Program

LTRC 2000

22nd Annual
Language Testing Research
Colloquium

THEME: "Interdisciplinary Interfaces with Language Testing"

Empire Landmark Hotel
Vancouver, British Columbia, Canada

March 8-11, 2000
ACKNOWLEDGEMENTS

LTRC-2000 Organizing Committee

Lyle Bachman, LTRC 2000 Co-Chair
Bonny Norton, LTRC 2000 Co-Chair
Adrian Palmer, LTRC 2000 Co-Chair

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

Cambridge University Press
Department of Language and Literacy Education, University of British Columbia
English Language Institute, University of British Columbia
International Language Testing Association
Teachers College, Columbia University
TOEFL Policy Council, Educational Testing Service
University of Cambridge Local Examinations Syndicate

We are also grateful to the following people who served as proposal reviewers:

Charles Alderson
Micheline Chalhoub-Deville
Caroline Clapham
Alister Cumming
Fred Davidson
Dan Douglas
Jana Fox
Glenn Fulcher
Liz Hamp-Lyons
Ari Huhta
Antony Kunnan
Sari Luoma
Brian Lynch
Jim Purpura
John Reid
Steve Ross
Elana Shohamy
Charlie Stansfield
Carolyn Turner
ACKNOWLEDGEMENTS

We also wish to thank the following organizations for participating in the book exhibit:

Cambridge University Press
Educational Testing Service (ETS)
NCELTR, Macquarie University
Nelson, Thomson Learning
Oxford University Press
Pearson Education ESL (USA)
Pearson Educational (Canada)
University of Cambridge Local Examinations Syndicate (UCLES)
University of Melbourne Language Testing Research Centre

We would like to thank the people listed below, who assisted in various ways with the organization of the conference and the many tasks involved in making LTRC 2000 happen:

Nathan Carr University of California, Los Angeles
Glenn Fulcher University of Surrey
Jee Wha Kim Columbia University
James Purpura Columbia University
Rowena Tate Conference Services, University of British Columbia
Matthew Joseph Wagner Columbia University
MONDAY, MARCH 6
LTRC Workshop 1
Advances and New Developments in G-Theory: Applications of Structural Equation Modeling
Dr. George A. Marcoulides California State University, Fullerton
Location: Education Computing Services, Scarfe Building
University of British Columbia

<table>
<thead>
<tr>
<th></th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 - 12:00</td>
<td>Workshop (by preregistration only)</td>
<td>Scarfe Bldg. Room 1006</td>
</tr>
<tr>
<td>12:00- 1:00</td>
<td>LUNCH</td>
<td></td>
</tr>
<tr>
<td>1:00 - 5:00</td>
<td>Workshop continued</td>
<td>Scarfe Bldg. Room 1006</td>
</tr>
</tbody>
</table>

TUESDAY, MARCH 7
LTRC Workshop 2
Applications of Multidimensional Item Response Theory
Dr. Terry Ackerman University of North Carolina, Greensboro
Location: Education Computing Services, Scarfe Building
University of British Columbia

<table>
<thead>
<tr>
<th></th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 - 12:00</td>
<td>Workshop (by preregistration only)</td>
<td>Scarfe Bldg. Room 1101</td>
</tr>
<tr>
<td>12:00- 1:00</td>
<td>LUNCH</td>
<td></td>
</tr>
<tr>
<td>1:00 - 5:00</td>
<td>Workshop continued</td>
<td>Scarfe Bldg. Room 1101</td>
</tr>
<tr>
<td>5:00 - 6:30</td>
<td>LTRC Registration: Empire Landmark Hotel</td>
<td>Outside the Pavilion Room</td>
</tr>
<tr>
<td>6:30 - 8:00</td>
<td>Welcoming Reception co-hosted by the English Language Institute and the Department of Language and Literacy Education, University of British Columbia</td>
<td>The Pavilion Room</td>
</tr>
</tbody>
</table>

3
WEDNESDAY, March 8

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 a.m. - 4:00 p.m.</td>
<td>Registration</td>
<td>Outside Room 520/522</td>
</tr>
<tr>
<td>8:45 - 10:15</td>
<td>Opening Session</td>
<td>Room 520/522</td>
</tr>
<tr>
<td></td>
<td>Welcome: Bonny Norton University of British Columbia</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Introduction of Keynote Speaker: Lyle Bachman University of California, Los Angeles</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Keynote Address: Examining Dialogue: Another Approach to Validating Inferences Drawn from Test Scores? Merrill Swain Ontario Institute for Studies in Education, University of Toronto</td>
<td></td>
</tr>
<tr>
<td>10:15 - 10:45</td>
<td>Break</td>
<td></td>
</tr>
<tr>
<td>10:45 - 12:15</td>
<td>Paper Session 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chair: Helmut Vollmer University of Osnabrück</td>
<td></td>
</tr>
</tbody>
</table>

Setting Standards Where There Are None
J Charles Alderson The British Council, Budapest, Hungary

A New Standard Setting Method for Multiple Cut-Off Scores
Fellianka Kaftandjieva University of Jyväskylä
Norman Verhelst National Institute for Educational Measurement, The Netherlands

Diagnosis before Cure: The Interface between Language Testers and Medical Educators
Elisabeth Grove and Annie Brown University of Melbourne

12:15 - 1:30 Lunch

Cnc: Council of Europe Terms will be the currency of the future
[We need to do ECCE in terms of this]
WEDNESDAY, March 8

1:30 - 3:30  
**Symposium 1: The National College English Test**  
Organizer: Yang Huizhong *National College English Testing Committee of China and Shanghai Jiao Tong University*

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Location</th>
</tr>
</thead>
</table>
| 1:30 - 3:30 | **The Validation Study of the National College English Test**  
Yang Huizhong *National College English Testing Committee of China and Shanghai Jiao Tong University* | The Pavilion Room         |
| 1:30 - 3:30 | **An Introspection Study of the Validity of the CET Reading Comprehension Test**  
Jin Yan *National College English Testing Committee of China and Shanghai Jiao Tong University* | The Pavilion Room         |
| 1:30 - 3:30 | **Towards Improving Marking Reliability in the CET Test**  
Yang Huizhong *National College English Testing Committee of China and Shanghai Jiao Tong University*  
Xia Guozuo *National College English Testing Committee of China and Fu Dan University* | The Pavilion Room         |
| 1:30 - 3:30 | **The CET-SET Test—Testing English Speaking Abilities**  
Yang Huizhong *National College English Testing Committee of China and Shanghai Jiao Tong University*  
Guo Jieke *National College English Testing Committee of China and South-China University of Science and Technology* | The Pavilion Room         |
| 3:30 - 3:45 | Break                                                                  | The Pavilion Room         |
| 3:45 - 5:15 | **Paper Session 2**  
Chair: Glayol Ekbatani *St John's University* | The Pavilion Room         |
| 3:45 - 5:15 | **Ethical Dilemmas in Language Testing: What Can We Actually Do?**  
Liz Hamp-Lyons and Tom Lumley *Hong Kong Polytechnic University* | The Pavilion Room         |
| 3:45 - 5:15 | **Content and Context Visuals and Performance on TOEFL CBT Listening Comprehension Stimuli**  
April Ginther *Purdue University* | The Pavilion Room         |
| 3:45 - 5:15 | **The Role of Digital Video Media in Response to Task Demands**  
Paul Gruba *The University of Melbourne* | The Pavilion Room         |
| 5:30 - 7:30 | **Reception and Launch of Cambridge Language Assessment Series, hosted by Cambridge University Press** | The Pavilion Room         |

5
<table>
<thead>
<tr>
<th>Time</th>
<th>Event Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 a.m. - 1:30 p.m.</td>
<td>Registration</td>
</tr>
<tr>
<td>9:00 - 10:30</td>
<td>Paper Session 3</td>
</tr>
<tr>
<td></td>
<td>Chair: Hossein Farhady Iran University of Science and Technology</td>
</tr>
<tr>
<td></td>
<td>Web-Based Language Testing: Opportunities and Challenges</td>
</tr>
<tr>
<td></td>
<td>Carsten Roever University of Hawai’i at Manoa</td>
</tr>
<tr>
<td></td>
<td>Detecting DIF (Differential Item Functioning) in a Speaking Test with Polytomous Response Data</td>
</tr>
<tr>
<td></td>
<td>Minkyung Kim University of California, Los Angeles</td>
</tr>
<tr>
<td>10:30 - 10:50</td>
<td>Break</td>
</tr>
<tr>
<td>10:50 - 12:10</td>
<td>Paper Session 4</td>
</tr>
<tr>
<td></td>
<td>Chair: Hossein Farhady Iran University of Science and Technology</td>
</tr>
<tr>
<td></td>
<td>Assessment Criteria in a Large-Scale Writing Test: What Do They Really Mean to the Raters?</td>
</tr>
<tr>
<td></td>
<td>Tom Lumley Hong Kong Polytechnic University</td>
</tr>
<tr>
<td></td>
<td>Assessment Criteria for NNS and NS Essays: Do uniform standards work?</td>
</tr>
<tr>
<td></td>
<td>Sally O'Hagan The University of Melbourne</td>
</tr>
<tr>
<td></td>
<td>Native and Nonnative Speaking EFL Teachers' Evaluation of Chinese Students' Writing</td>
</tr>
<tr>
<td></td>
<td>Ling Shi University of British Columbia</td>
</tr>
<tr>
<td>12:10 - 1:30</td>
<td>Lunch</td>
</tr>
<tr>
<td>12:00 - 6:00</td>
<td>Book Exhibits</td>
</tr>
<tr>
<td></td>
<td>Room 526</td>
</tr>
</tbody>
</table>
THURSDAY, MARCH 9

