><a href="mailto:sangshin@ucla.edu">sangshin@ucla.edu</a></td>
</tr>
<tr>
<td>Shin</td>
<td>Sapt-Kenai</td>
<td>UCLA</td>
<td>US</td>
<td><a href="mailto:sung74@ucla.edu">sung74@ucla.edu</a></td>
</tr>
<tr>
<td>Last Name</td>
<td>First Name</td>
<td>Affiliation</td>
<td>Country</td>
<td>E-mail address</td>
</tr>
<tr>
<td>-----------</td>
<td>------------</td>
<td>--------------------------------------------------</td>
<td>---------</td>
<td>-------------------------------------</td>
</tr>
<tr>
<td>Shiotsu</td>
<td>Yoshikiko</td>
<td>Kurume University</td>
<td>Japan</td>
<td><a href="mailto:ts@jp.org">ts@jp.org</a></td>
</tr>
<tr>
<td>Slovakek</td>
<td>Simson</td>
<td>CSULA</td>
<td>US</td>
<td><a href="mailto:aslovak3@calstatela.edu">aslovak3@calstatela.edu</a></td>
</tr>
<tr>
<td>Song</td>
<td>Xiaomei</td>
<td>Queen’s University</td>
<td>Canada</td>
<td><a href="mailto:qas@link.queensu.ca">qas@link.queensu.ca</a></td>
</tr>
<tr>
<td>Song</td>
<td>Minkyung</td>
<td>UCLA</td>
<td>US</td>
<td><a href="mailto:songmyu@ucla.edu">songmyu@ucla.edu</a></td>
</tr>
<tr>
<td>Spencer</td>
<td>Mary</td>
<td>University of Guam</td>
<td>US</td>
<td><a href="mailto:mapencer@uogu.uog.edu">mapencer@uogu.uog.edu</a></td>
</tr>
<tr>
<td>Stanford</td>
<td>Charles W.</td>
<td>Second Language Testing Inc.</td>
<td>US</td>
<td><a href="mailto:cstangfield@slti.com">cstangfield@slti.com</a></td>
</tr>
<tr>
<td>Stoneman</td>
<td>Bernadette</td>
<td>Hong Kong Polytechnic University</td>
<td>HK</td>
<td><a href="mailto:egstoneman@polyu.edu.hk">egstoneman@polyu.edu.hk</a></td>
</tr>
<tr>
<td>Strong-Krause</td>
<td>Diane</td>
<td>Brigham Young University</td>
<td>US</td>
<td><a href="mailto:dsgy@email.byu.edu">dsgy@email.byu.edu</a></td>
</tr>
<tr>
<td>Tannenbaum</td>
<td>Deborah</td>
<td>CSULA</td>
<td>US</td>
<td><a href="mailto:dobuchal@glendale.edu">dobuchal@glendale.edu</a></td>
</tr>
<tr>
<td>Taskesen</td>
<td>Cynthia</td>
<td>DOJ/FBI</td>
<td>US</td>
<td>n0-e-mail-address</td>
</tr>
<tr>
<td>Taylor</td>
<td>Lynda</td>
<td>UCLES</td>
<td>UK</td>
<td><a href="mailto:taylor.l@ucles.org.uk">taylor.l@ucles.org.uk</a></td>
</tr>
<tr>
<td>Taylor</td>
<td>Linda</td>
<td>CASAS</td>
<td>US</td>
<td><a href="mailto:taylor@casas.org">taylor@casas.org</a></td>
</tr>
<tr>
<td>Tanke</td>
<td>Margaret</td>
<td>Mt. San Antonio College</td>
<td>US</td>
<td><a href="mailto:mtsako@mtaer.edu">mtsako@mtaer.edu</a></td>
</tr>
<tr>
<td>Thomas</td>
<td>Jillanne</td>
<td>CSUF</td>
<td>US</td>
<td><a href="mailto:jillanne@student.fullerton.edu">jillanne@student.fullerton.edu</a></td>
</tr>
<tr>
<td>Thrasher</td>
<td>Randy</td>
<td>Okinawa Christian University</td>
<td>Japan</td>
<td><a href="mailto:thrasher@ocjc.ac.jp">thrasher@ocjc.ac.jp</a></td>
</tr>
<tr>
<td>Tintos-Rose</td>
<td>Kathy</td>
<td>CPS Human Resource Services</td>
<td>US</td>
<td><a href="mailto:kathy@cpa.ca.gov">kathy@cpa.ca.gov</a></td>
</tr>
<tr>
<td>Tsai</td>
<td>Ya-Lin Olivia</td>