1:30 - 3:30  Symposium 2: Putting Tasks to the Test
Organizer: Dr. Cathie Elder University of Melbourne

Presenters:
Noriko Iwashita (on behalf of Tim McNamara, Noriko Iwashita and Cathie Elder) University of Melbourne
Helen Slatyer (on behalf of Geoff Brindley, Gill Wigglesworth and Helen Slatyer) Macquarie University
John Norris (on behalf of J. D. Brown, John Norris and Tom Hudson University of Hawai‘i at Manoa

Respondents:
Prof. Lyle F. Bachman University of California, Los Angeles
Prof. Peter Skehan Thames Valley University
Prof. Peter Robinson Aoyama Gakuin University

3:30 - 4:00  Break

4:00 - 4:30  ILTA Presidential Address
Alan Davies University of Edinburgh

4:30 - 5:05  Work in Progress Plenary Session
Chair: Bonny Norton University of British Columbia

Developing Collaborative Assessment Tasks Across Disciplines
Vivien Berry University of Hong Kong

Issues in Video Listening Assessment
Christine Feak and Julia Salehzadeh University of Michigan, Ann Arbor

A New Generation of Language Testers Using Structural Equation Modeling
Shoichi Gregory Kamei University of California, Los Angeles
Amy D. Yamashiro Nihon University, Tokyo
Greta Gorsuch Texas Tech University, Lubbock
David Beglar Temple University Japan, Osaka
George A. Marcoulides California State University at Fullerton
THURSDAY, MARCH 9

Developing an Educational Assessment Framework
Jo Lewkowicz University of Hong Kong

Test Takers' Performance on and Affective Responses to Paper-and-Pencil and Computer-Administered Japanese Reading Tests
Yasuyo Sawaki University of California, Los Angeles

A Cognitive-Psychometric Approach to Construct Validation of Web-Based Language Assessment
Lihshing Wang, Lyle Bachman, Nathan Carr, Greg Kamei, Mikyung Kim, Lorena Liosa, Yasuyo Sawaki, Sang Keun Shin, Viphavee Vongpumivit, Xiaoming Xi and David Yessis University of California, Los Angeles

Investigation of Size of Vocabulary and Lexical Richness in Candidate Output on a Semi-Direct Oral Proficiency Test
Jessica Wu The Language Training & Testing Center, Taipei

A Sociological Dimension in Tests of English for Immigrant Teachers
Elaine Wylie Griffith University

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Rooms</th>
</tr>
</thead>
<tbody>
<tr>
<td>5:05 - 6:00</td>
<td>Work in Progress Break-out Session</td>
<td>524 and 525</td>
</tr>
<tr>
<td>6:00 - 7:30</td>
<td>Dinner Break</td>
<td></td>
</tr>
<tr>
<td>7:30 - 8:30</td>
<td>Session on the Mentoring of Graduate Students in Language Testing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Moderator: Adrian Palmer University of Utah</td>
<td></td>
</tr>
</tbody>
</table>

Participants:
Charles Alderson and Jay Banerjee University of Lancaster
Lyle Bachman and Greg Kamei University of California, Los Angeles
J.D. Brown and John Norris University of Hawai‘i at Manoa
Fred Davidson and Yeonsuk Cho University of Illinois at Urbana-Champaign
Alan Davies and Ardeshr Geranpayeh University of Edinburgh
Tim McNamara and Paul Gruba University of Melbourne
FRIDAY, MARCH 10

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 a.m. - 11:00 a.m.</td>
<td>Registration</td>
<td>Outside Room 520/522</td>
</tr>
<tr>
<td>8:00 - 9:00</td>
<td>LTRC and ILTA Business Meetings</td>
<td></td>
</tr>
<tr>
<td>9:00 - 10:30</td>
<td>Paper Session 5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chair: Fred Davidson <strong>University of Illinois at Urbana-Champaign</strong></td>
<td></td>
</tr>
</tbody>
</table>

**Test Qualities: Expanding the Engineering Approach**
Antony John Kunnan **California State University, Los Angeles**

**Language Assessment: Inter, Itra, and Supra Disciplinary Interfaces**
Hossein Farhady **Iran University of Science and Technology**

**The Acquisition and Retention of ESL/Acculturation and Globalization Factors for Japanese Year Abroad University Students**
Stephen Carey **University of British Columbia**

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:30 - 10:50</td>
<td><strong>Break</strong></td>
<td></td>
</tr>
<tr>
<td>10:50 - 12:10</td>
<td>Paper Session 6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chair: Angela Hasselgran <strong>University of Bergen</strong></td>
<td></td>
</tr>
</tbody>
</table>

**Using Observation Checklists to Validate Speaking-Test Tasks**
Barry O'Sullivan **University of Reading**
Nick Saville **University of Cambridge Local Examinations Syndicate**

**Gender Bias in Oral Proficiency Testing**
Kieran O'Loughlin **University of Melbourne**

**Expanding the Construct of Reading Comprehension: Beyond Basic**
Mary McGroarty and Latrica Trites **Northern Arizona University**

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>12:10 - 1:30</td>
<td>Lunch</td>
<td></td>
</tr>
<tr>
<td>12:00 - 6:00</td>
<td><strong>Book Exhibits</strong></td>
<td>Room 526</td>
</tr>
</tbody>
</table>
FRIDAY, MARCH 10

1:30 - 3:30  Symposium 3: Context and Method in Washback Studies
Organizer: Liying Cheng University of Alberta

Presenters:
Nick Saville University of Cambridge Local Examinations Syndicate
Marvin Wideen Simon Fraser University
Carolyn E. Turner McGill University
Luxia Qi City University of Hong Kong

3:30 - 4:00  Break

4:00 - 5:00  Poster Plenary Session
Chair: Micheline Chalhoub-Deville University of Iowa

Developing a Web-based Language Placement Examination System
Lyle Bachman, Nathan Carr, Greg Kamei, Mkyung Kim, Lorena Llosa, Yasuyo Sawaki, Sang Keun Shin,
Sung-Ock Sohn, Viphavee Vongpumivitch, Liuhshing Wang, Xiaoming Xi and David Yessis University of
California, Los Angeles

A Self-Instructional Multimedia Rater Training Program
Helen S. Carpenter and Dorry M. Kenyon Center for Applied Linguistics

Assessing Integrated Academic Skills for International Undergraduates
Dan Douglas, Cynthia Myers, Volker Hegelheimer, Felicidade Van Acker
Iowa State University

History of the Baccalauréat
Samira Elatia University of Illinois at Urbana-Champaign

Reverse Engineering of Two Tests of French as a Second Language
Sandhya Gopalan and Samira Elatia University of Illinois at Urbana-Champaign

Border Crossings: What Multiple Proficiency Ratings Reveal
Gene Halleck Oklahoma State University
FRIDAY, MARCH 10

Multimedia Performance-Based Language Assessment: The Computerized Oral Proficiency Instrument (COPI)
Valerie A. Malabonga and Dorry Kenyon Center for Applied Linguistics

Assessment of Dimensionality of Language Test Data
Serena Pyo University of Illinois at Urbana-Champaign

The Journey toward Standardization: The University of British Columbia, English Language Institute Experience
Joy Lin Salzberg, Pat Burn, Pegram Deupree, Lori Hargreaves, Elaine Klein, Laurel Saba and Joyce White University of British Columbia

Software Demonstration for Creating Computer-Delivered/Human-Rated Speaking Exams
Diane Strong-Krause and Troy Cox Brigham Young University

From Rating to Assessment: The Rater Perspective on Assessing Writing in Finnish as a Second Language
Mirja Tarnanen University of Jyväskylä

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>5:00 - 5:45</td>
<td>Poster Break-out Session</td>
</tr>
<tr>
<td>5:45 - 6:00</td>
<td>Closing Remarks:</td>
</tr>
<tr>
<td></td>
<td>Lyle Bachman University of California, Los Angeles</td>
</tr>
<tr>
<td></td>
<td>Adrian Palmer University of Utah</td>
</tr>
<tr>
<td></td>
<td>Bonny Norton University of British Columbia</td>
</tr>
<tr>
<td>6:30</td>
<td>Buses depart hotel for Conference Banquet at Grouse Mountain</td>
</tr>
<tr>
<td>7:00</td>
<td>Banquet and Awards Presentations</td>
</tr>
<tr>
<td></td>
<td>Entertainment Produced and Directed by Ted Rodgers, aided and abetted by LFB, ASP and other co-conspirators</td>
</tr>
<tr>
<td></td>
<td>Wine compliments of UCLES</td>
</tr>
</tbody>
</table>
SATURDAY, MARCH 11

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:45 - 12:00</td>
<td>LTRC-AAAL PLENARY PANEL</td>
<td>Room 520/522</td>
</tr>
<tr>
<td></td>
<td>Interdisciplinary Interfaces with Language Testing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Moderator: Lyle Bachman LTRC Program Co-Chair</td>
<td></td>
</tr>
<tr>
<td>8:45 - 9:00</td>
<td>Introduction of Panelists</td>
<td></td>
</tr>
<tr>
<td>9:00 - 10:30</td>
<td>First Session</td>
<td></td>
</tr>
</tbody>
</table>

Panelists:
Diane Larsen-Freeman School for International Training
John Schumann University of California, Los Angeles
Larry Selinker University of London
Richard Young University of Wisconsin

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:30 - 10:45</td>
<td>Break</td>
</tr>
<tr>
<td>10:45 - 11:50</td>
<td>Second Session</td>
</tr>
</tbody>
</table>

Panelists:
Bernard Mohan University of British Columbia
David Nevo Tel Aviv University
Bruno Zumbo University of Northern British Columbia

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:50 - 12:00</td>
<td>Closing Remarks</td>
</tr>
<tr>
<td></td>
<td>Patricia L. Carrell AAAL 1st Vice President and AAAL Conference Chair</td>
</tr>
</tbody>
</table>
Abstracts

Setting Standards Where There Are None
J Charles Alderson
The British Council, Budapest, Hungary

Research Paper

Abstract

In Central Europe, education has undergone considerable upheaval since the change of political systems at the end of the 1980s. One of the most radical is the revision and reform of school-leaving examination systems. From school- and teacher-based subjective assessments conducted in a largely ritualistic fashion, with no explicit criteria for evaluation, standards to refer to, or quality control procedures of any sort, systems are now being developed that introduce modern ideas of testing and quality assurance, external quality control procedures and alignment, in theory, to common European standards.

In this context, it is well known that teachers' grades in final examinations vary enormously, from school to school, region to region and even within the same school. Indeed, a growing realisation that this variation is unacceptable in the New Europe is one of the driving forces for examination reform. Several governments have decreed that school-leaving examinations should be developed at at least two levels: Intermediate and Advanced, with no usable definition of what these terms might mean.

This paper will present research conducted as part of the reform of the Hungarian School-leaving English Examination in an attempt to define what these levels might mean, and what the consequences of such definitions might be. Both qualitative and quantitative research methodologies were utilised in experiments in standard-setting procedures. "Standardised" anchor items and IRT analyses were used to calibrate items and people, in order to map out the territory of ability and difficulty displayed in the results of large-scale piloting of experimental tasks intended for the new examinations. Expert judgements were also gathered in relation to the level definitions of the Common European Framework, the test specifications, and the presumed difficulty, for the school-leaving population, of the experimental tasks.

The results show the difficulty and dangers in seeking a priori to impose two levels on examinations intended to cover a highly heterogeneous population. The paper concludes with a discussion of the political consequences of the results, and recommendations for the resolution of the conflict between government requirements and empirical data. Both the methodologies used, and the implications of the results, have considerable relevance to the definition of levels of performance, especially where no such definitions have previously existed, or where they have simply been assumed and not been questioned.

J Charles Alderson
c/o The English Language Unit
The British Council
Benczur utca 26
1068 Budapest
Hungary
Fax: (361) 342 5728 or 352 8779
c.alderson@lancaster.ac.uk or alderson@isis.elte.hu
Developing a Web-based Language Placement Examination System

Lyle Bachman, Nathan Carr, Greg Kamei, Mikyung Kim, Lorena Llosa, Yasuyo Sawaki, Sang Keun Shin, Sung-Ock Sohn, Viphavee Vongpumivitch, Lisheng Wang, Xiaoming Xi and David Yessis

University of California, Los Angeles

Poster Session

Abstract

Accurately assessing and diagnosing the language ability of students for the purpose of placing them into appropriate language courses is an essential component of many foreign language programs at tertiary institutions. Recent advances in the technologies of language test design and development, web-based delivery of interactive multimedia, multiformat materials, and item response theory have created the possibility of developing language tests that not only make more efficient and effective use of resources, but also facilitate the engagement of test takers in a greater variety of potentially more authentic language use tasks than is possible with paper-and-pencil formats.

This poster session will describe the design and development of a web-based language placement examination system for three language programs—ESL, Japanese and Korean—at a university in the United States. The design of this system is being driven by considerations such as test purpose, construct definition, task specifications, test taker characteristics, overall usefulness, and resources that will be required and that are available for the system. While specific features vary considerably across the three language programs, we are developing a system that is flexible enough to accommodate the needs of each. The objectives of the project are to:

- design and develop a web-based system for 1) authoring, editing, storing and retrieving test tasks, 2) administering language placement examinations on-line, and 3) managing examination data processing, analysis and reporting, that will be flexible enough to meet the needs of all the language programs at the university;
- design and develop placement examinations for ESL, Japanese and Korean that meet accepted measurement standards.

This project is particularly challenging because it involves the following:

1. the integration of learning, pedagogical, curricular and assessment considerations;
2. the integration of content considerations with the potential and limitations of web-based and IRT technologies;
3. four general types of language use tasks (listening, speaking, reading and writing); and
4. three different languages with three widely differing orthographies.

Lyle F. Bachman
Dept. of Applied Linguistics & TESL
UCLA
3300 Rolfe Hall
405 Hilgard Avenue
P. O. Box 951531
Los Angeles, CA 90095-1531
Fax: 310 206-4118
bachman@ucla.edu

LYLE F. BACHMAN is Professor of Applied Linguistics and TESL at the University of California, Los Angeles. His current professional interests include the development and validation of tests of language ability, the technology of test design and development, and the interfaces between language testing research and other areas of applied linguistics research. His publications include Fundamental Considerations in Language Testing, Language Testing in Practice (with Adrian S. Palmer), and Interfaces
between Second Language Acquisition and Language Testing Research (co-edited with Andrew D. Cohen.) He has also published numerous articles in the area of language testing and evaluation. He is currently editor of the journal Language Testing, and co-editor of the Cambridge Language Assessment Series.

SUNG-OCK SOHN is Associate Professor of East Asian Languages and Cultures and the Korean Language Program Coordinator at the University of California, Los Angeles. Her current research interests include Korean linguistics and heritage language issues.

LIH-SHING WANG is Professor of Psychology at National Chung Cheng University, Taiwan, and currently Visiting Professor in the Departments of Applied Linguistics & TESL and Psychology at the University of California, Los Angeles. Her current research interests are in the areas of cognitive-psychometric modeling, computer adaptive assessment, and intelligent assessment.

NATHAN CARR, GREG KAMEI, MIKYUNG KIM, LORENA LLOSA, YASUYO SAWAKI, SANG KEUN SHIN, VIPHAVEE VONGPUMIVITCH, XIAOMING XI, and DAVID YESSIS are all graduate students in Applied Linguistics, and members of the Language Assessment Laboratory, Department of Applied Linguistics & TESL, University of California, Los Angeles.
Developing Collaborative Assessment Tasks Across Disciplines
Vivien Berry
The English Centre, University of Hong Kong

Work in Progress

Abstract

In this session I will discuss the development of an assessment system consisting of different task types employing text-based, audio-visual and IT channels of delivery. These are being designed to provide multiple evidences of both English language and academic subject achievement. This ongoing project is premised on the belief that language and content are inseparable in teaching and assessment and that innovation in the development of curricula content requires innovative methods of assessment to complement it. It also supports the notion that there is no 'best' way of learning, teaching or assessing and that learners learn in different ways according to a variety of factors. The aims of the project are to:

- encourage student autonomy by involving students in assuming responsibility for providing evidence of their own learning
- develop multiple methods of providing evidence of both academic achievement and of the broader language ability of interest to future employers
- support curricula innovation through the development of an assessment system which can provide multiple evidences over time.
- identify and develop alternative channels of delivery for assessment tasks and determine comparable difficulty levels
- create a mechanism for assisting learners to develop skills in choosing the methods of assessment which will be best suited both to their individual learning styles and the specific competencies for which they wish to provide evidence.

Progress to date will be briefly outlined and advice and suggestions on how best to develop the next stage of the project will be solicited.

Vivien Berry
44 High West,
142 Pokfulam Road,
Hong Kong
Fax: 852-2819-1534
vberry@hkucc.hku.hk

Vivien Berry has taught undergraduate courses in Study Skills and MA courses in language testing and evaluation and research methods at the University of Hong Kong for the past seven years. Prior to that, she was involved in teaching and assessment in Japan, the Middle East and Turkey. Her main research interests are in developing integrative assessment instruments for tertiary level English language and science and investigating how individual learner differences interact with other test facets and affect outcomes on oral test tasks.
The Acquisition and Retention of ESL/Acculturation and Globalization Factors for Japanese Year Abroad University Students
Stephen Carey
University of British Columbia

Abstract
There are few comprehensive studies of year-abroad academically credited ESL/acculturation programs which combine qualitative and quantitative analyses of the process of ESL acquisition for Japanese university students in Western Canada. There are even fewer studies that combine this scope of research of a large scale program with a comprehensive follow-up of these students during and following their return to Japan for the extended period of of 7 years. In this research project, objective measurements were taken of students ESL acquisition prior and during their initial year enrolled in the program at UBC. These conventional TOEFL and OPI scores were supplemented by qualitative interview data which focused on the students identification of critical ESL acquisition experiences. Interviews and questionnaires were also administered for a cross-sectional study of these Japanese students after they had returned to Japan for from one to seven years. There were significant gender, and language of testing effects for the interview and questionnaire data. There were gender effects for acculturation and readjustment to Japanese culture. Males differed from females in accounts of significant ESL acquisition experiences.

Stephen Carey
Language Education Dept.
Faculty of Education
2125 Main Mall
U.B.C.
Vancouver, V6T 1z4
Canada
Fax: 689-7310
stephen.carey@ubc.ca

Stephen Carey teaches in the Department of Language Education at UBC. His interests include the use of electronic bulletin boards for enhancing ESL acquisition by Asian students.
A Self-Instructional Multimedia Rater Training Program
Helen S. Carpenter and Dorry M. Kenyon
Center for Applied Linguistics

Poster Session

Abstract

In this poster session, a new multimedia rater training software program will be demonstrated. The goal of the program is to provide interactive training in applying criteria for rating oral proficiency in a self-instructional format. While there are many innovations in testing using computer-delivered applications, this program seeks to use the computer to improve the rating aspect of performance-based language assessment. The program also provides a prototype that may be applicable to a variety of assessments where raters need to be well trained to apply criteria and rubrics.