<td>UCLA</td>
<td>US</td>
<td><a href="mailto:olivietai@msn.com.tw">olivietai@msn.com.tw</a></td>
</tr>
<tr>
<td>Turner</td>
<td>Carolyn</td>
<td>McGill University</td>
<td>Canada</td>
<td><a href="mailto:carolyn.turner@mcgill.ca">carolyn.turner@mcgill.ca</a></td>
</tr>
<tr>
<td>Ureno</td>
<td>Salvador</td>
<td>CSULA</td>
<td>US</td>
<td><a href="mailto:salurano@aol.com">salurano@aol.com</a></td>
</tr>
<tr>
<td>van Deemter</td>
<td>R</td>
<td>Leiden University</td>
<td>The Netherlands</td>
<td><a href="mailto:r.vandeemter@st.leidenuniv.nl">r.vandeemter@st.leidenuniv.nl</a></td>
</tr>
<tr>
<td>van Moore</td>
<td>Alistair</td>
<td>Kanda University</td>
<td>Japan</td>
<td><a href="mailto:almoore@hotmail.com">almoore@hotmail.com</a></td>
</tr>
<tr>
<td>Vongruminitch</td>
<td>Vipheen</td>
<td>UCLA</td>
<td>US</td>
<td><a href="mailto:vongrumen@ucla.edu">vongrumen@ucla.edu</a></td>
</tr>
<tr>
<td>Wagner</td>
<td>Mathew Elvis</td>
<td>Teachers College, Columbia University</td>
<td>US</td>
<td><a href="mailto:waynemaid@yahoo.com">waynemaid@yahoo.com</a></td>
</tr>
<tr>
<td>Weigle</td>
<td>Sara Cushing</td>
<td>Georgia State University</td>
<td>US</td>
<td><a href="mailto:swegle@gasu.edu">swegle@gasu.edu</a></td>
</tr>
<tr>
<td>Wolch</td>
<td>Kevin</td>
<td>CSULA</td>
<td>US</td>
<td><a href="mailto:kpwoehoe@charter.net">kpwoehoe@charter.net</a></td>
</tr>
<tr>
<td>White</td>
<td>Jeffrey A.</td>
<td>Los Angeles Unified School District</td>
<td>US</td>
<td><a href="mailto:jeff.white@lausd.net">jeff.white@lausd.net</a></td>
</tr>
<tr>
<td>Wiernman</td>
<td>Cynthia</td>
<td>Borough of Manhattan Community College, City University of New York</td>
<td>US</td>
<td><a href="mailto:engaspoulos@aol.com">engaspoulos@aol.com</a></td>
</tr>
<tr>
<td>Woo</td>
<td>Hui-Jeong</td>
<td>University of Illinois, Urbana-Champaign</td>
<td>US</td>
<td><a href="mailto:hwoo@uiuc.edu">hwoo@uiuc.edu</a></td>
</tr>
<tr>
<td>Xi</td>
<td>Xiaoming</td>
<td>RTS</td>
<td>US</td>
<td><a href="mailto:xxi@ets.org">xxi@ets.org</a></td>
</tr>
<tr>
<td>Yamashiro</td>
<td>Amy</td>
<td>University of Michigan</td>
<td>US</td>
<td>no e-mail address</td>
</tr>
<tr>
<td>Yoshizawa</td>
<td>Kiyomi</td>
<td>Kanazawa University</td>
<td>Japan</td>
<td><a href="mailto:yoshizawa@ipku.kanazawa-u.ac.jp">yoshizawa@ipku.kanazawa-u.ac.jp</a></td>
</tr>
<tr>
<td>Yu</td>
<td>Jing</td>
<td>University of Michigan</td>
<td>US</td>
<td><a href="mailto:yjingu@masu.edu">yjingu@masu.edu</a></td>
</tr>
<tr>
<td>Yun</td>
<td>Youngmi</td>
<td>University of Illinois, Urbana-Champaign</td>
<td>US</td>
<td><a href="mailto:yunyuan@uiuc.edu">yunyuan@uiuc.edu</a></td>
</tr>
<tr>
<td>Zhou</td>
<td>Yukia</td>
<td>Tokyo University of Foreign Studies</td>
<td>Japan</td>
<td><a href="mailto:zhousyuja@bmtmail.com">zhousyuja@bmtmail.com</a></td>
</tr>
</tbody>
</table>
Second Language Research
Language Teaching Research
Child Language Teaching & Therapy
Language Testing