The poster will show the conceptual framework for the program and present data collected from users during pilot sessions. The program itself will be available for inspection on a laptop.

Helen S. Carpenter
4646 40th Street, NW
Washington, DC 20016-1859
USA
Fax: (202) 362-3740
helen@cal.org

Helen S. Carpenter coordinates the multimedia rater-training program at the Center for Applied Linguistics through the National Capital Language Resource Center. She is a doctoral student in Applied Linguistics at Georgetown University.

Dorry M. Kenyon heads the Center for Applied Linguistics' foreign language testing division. His current research and development interests lie in applying technology to enhance performance-based language proficiency assessments.
Context and Method in Washback Studies

Organizer: Liying Cheng University of Alberta

Presenters:
Nick Saville University of Cambridge Local Examinations Syndicate
Marvin Wideen Simon Fraser University
Carolyn E. Turner McGill University
Luxia Qi City University of Hong Kong

Abstract

We live in a testing world. Our education system is awash with various high-stakes testing, be they standardized multiple-choice testing or portfolio assessments (Linn, 1992; Murphy et al, 1997). Teachers are held accountable through programs such as achievement indicators and performance incentives in the hope that our students will do better (Council of Ministers of Education, 1999; McEwan, 1995; Stecher & Barron, 1999; Wideen et al, 1997). The belief that assessment can leverage educational change has often led to top-down educational reform strategies in order to bring about changes in teaching and learning by bringing about changes in testing (Noble & Smith, 1994; HKEA, 1994).

This educational phenomenon, commonly called 'washback' in applied linguistics, refers to the influence of testing on teaching and learning. There is evidence to suggest that high-stakes testing has washback effects on teaching and learning (Alderson & Hamp-Lyons, 1996; Alderson & Wall, 1993; Cheng, 1997, 1998 & 1999; Shohamy, 1993; Wall & Alderson, 1993; Watanabe, 1996a, 1996b). The extensive use of and/or misuse of examination scores for various educational and social purposes has made washback a distinct, complex phenomenon both in general education and in language education. It has been thus linked to the validity of the test-systemic validity (Fredericksen & Collins, 1989), and Messick's notion of consequential validity (Messick, 1996).

Over the past two decades, steadily increasing number of research studies have been carried out in many parts of the world (e.g. Australia, Brazil, Canada, Germany, Greece, Hong Kong, Japan, Sri Lanka, the UK, and the US) to investigate the nature and mechanism of washback, how the phenomenon occurs and to which educational contexts and testing cultures are most relevant. These studies have also investigated the impact of testing on multifaceted aspects of teaching and learning. Therefore, implications and recommendations can be made available to policy makers, education and testing organizations within the contexts under study. At the same time, research methodology for such an interdisciplinary study has brought about new challenges to researchers investigating this phenomenon (Watanabe, 1996a; Rogers, 1990).

This symposium aims to provide a forum for further discussion of the above phenomenon at the intersection of language testing and teaching practice/programs. Participants in this symposium will shed light, by making reference to their own empirical studies, on washback, highlighting the phenomenon as an evaluation, as a curriculum innovation, and as a top-down educational reform. They will also share with the audience the essentiality of an in-depth understanding of the complicated issues embedded in classrooms within specific education contexts under study. In addition, they will discuss the various quantitative and qualitative methods being used in their studies. The presentations will include the IELTS impact study by UCLES; the impact of large-scale testing on instructional activity in British Columbia secondary schools; the initial steps of a study investigating washback from empirically derived rating scales for ESL speaking in Quebec provincial exams; and the study of the intended washback of the National Matriculation English Test in China.

Liying Cheng
Department of Educational Psychology
University of Alberta
6-110th Education North
Liying Cheng is based with the Center for Research in Applied Measurement and Evaluation and the graduate ESL program at the Department of Educational Psychology, University of Alberta. She is educated in China, UK, Hong Kong and Canada, and has worked more than 15 years as an ESL/EFL teacher, teacher educator, program developer and researcher in the above places. Some of her current research projects include an impact study of high-stakes public examinations; an investigation of ESL/EFL teachers’ classroom assessment practices; an interpretation of differential item and bundle functioning on translated tests; and an investigation into EAP requirements and academic achievement at the tertiary level.

Nick Saville has over 20 years experience in the field of linguistics, English language teaching and language testing. After completing a degree in Linguistics, he worked for six years at a university in Italy where he gained extensive practical experience in setting and administering English language tests. In 1985 he returned to university to study for a Masters degree in TEFL and completed a dissertation in the field of language testing. In 1987, he began working for the University of Cambridge Local Examinations Syndicate (UCLES) as Project Manager for a test development project in Tokyo. After 2 years in Japan he moved back to Cambridge to join the EFL Division working on test validation projects. Since 1994, he has been Group Manager for Test Development and Validation and is now an Assistant Director for UCLES EFL.

Marvin Wideen is a professor at the Institute for Studies in Teacher Education, Faculty of Education, Simon Fraser University.

Carolyn E. Turner is Associate Professor and Graduate Program Director, Department of Second Language Education, McGill University, Montreal, Quebec, Canada. Her main focus/commitment is language testing and evaluation within educational settings. She carries this out through her teaching, research, service to the community, and publications. She works with outside organizations on testing matters (e.g., Ministry of Education in Quebec, Educational Testing Service and the TOEFL family of tests). She is working on large-scale educational projects involving empirically derived rating scales and the impact of their use. She is active in the language testing community (e.g., Language Testing journal Editorial Advisory Board; ILTA Executive Board; Language Testing Research Colloquium, LTRC, ongoing participation). A recent publication pertinent to this symposium is: Turner, C. E. (In press). The need for impact studies of L2 performance testing and rating: Identifying areas of potential consequences at all levels of the testing cycle. In A. Brown et al. (Eds.), Experimenting with uncertainty: Language testing essays in honor of Alan Davies (pp.127-139). Cambridge, Cambridge University Press.

Luxia Qi has a professional and personal interest in language education. She holds two MAs in applied linguistics, one is awarded by the Guangdong University of Foreign Studies (GUFS) in China and the other by the University of Lancaster in England. She has taught in two universities in China, first the Guiyang Normal University and then GUFS. Besides teaching, she has been involved in language test development for the National Education Examination Authorities in China for the last 10 years. Her teaching and research are in the areas of reading, testing and second language acquisition. These experiences have made her fully aware of the influence of testing on teaching, learning and people’s lives. This is the main reason why she is currently doing Ph.D. studies in the English Department of the City University of Hong Kong focusing on the issue of washback.
Assessing Integrated Academic Skills for International Undergraduates
Dan Douglas, Cynthia Myers, Volker Hegelheimer, Felicidade Van Acker
Iowa State University

Poster Session

Abstract

The integration of language learning with broader curricular goals (Richards 1998) is a powerful motivating factor in student learning. Integrated tasks are defined as ones in which all the tasks are thematically linked and where the input that has been provided forms the basis for the response(s) to be generated by test takers (Lewkowicz 1997: 121). The use of integrated tasks is justified on the grounds of providing learners with material and ideas to write about (Read 1990; Weir 1993) and to enhance the interpretation of test performance as predictive of performance in non-test, academic situations (Wesche 1987).

Based on a model developed by Davidson and colleagues (http://ux6.cso.uiuc.edu/~fgd/vressay.htm), we are producing an integrated multi-media EAP instructional and assessment package to teach and assess academic listening, reading, and writing to international undergraduates across the university curriculum. Faculty in several disciplines videotaped short lectures for small groups of students on introductory topics in their fields. They also provided short, thematically related readings to accompany the lectures. For each set of lectures and readings, learners are given instructions for a writing task appropriate for the discipline and topic, involving the integration of the lecture and reading material. The formative assessment instruments provide diagnostic feedback for learners in such areas as paraphrasing skills, appropriate rhetorical organization, and use of quotes and avoidance of plagiarism. The assessment criteria include topical knowledge as well as aspects of language ability. The lectures, reading passages, and writing tasks are being placed on CD-ROM for use in academic English classes and for self-access by students in writing labs. In using these materials, students are exposed to listening, reading and writing across disciplines to help them more easily transfer academic skills from the English class to their content classes.

The poster will outline the theoretical rationale for the integrated multi-media procedure and demonstrate the instructional and assessment materials.

Dan Douglas
English Department
Ames, IA 50011
Fax: (515) 294-6814
E-mail: dan doug@iastate.edu

Name: Volker Hegelheimer
English Department
Ames, IA 50011
Fax: (515) 294-6814
E-mail: volkerh@iastate.edu

Cynthia Myers
English Department
Ames, IA 50011
Fax: (515) 294-6814
E-mail: cly Myers@iastate.edu

Name: Felicidade Van Acker
English Department
Ames, IA 50011
Fax: (515) 294-6814
E-mail: fva@iastate.edu

Dan Douglas, Cynthia Myers, and Volker Hegelheimer teach in the TESL/Applied Linguistics Program at Iowa State University. Felicidade Van Acker is an MA candidate and research assistant in TESL/Applied Linguistics at Iowa State.
History of the Baccalauréat
Samira Elatia
University of Illinois at Urbana-Champaign

Abstract

The Baccalauréat is a comprehensive exam given originally in France. This exam dates back to 1808 when Napoléon Bonaparte was looking for an exam that will enable the “empire” to select civil servants. It was based on the idea of “merit” not privileged. The Baccalauréat is an exit exam that is given to students in their last year of High-School. The Baccalauréat spread in many countries around the world, mostly francophone countries that were ex-colonies of France. Almost two centuries old, the Baccalauréat underwent several changes and is still undergoing change specially with regards to languages.

Languages have always been an important part of the Baccalauréat exam. In the beginning, the Baccalauréat was exclusively an oral exam and languages were a major component of the test. Candidates were not tested in one language but in several languages: Latin, Greek and other important languages of Europe such as English, German and Italian. There were some subjects that were studied in foreign languages, namely philosophy. In the modern Baccalauréat, languages are still very important. In all the subdivisions and specialties of the Baccalauréat, whether it is mathematical sciences, natural sciences, or literature, there is a foreign language test. Students or candidates chose which language they want to be tested on. The Baccalauréat offers exams in English, Latin, Spanish, German, Italian, Arabic and other languages. There is also a Baccalauréat of languages, where the main components of the exam are languages, three or more languages.

In this poster presentation, I will display a historical development of the Baccalauréat exam in France. I will also show through this historical display, the role and the importance of languages in the Baccalauréat exam. Following is an outline of the components of the presentation:
1. A definition and origin of the word Baccalauréat from the Middle Ages to the 19th century.
2. A chronological development of the Baccalauréat since its birth in 1808 under the supervision of Napoléon Bonaparte.
3. A graphic display of the three major periods in the development of the Baccalauréat:
   a. First period from 1808 to 1840
   b. Second period from 1840 to 1874
   c. Third period from 1874 to 1902
4. The increasing number of the people passing the Baccalauréat exam before 1900 and after 1900.
5. The Baccalauréat today:
   a. Charts of the different section of the Baccalauréat.
   b. Presentation of the number of “candidates” sitting and passing the Baccalauréat.
   c. A short comparison between the Baccalauréat in France and in another Francophone country.

Samira Elatia
2090 Foreign languages Building
707 S.Mathews Ave.
Urbana, IL 61801
USA
Fax: (217) 244-2223
elatia@uiuc.edu
Putting Tasks to the Test

Organizer: Dr. Cathie Elder University of Melbourne

Presenters:
Noriko Iwashita (on behalf of Tim McNamara, Noriko Iwashita and Cathie Elder) University of Melbourne
Helen Slatyer (on behalf of Geoff Brindley, Gill Wigglesworth and Helen Slatyer) Macquarie University
John Norris (on behalf of J. D. Brown, John Norris and Tom Hudson University of Hawai'i at Manoa

Panelists:
Prof. Lyle F. Bachman University of California, Los Angeles
Prof. Peter Skehan Thames Valley University
Prof. Peter Robinson Aoyama Gakuin University

Abstract
How relevant for Language Testing are SLA research findings about tasks? Can these fields usefully inform one another or are the requirements of acquisition, teaching and testing too different? While language testing relies increasingly on complex tasks of the kind investigated by SLA researchers and advocated for use within L2 pedagogy, the impact of task characteristics on test performance and the constraints which different task choices may have on score inferences have not yet been widely discussed. In this symposium we invite prominent researchers from SLA and Language Testing to present and/or critique the findings of recent research on task-based assessment and to debate a range of issues at the interface between task-based language teaching and testing.

The symposium comprises three paper presentations followed by a response from three panelists. The first paper reports on a large scale experimental study conducted for the Educational Testing Service which explores aspects of task structure and task performance conditions which have been claimed by SLA researchers to affect task difficulty and which attempts to reproduce these findings in relation to narrative tasks in an oral proficiency testing context. Findings reveal that variations in task structure and task performance conditions do not affect performance in the way that had been predicted.

The second paper focuses on the role played by certain key task variables in the assessment of reading and listening skills, using tasks drawn from a pre-existing curriculum and assessment framework. The effect of systematic manipulation of these variables (which include contextual support, background knowledge, nature of the input and item format) on test scores is reported. Findings reveal a complex interaction between the text and items which overrides the potential effect of the task variables.

The third paper summarizes findings from recent research into the development and use of task-based performance assessment for achievement testing purposes. Particular attention is paid to the relationship between a priori estimates of task difficulty (based on cognitive processing demands) and evaluations of actual task performance from several interpretative perspectives. Apparent discord between task difficulty estimates and subsequent performance evaluations is discussed in the light of task types selected, and the way these tasks are operationalized in the test situation.

Panelists will comment on the implications of each presentation for our conceptualisations of tasks in both SLA and Language Testing and for the specification and design of tasks for teaching, testing and research purposes. Proposals for collaborative research between scholars in SLA and language testing will also be put forward.

Dr. Cathie Elder
Language Testing Research Centre
Department of Linguistics and Applied Linguistics
2nd Floor
138-146 Cardigan St Parkville 3053
Victoria AUSTRALIA

23
Panelists

Lyle F. Bachman is Professor of Applied Linguistics and TESL at the University of California, Los Angeles. His current professional interests include the development and validation of tests of language ability, the technology of test design and development, and the interfaces between language testing research and other areas of applied linguistics research. His publications include Fundamental Considerations in Language Testing, Language Testing in Practice (with Adrian S. Palmer), and Interfaces between Second Language Acquisition and Language Testing Research (co-edited with Andrew D. Cohen.) He has also published numerous articles in the area of language testing and evaluation. He is currently editor of the journal Language Testing, and co-editor of the Cambridge Language Assessment Series.

Peter Robinson is Professor of Linguistics and SLA in the Department of English, Aoyama Gakuin University in Tokyo. His research interests are in the role of cognitive factors during instructed language learning.

Peter Skehan is Professor of English Language Teaching at Kings College, London. He has held academic posts at Birmingham University, the Institute of Education London University, Thames Valley University and Kings College, London. His current interests are individual differences in language learning; language testing; and second language acquisition, especially task-based instruction. Recent publications include Individual Differences in Second Language Learning (Arnold, 1989) and A Cognitive Approach to Language Learning (OUP, 1998).

Paper authors (NB Presenters of each paper are marked with an asterisk)

Hawai’i

*John Norris is a Mellon Fellow at the National Foreign Language Center, Washington DC, where he is completing his Ph.D. (SLA, University of Hawaii) dissertation research on assessment validation in university-level foreign language education contexts. In addition to methods for validity inquiry, his research interests include task-based and other forms of performance assessment, L2 instructional effectiveness, and research synthesis and meta-analysis. He has worked in language education in Brazil, Hawaii, Japan, and Washington DC.

James Dean ("TD") Brown, Professor on the graduate faculty of the Department of ESL at the University of Hawaii at Manoa, has published numerous articles on language testing and curriculum development, and several books on: reading statistical language studies (Cambridge U. Press), language curriculum development (Heinle & Heinle), language testing (Prentice Hall; also translated into Japanese by Wada, Taishukan Shoten publishers), an edited collection with Yamashita on language testing in Japan (JALT), two with Hudson and Detmer on testing pragmatics (U. of Hawaii Press), one with Norris and Hudson on performance testing (U. of Hawaii Press), and an edited collection of ideas for classroom testing (TESOL).

Thom Hudson is Associate Professor on the graduate faculty of the Department of ESL at the University of Hawai’i at Manoa. His research interests are second language testing, reading, curriculum, materials and research methods. He has presented his work at several conferences and colloquia, and has published articles on language testing, second language reading, and second language pragmatics. He has taught in the U.S., Egypt, Mexico, and Japan.

Melbourne
Tim McNamara is Associate Professor in Applied Linguistics and Director of the Language Testing Research Centre at the University of Melbourne. Tim's research interests include performance assessment in second languages particularly in occupational and academic contexts; the application of Rasch measurement in language testing; and theories of performance in second language performance assessment.

*Noriko Iwashita is a Research Fellow at the Language Testing Research Centre in the Department of Linguistics and Applied Linguistics at the University of Melbourne and a former lecturer in Japanese as a foreign language. She has recently completed her PhD thesis on the role of conversation in the acquisition of Japanese as a second language and has conducted a number of major research projects at the interface between language testing and second language acquisition.

Cathie Elder is Associate Professor at the Institute of Language Teaching and Learning at the University of Auckland and a former Director of the Language Testing Research Centre at the University of Melbourne. Her research interests span the areas of language policy, language testing, bilingualism and the teaching and assessment of languages other than English in schools.

Macquarie

Geoff Brindley is a Senior Lecturer in Linguistics and Research Co-ordinator in the National Centre for English Language Teaching and Research, Macquarie University, Sydney, Australia. He is the author of a variety of publications on language proficiency assessment, second language acquisition and curriculum design.

*Helen Slatyer is a research officer at NCELTR/Macquarie University and has been involved in a number of research projects which investigate the design and difficulty of assessment tasks for the Certificates in Spoken & Written English (CSWE) of the Adult Migrant English Program. She also lectures in bilingualism and community-based interpreting for the Master in Applied Linguistics degree and Graduate Diploma in Translation and Interpreting in the Linguistics Department at Macquarie University.