These journals have brought together a wide range of topics of interest to all those involved in linguistics research, stimulating debate as well as providing a vital reference source for busy academics. Published regularly throughout the year to ensure you are kept up to date, these journals provide authoritative reviews of all aspects of linguistics.

www.arnoldpublishers.com/journals

Try our journals for FREE! - contact us today at: ejournal@hodder.co.uk or 020 7873 6358

ARNOLD
Research Notes

The quarterly publication reporting on matters relating to research, test development and validation within Cambridge ESOL

If you are interested or involved in English language assessment, Research Notes is an invaluable resource which gives you up-to-date information to support your teaching, teacher training or research activity.

Research Notes 15
Focus on specialised exams

- Test comparability
- Analysing domain-specific lexical categories
- IELTS Writing: revising assessment criteria and scales
- An IELTS Impact Study
- The YLE Review
- Creating a virtual community of assessment practice
- Reliability in FCE objective papers
- 2003 IELTS Master’s Award winner

New

You can now find articles online that match your particular area of interest, searching by:
- exam (KET, CAE)
- skill (Listening, Speaking)
- area (Candidates, Research Methods)

To find articles, subscribe or download back issues of Research Notes, please visit the website.

Research Notes 14
Focus on Teaching Awards

- Cambridge ESOL Teaching Awards: current perspectives, future trends
- The Distance DELTA
- The teacher educator as reflective practitioner
- Language awareness and assessment
- In-service language teaching in Brazil using CELT
- Teacher support
- Interaction in the paired FCE speaking test

www.CambridgeESOL.org/researchnotes
Next generation TOEFL®
Focus on communication

Now institutions can feel more confident in their students’ ability to use language, while students gain the confidence they need to succeed.

Next Generation TOEFL
More Than Just a Test:
• Assesses reading, listening, writing, and speaking
• Integrated tasks in all four skills more accurately measure students’ ability to use language
• Content simulates actual university experiences
• Score reports provide detailed information on the test taker’s ability
• New teaching and learning tools support English language learning

Next Generation TOEFL Timeline:

JULY 2004 – LEARN AND PRACTICE ONLINE
Students and teachers can experience a practice version of the test on the Web.

SEPTEMBER 2005 – WORLD PREMIERE!
The next generation TOEFL test will be administered at official ETS test centers for admissions purposes. The test will be given under secure conditions and scores will be reported directly to institutions.

To find out more, visit www.ets.org/toefl23.html or contact:

In the USA:
Phone: 1-800-241-5393
e-mail: efl@ets.org

Outside the USA:
Phone: 1-609-693-3232
Fax: 1-609-693-2366
e-mail: internationaldevelopment@ets.org

ETS TOEFL®

Copyright © 2004 by Educational Testing Service. All rights reserved. ETS, the ETS logo and TOEFL are registered trademarks of Educational Testing Service.

Educational Testing Service  M204