Gillian Wigglesworth is Senior Lecturer in Linguistics and Convenor of the Master of Applied Linguistics programs at Macquarie University. Her research interests encompass first and second language acquisition, and the evaluation and assessment of second languages. She has published both in Australian and internationally refereed journals and books on issues related to language assessment, second language acquisition and linguistic and cognitive development of school age children. She has co-edited three books, one on language and gender, and two on issues in language testing and evaluation, and has presented at numerous conferences, and been invited to present seminars and workshops both in Australia and overseas.
Issues in Video Listening Assessment
Christine Feak and Julia Salehzadeh
University of Michigan, Ann Arbor

Work in Progress

Abstract

Despite all the recognized benefits of using video in the language classroom, the use of video in listening assessment has been limited (Brindley 1998; Progosh 1998). Concomitantly, research into issues raised by a video test as opposed to an audio-only test, has been limited. What little research has been done on video assessment, has focused on test-takers views (Progosh 1996) and the validity (Shin 1998) of a video assessment. While these issues are not unimportant, the use of video, rather than audio alone raises many more substantive issues, particularly with regard to the actors used in the video presentation.

The development of a video assessment poses significant challenges to test designers that do not exist for an audio alone test. For instance, content issues aside, a main concern in an audio assessment is the character of the voices used. Should male or female voices be used? Or does it not matter? Should the pronunciation be accented or unaccented (however these may be defined)? Although developers of a video test need to deal with these issues as well, they must also consider non-verbal or para-linguistic issues that are masked by an audio alone assessment. For example, special consideration must be given to the age, gender, and ethnicity of the actors as well as to the kinds of roles they will adopt.

This work in progress is concerned with a recently developed video listening assessment (VLA) designed to place students into or exempt them from EAP listening course work at a large midwestern university. This assessment presents short segments from a small university film class, which are then followed by various kinds of multiple choice questions. The classroom segments include a first-day lecture and teacher-led discussions in which the students actively participate.

As part of the trialing process for this new assessment, learners are asked to complete a questionnaire, which among other aspects of the assessment, asks them to comment on how the students are portrayed. While response to the VLA has been overwhelmingly positive, a small number of test-takers has voiced some concern over how members of certain groups are represented. Thus, one aspect of our VLA research is focusing on this issue, particularly with regard to the following questions. Are certain groups somehow portrayed negatively? To what extent do we need to consider the criticisms of small number of students? Can we ever develop a video test that would not raise any concerns about how a particular group is represented in a video? Should we abandon our video test because of concerns raised by a relatively small number of students?

Developers of the video test will discuss these and, if time permits, other issues related to video assessment with session participants.

3039 NUB
English Language Institute
University of Michigan
Ann Arbor, MI 48109
Fax: (734) 763-0369
cfeak@umich.edu

Christine Feak is a Lecturer at the English Language Institute, University of Michigan, as well as a Research Assistant in the ELI’S Testing and Certification Division. She is active in the development of the Institute’s Academic English Assessment, which is designed to place university students into or exempt them from English for Academic Purposes course work.

Julia Salehzadeh is a Lecturer at the English Language Institute, University of Michigan, where she teaches courses in Listening Comprehension. She is also the lead Research Assistant in the development of the Institute’s video listening assessment.
Using a Communicative Framework to Compare
Language Proficiency Tests: The Validity Problem
Ardeshir Geranpayeh
University of Edinburgh

Abstract
A fundamental concern in the comparability of language proficiency tests is to find a valid
framework with which to compare language tests meaningfully. Bachman and his colleagues, in the
course of the Cambridge-TOEFL comparability study (CTCS), designed an empirical framework based on
Bachman's (1990) facets of test methods (TM) and communicative language ability components (CLA).
This paper examines the usefulness of such a communicative framework in the comparability of two
language proficiency tests: IELTS and TOEFL. The framework used in this research was a modified
version of the one used in CTCS; the modification being made partly on the basis of an ongoing research
in UCLES and partly on the limitation of research resources. The two content analysis instruments used in
the framework allowed us to compare 34 different aspects of the language tests; 13 components of
communicative language ability and 21 facets of test methods. The results of the content analysis show
that the instruments were highly susceptible to the interpretation of expert judges of what an item was
testing and of what the instructions required them to do. The inter-rater reliability for two of the judges
who rated the CLA components and some of the TM facets was very high; the two had worked together
on the research project for over two years. However, that of the three judges who rated some of the TM
facets was very low; they had individual briefings about the clarity of the instructions but no joint
meetings.

Ardeshir Geranpayeh
University of Edinburgh
Department of Theoretical & Applied Linguistics
14 Buccleuch Place,
Edinburgh EH8 9LN
Fax: +44 131 650 6526
Email: ardeshir.geranpayeh@ed.ac.uk

A PhD candidate about to submit the thesis on comparability of language proficiency tests. Research
interests: language proficiency tests comparability, test takers characteristics, CALL, and SLA. Sex: Male.
Age: 42.
Content and Context Visuals and Performance on TOEFL CBT Listening Comprehension Stimuli
April Ginther
Purdue University

Abstract
The listening comprehension section of the TOEFL has traditionally involved audio presentations of language without accompanying visual stimuli. Now that TOEFL is computer based, listening comprehension items are being created that include both audio and visual information. This study was conducted in order to begin to understand the effects of different types of visual presentations that might accompany the audio portion of the stimuli. A nested cross-over design (subjects nested in proficiency, level, and form) was used to examine the effects of language proficiency (high or low), still photos (present or absent), and type of stimuli (dialogues/short conversations, academic discussions, mini-talks with context visuals, mini-talks with content visuals) on performance on standard multiple choice items. Three two-way interactions were significant: proficiency by type of stimuli, type of stimuli by time, and type of stimuli by visual condition. The weakest of these interactions, type of stimuli by visual condition was the most interesting and indicated that the presence of visuals results in facilitation of performance when the visuals bear information that compliments the audio portion of the text. In addition, subjects were administered a series of questionnaires concerning their preferences with respect to the visual or non-visual experimental conditions. The majority of the subjects indicated a strong preference for the presence of visuals.

April Ginther
English Department
Heavilon Hall
Purdue University
West Lafayette, IN 47906
aginther@purdue.edu
Reverse Engineering of Two Tests of French as a Second Language
Sandhya GOPALAN and Samira ELATIA
University of Illinois at Urbana-Champaign

Abstract
Reverse Engineering is a process by which a test or a task is analyzed for the purpose of re-designing its specification. Reverse Engineering can serve to critique a test/task in order to improve it or to simply re-design a previously non-existent blueprint so that it can generate more items.

In this poster presentation, we will present two different reversed-engineered test specifications for two listening comprehension tests of French. The tests that we studied were geared towards intermediate level French classes from two different universities in Illinois. In the first test, we will develop a blueprint for one item in the test that has generated many items and that has been included in the exam for several years. In the second test, an achievement test (final exam) of listening comprehension, we will build a test specification that will allow us to analyze and critique the components of the three sections of the exam.

The poster presentation will contain the following categories that are considered essential in a test specification process:

I. Title of the specification
II. Context
III. Mandate
IV. General Description
V. Proctor instructions for administering this listening task
VI. Prompt Attributes
VII. Response Attributes
VIII. Choosing the ‘bad’ and the ‘good’ items.
IX. Recommendations for selecting the listening comprehension text
X. Sample item
XI. Grading and the rater.
XII. Graphs of the results of the test.

In all the sections of the presentation, reliability and validity issues will be highlighted with respect to each item.

Samira Elatia
2090 Foreign languages Building
707 S.Mathews Ave.
Urbana, IL 61801
USA
Fax: (217) 244-2223
elatia@uiuc.edu
Diagnosis before Cure: The Interface between Language Testers and Medical Educators
Elisabeth Grove and Annie Brown
University of Melbourne

Abstract
This paper considers the practical and theoretical issues involved in developing and validating an EAP test of communication skills. Such tests inevitably involve crossing the boundaries set for general tests of second language proficiency, and also call into question what may be understood by 'communication'. The subject of the paper, a two-tier diagnostic test developed for the Faculty of Health Sciences of an anglophone university, is used to assess the English communication skills of incoming first-year students, native and non-native speakers alike. Its purpose is to identify those students whose level of communicative ability puts them at risk of under-achievement in the course, so that appropriate support and remediation can be offered as early as possible.

The interface between the language testers and medical educators was essential to the development of the test and also at times a source of creative discomfort to both parties. We argue that the negotiations between the two teams of participants, which involved accommodating differing views of communication, as well as acknowledging the practical and theoretical limits of testability, resulted in a test that is both innovative and flexible. The test tasks, assessment criteria and reporting procedures reflect a much more explicit construct of communicative effectiveness than is normally the case in traditional EAP tests. Qualitative data and statistical comparisons between test scores and subsequent course results are presented in support of test validity and the further contention that the test represents a de facto critique of commonly-used and notoriously vague measures of 'communicative competence'.

Elisabeth Grove
Language Testing Research Centre
138-146 Cardigan St.
Carlton, 3053
Victoria, AUSTRALIA
Fax: 001-61-3-344 5163
l.grove@linguistics.unimelb.edu.au

Annie Brown
Language Testing Research Centre
The University of Melbourne
Level 2, 138-146 Cardigan St.
Carlton
Victoria 3052
Australia
Fax #61 3 9344 5163
a.brown@linguistics.unimelb.edu.au

Elisabeth Grove is Research Fellow in the Language Testing Research Centre, at the University of Melbourne. Since joining the LTRC, she has had extensive experience of LSP testing projects. She has run assessor training programmes, taught courses in language assessment and undertaken research into LSP tests of reading and competency-based language curriculum and assessment. She has a particular interests in the assessment of communication in the medical setting.

Annie Brown is Deputy Director of the Language Testing Research Centre at the University of Melbourne, where she has been employed since its inception in 1990. During this time she has been directly involved in or responsible for a number of test development and validation projects, as well as teaching Language Testing at postgraduate and certificate levels. She is currently completing a PhD in the area of oral language testing, investigating the interlocutor effect. Her other research interests include language program evaluation and discourse analysis.
The Role of Digital Video Media in Response to Task Demands

Paul Gruba
The University of Melbourne

Abstract

Despite the widespread use of video and multimedia in second language instruction, little is known about how learners attend to dual-coded media and, in particular, how dynamic images influence understanding. In light of this lack of research, the aim of this investigation was to examine the role of visual elements in second language listening comprehension when digital video was used as a mode of presentation.

To achieve this aim, the investigator worked closely with Japanese language instructors to select three authentic Japanese news (NHK) broadcasts. These video clips were then digitized for computer display. Following an extensive trial study and the development of open-ended questions, twelve tertiary students of Japanese were directed to provide immediately introspective verbal reports as they attended to the videotexts. Data was analysed using both quantitative and qualitative approaches. Based on constructivist perspectives of comprehension, two frameworks were developed regarding 1) initial front-to-back comprehension and 2) engagement with videotext in response to open-ended task demands.

The results of the investigation point out that visual elements work in a number of ways that go beyond merely 'supporting' verbal elements; they are better thought of as integral resources to comprehension whose influence shifts from primary to secondary importance as a listener develops a mature understanding of the videotext. The investigation concludes with discussion regarding video-mediated listening assessment, computer-based language testing and, briefly, media studies as a point of departure for research in second language settings.

Centre for Communication Skills & ESL
The University of Melbourne
138 Cardigan Street, Carlton VIC 3010
AUSTRALIA
Fax: +61-3-9349-3501
p.gruba@language.unimelb.edu.au

After earning a MA/TESL from the University of California, Los Angeles, Paul directed a language testing research project at Kanda University of International Studies in Chiba, Japan. The results of the project lead to the implementation of the video-mediated Kanda English Proficiency Test (KEPT) as the basis for ranking student proficiencies at the university. Based on these experiences, Paul recently completed a doctorate at the University of Melbourne that investigated the role of digital media in second language listening comprehension. At present, Paul works as a lecturer on the 'Communication Across the Curriculum' Project that aims to integrate communication skills into discipline-specific teaching. He also assists with the development of subjects in Information Technology and Communication, Language and Media and short courses concerning thesis writing for postgraduate students.
Border Crossings: What Multiple Proficiency Ratings Reveal
Gene Halleck
Oklahoma State University

Abstract
Critics of the American Council on the Teaching of Foreign Languages (ACTFL) Speaking Proficiency Guidelines have questioned whether they are reliable predictors of actual performance or whether they are just the result of decades of language teachers' intuitions. Some have called into question the validity of a single, global rating (Bachman 1988) suggesting that it is empirically unjustified and that the assumption in the Guidelines that skills are acquired in a linear fashion, is unwarranted (Kramsch 1986). Such criticisms call for empirical research in which a holistic rating derived by applying the Guidelines could be compared to a number of different abstract rating scales such as those proposed by Bachman and Palmer (1983) and Bachman and Savignon (1986). This poster illustrates an analysis of 30 Oral Proficiency Interviews rated according to the Guidelines and then also rated for grammatical, discourse, and sociolinguistic competence.

The ratings of ten interviews on each side of a "major border" and ten interviews on each side of a "minor border" were compared. In other words, I compared the grammatical, discourse and sociolinguistic competence of ten Superior interviews with ten Advanced High interviews, and then compared the Advanced High interviews with ten interviews that were only rated at the Advanced level.

According to ACTFL, the twenty interviews with the designation of "Advanced" (i.e. the Advanced High and the Advanced interviews) should have had more in common with each other than the Advanced High interviews should have had with the Superior interviews. But by clearly distinguishing different aspects of oral proficiency and assigning separate ratings to each of these abilities an interesting picture emerged. The results show that the language elicited in the Advanced High interviews is more similar to that of the interviews across the major border (Advanced High/ Superior) than it is to the language of the interviews across the minor border (Advanced High/Advanced). Yet ACTFL's system of a single, global rating could not reflect this.

Thus these empirical data call into question not only the validity of designating some borders as "major" and others as only "minor" but also point out that by assigning a single, global rating to describe the performance of an entire interview, we are missing important information about the candidate's communicative competence.

Gene Halleck
205 Morrill Hall
Stillwater, OK 74078
USA
Fax: 495-744-6326
GHAL100109@aol.com

Gene Halleck is an Associate Professor of TESL/Linguistics in the English Department at Oklahoma State University, and is the Director of the ITA Orientation, Testing, and Training Program.


**Ethical Dilemmas in Language Testing: What Can We Actually Do?**

Liz Hamp-Lyons and Tom Lumley

*Hong Kong Polytechnic University*

**Abstract**

In Hong Kong, where English is both a second language and the language of instruction in universities, increasing concern is being expressed about the English proficiency of graduates once they enter the workplace. One response to this concern has been the development of an exit assessment, designed to predict English language effectiveness beyond the academic setting: the Graduating Students' Language Proficiency Assessment (GSLPA) - English. This test is designed primarily for the information of prospective employers, and consequently focuses on English communication on the professional workplace, rather than on EAP as is usual in universities. The implementation of this instrument provides an ideal context for the investigation of the consequences of both test interpretation and test use (Messick, 1989). The GSLPA (English) is used as a case study to understand some of the principles of accountability, ethics and washback, from the perspective of the language tester. This paper examines how these can conflict in practice with the real-world demands and concerns of a variety of stakeholders (who sometimes represent very divergent positions): employers, policy makers, language educators, others involved in language assessment, academics, tertiary institutions and students. The paper proposes principles and practices for addressing areas of conflict, which include:

- the demand for evidence of the effectiveness of university language enhancement programmes;
- proposals to use the GSLPA in other contexts;
- the use of other language assessment instruments for similar purposes;
- cost;
- the need to reform an education system already too dependent on exams; and
- the concern about comparisons across universities.

**Asian Centre for Language Assessment Research**

Department of English

Hong Kong Polytechnic University

Hung Hom

Kowloon

Hong Kong

Fax: 852-2333-6569

ehl@polyu.edu.hk

Tom Lumley is Assistant Professor at Hong Kong Polytechnic University, and an Associate of the Asian Centre for Language Assessment Research there. Before that he was a Research Fellow at the Language Testing Research Centre, University of Melbourne, for a number of years. He has had extensive involvement in a wide range of language assessment research and development projects in numerous contexts in Australia and Asia. His research interests are in performance assessment, especially of writing, and in specific-purpose assessment, as well as in applications of Rasch measurement in language assessment. His doctoral studies combine several of these threads with analysis of verbal protocols to explore the process of rating from the perspective of the rater, in a large-scale test of English.
A New Standard Setting Method for Multiple Cut-Off Scores

Fellianka Kaftandjieva and Norman Verhelst
Center for Applied Language Studies, University of Jyväskylä
National Institute for Educational Measurement, The Netherlands

Research Paper

Abstract

A common practice in language testing is that the test results are reported in terms of 'proficiency levels' described a priori as a scale with a number of ordered categories (usually more than two) where the current standard-setting methods for dichotomous classification cannot be applied.

This paper aims:

- to propose a new standard-setting method which will enable test score conversion to multiple levels of language proficiency, and
- to present the results of its application for test score conversion, for two different foreign language tests, to scales of language proficiency presented in the Council of Europe Common Framework.

The method can be classified as a test-centered continuum method where the cut-off points are set to minimize the preliminarily defined loss function that combines two sources of information:

1. Item parameter estimations, based on the IRT modelling, and
2. Experts' judgements about the relation between item mastery and the level of proficiency.

Two 60-item foreign language tests were used in this study with an incomplete design including 4 booklets. Each of the items was piloted with at least 200 examinees, i.e. a total of 419 examinees took part in the study.

Seven experts participated in the judgement process. After a training session, they had to decide independently for each of the items included in the tests whether a test-taker on level X should be able to answer it correctly.

Preliminary results show that the proposed method is easy to implement and overcomes many of the disadvantages of the other standard-setting methods.

Fellianka Kaftandjieva
Centre for Applied Language Studies
University of Jyväskylä
P.O.Box 35
40351 Jyväskylä
Finland
Fax: +358-14-603521
felkatt@cc.jyu.fi

Fellianka Kaftandjieva: Ph.D. in Education, MA in Mathematics, Associate Professor in Educational Measurement and Evaluation at Sofia University (Bulgaria) and a researcher at the Centre for Applied Language Studies, University of Jyväskylä (Finland). Teaches courses in Research Methods, Test Theory, Statistics and Gifted Education. Main interests - Item Response Modelling, Differential Item Functioning.

Norman Verhelst: Ph.D. in Psychometrics, MA in Psychology, Professor in Psychometrics at Twente University (The Netherlands) and a senior researcher at CITO, Arnhem (The Netherlands). Teaches courses in Psychometrics and Test Theory. Main interests - Item Response Modelling, Computer Programming.
A New Generation of Language Testers Using Structural Equation Modeling
Shoichi Gregory Kamei
University of California, Los Angeles
Amy D. Yamashiro
Nihon University, Tokyo
David Beglar
Temple University Japan, Osaka

Greta Gorsuch
Texas Tech University, Lubbock

George A. Marcoulides
California State University at Fullerton

Abstract
Advances in computer technology and statistical software have made structural equation modeling (SEM) accessible to a wider audience of researchers and graduate students. Because of this, new generations of language testers require proper training to use SEM appropriately. This symposium-style presentation presents four language testing dissertation projects that employed SEM and commentary by a designated respondent who is an acknowledged expert in SEM methodology and training.

The first two studies, presented jointly in "Applying Structural Equation Modeling to MTMM Data: Concepts and Issues," employ correlated uniqueness (CU) models to analyze multitrait-multimethod (MTMM) data. One study employed a correlated trait, correlated uniqueness (CTCU) model and utilized multiple sample confirmatory factor analysis (CFA) to investigate the stability of a telephone oral interview across administrations in a small sample context. The other study combined correlated uniqueness (CU) with the use of a higher order latent factor to validate a criterion-referenced oral performance test.

The next presentation, "Construct Validation of a Second Language Vocabulary Test," describes a structural model utilizing a historico-developmental analysis of the words on a vocabulary test designed to measure second language learners' breadth of vocabulary. One concern is that many fundamental lexical variables, such as word frequency, are non-normally distributed. As a result, these variables must be transformed to improve the results of subsequent analyses. Another issue in the development of this structural model is the model's level of complexity. The simple model used in this study is powerful in that it is generalizable to a wide number of contexts and to different types of vocabulary tests. However, the drawback of a "simple" model is that it does not adequately represent the full complexity of lexical acquisition and use.

The third presentation, "Analyzing 'SEM-Challenged' Data: Complex Models and Reluctant Respondents," explores the application of SEM on survey data from 876 Japanese high school English as a foreign language teachers on policy and non-policy topics, such as university entrance examinations, that are thought to influence the teachers' classroom instruction. SEM was employed for survey validation using a complex model focusing on single indicator factors. This study found that teachers' reticence in responding to more "politically sensitive" items regarding their opinions of communicative and more traditional language learning activities resulted in measured variables with very limited variance and a concomitant decrease in the reliability of a fourth latent variable.

A short question-and-answer session follows each presentation. After the three presentations, the respondent comments on the student projects and discusses the training of new language testers in the use of SEM. This discussion will include considerations and warnings on the appropriate and responsible use of SEM, where SEM fits in the repertoire of language tester skills, and the validity of SEM approaches in language testing research and practice. Time is also allocated for a general discussion open to the floor.
Shoichi Gregory Kamei is a graduate student in the Interdepartmental Program in Applied Linguistics at the University of California, Los Angeles. His interests include language testing, language for specific purposes, distance education, and learner autonomy.

Amy D. Yamashiro, a lecturer at Nihon University, is a doctoral candidate at Temple University Japan (TUJ). She is the graduate student Editor of the TUJ Working Papers in Applied Linguistics. Her research interests include structural equation modeling, language testing, speech communication, and gender.

Greta Gorsuch is Assistant Professor of Applied Linguistics at Texas Tech University. She is former Editor of The Language Teacher, and is interested in large scale survey research, and developing testing courses for teachers.

David Beglar teaches in the Graduate College of Education at Temple University Japan where he is also a doctoral candidate. His research interests include language testing, structural equation modeling, vocabulary acquisition, and language for specific purposes.

George A. Marcoulides is Professor of Statistics at California State University at Fullerton and Adjunct Professor at the University of California at Irvine. He is the recipient of the UCEA William J. Davis Memorial Award for outstanding scholarship. He is currently Editor of the Quantitative Methodology Book Series, Editor of the Structural Equation Modeling journal, and on the editorial board of several other measurement and statistics journals. His research interests include generalizability theory and structural equation modeling.
Detecting DIF (Differential Item Functioning) in a Speaking Test with Polytomous Response Data

Mikyung Kim
University of California, Los Angeles

Abstract

The investigation of differential item functioning (DIF) has attracted a great deal of attention from test developers because DIF items pose a considerable threat to the validity of tests. In particular, it is crucial to detect DIF items in language proficiency tests in which test takers with diverse backgrounds are involved. To date, DIF analysis has been conducted mainly for multiple-choice items. However, given the recent interest in performance assessment, it is of great importance that DIF be investigated with constructed response items.

This paper aims to investigate DIF across two different broad ethnic groupings, Asian and European, in a speaking test in which the test takers' responses are rated polytomously. Data in this study were collected from 1,038 non-native speakers of English from Thailand, Japan, Hong Kong, Spain, France, and Switzerland who took the SPEAK (Speaking Proficiency in English Assessment Kit) test in 1988, as part of a large-scale international study. The methods used for DIF analysis were the likelihood ratio test and the logistic regression procedure. The primary categories of interest in this study were 'pronunciation', 'grammar', and 'fluency.'

Preliminary results have shown that 'pronunciation' functioned differentially across the Asian and European test takers. A content analysis of the DIF items suggested that the types and the numbers of scales might influence the test validity. The results of this study suggest that both quantitative (statistical) and qualitative (content) analyses should be considered in detecting DIF items in language proficiency tests. The study also provides methodological information on differences between two approaches to DIF analysis. It also provides test developers with useful information on the effects of scoring scales.

Mikyung Kim
P.O. Box 951531
3300 Rolfe Hall
Hilgard Ave. Los Angeles
California 90095-1531
kimmi@ucla.edu

I am a second-year Ph.D. student in Applied Linguistics at the University of California, Los Angeles (UCLA). My specialty is language assessment in conjunction with bilingualism. Within this specialty, my primary interest is in developing speaking tests and analyzing polytomous scoring.

I received a B.A. in English Language & Literature (1994), and an M.A. in English Linguistics (1996) at Korea University. I obtained another MA in TESL (1998) at UCLA. My theses focused on the reading procedure of foreign language learners and assessment of the speaking ability of bilingual students in a two-way immersion program.

I taught English to middle school students in Korea, and developed test items in Korean for bilingual students in Los Angeles. I also have taught Korean language courses at UCLA. Currently, I work at the Center for the Study of Evaluation, Standards and Student Testing on issues of concern to the limited English proficient population.
Test Qualities: Expanding the Engineering Approach

Antony John Kunnan

California State University, Los Angeles

Abstract

The concept of test qualities was interpreted fairly uniformly though narrowly in the last three decades by focusing on reliability and validity. Oller (1979), Henning (1987) and Bachman (1990) emphasized the view that reliability and validity are "essential measurement qualities," and Davies (1977) and Alderson et al., (1995) argued that there are contrary pressures of reliability and validity. The Mental Measurement Yearbook too has over the decades evaluated tests mainly in terms of reliability and validity. While this conceptualization has made substantial contributions to test development and research, it is narrowly conceived and ideologically primarily utilitarian. This approach to test qualities could be called the "engineering approach". As a marked departure from this approach, Messick (1989) in his paper on 'Validity' argued that the facets of validity should include both evidential and consequential bases. Bachman and Palmer (1996) too presented a list of Test Usefulness qualities that included reliability and validity but added authenticity, interactiveness, impact and practicality. These papers clearly urge the field to go beyond reliability and validity and to employ approaches other than the engineering approach.

In this paper, I will attempt to bring together elements of both evidential and consequential bases by arguing for the primacy of fairness over reliability and validity and by presenting a concept of fairness within a framework of social justice for language testing research. As a starting point, I will use the Code of Fair Testing Practices in Education prepared by the Joint Committee on Testing Practices (1988) which presents standards for test developers and users in developing and selecting tests, interpreting scores, striving for fairness and informing test takers. In addition, relevant qualities such as access and justice inadmissible under the earlier approach will be considered. Overall, I will argue in this approach that the evaluation of language tests and testing practice should be conducted not only by studying test performance but also by studying the relationships among all stakeholders such as test takers, developers and users. This approach to test qualities could be called the "ecological approach". I will conclude by arguing that with an appropriate balance between the two approaches, the concept of test qualities could be enriched and that both approaches could benefit from closer contact with each other.

Antony John Kunnan
546 North San Marino Ave.
San Gabriel, CA 91775
USA
Fax: 323-343-5336
akunnan@calstatela.edu

Antony Kunnan teaches in the TESOL Program at California State University, Los Angeles. He is the Test Reviews Editor for Language Testing. He has authored and edited many publications in language testing.
Developing an Educational Assessment Framework
Jo Lewkowicz
English Centre, the University of Hong Kong

Abstract
Tertiary institutions in Hong Kong are under increasing pressure to demonstrate value for their educational dollars. This pressure is translated into a need to provide evidence that students are benefiting from language enhancement courses. These courses are typically short, designed to facilitate students’ academic study or to prepare students for the workplace. Since the needs of students vary across faculty, the content and aim of these courses often differ considerably.

Estimating language improvement as a result of instruction is fraught with difficulties. These difficulties are intensified when the time available for courses is too short to enable a marked improvement in general language proficiency. Yet, the providers of enhancement courses are required to demonstrate course effectiveness to a range of stakeholders including administrators and students. There is hence an obvious tension between assessment for accountability and to support learning. This tension is exacerbated when assessment instruments are not tailor-made and where a single instrument is intended for a range of purposes.

The session will briefly outline the work to date on developing an educational assessment framework which is being designed in response to the difficulties outlined above. It will highlight some of the problems being encountered in ensuring that the framework is able to address the need for accountability and at the same time ensure educational quality. It will also consider and seek advice on factors other than those of reliability and generalizability which need to be built into the framework.

Dr Jo Lewkowicz
English Centre
The University of Hong Kong
Pokfulam Road
Hong Kong
Fax: (852) 2818 3544
Email: jolewkow@hkusa.hku.hk
Assessment Criteria in a Large-scale Writing Test: What Do They Really Mean to the Raters?

Tom Lumley
Hong Kong Polytechnic University

Abstract

Recent research (e.g., Weigle 1994, Milanovic, Saville & Shuhong, 1996) into the assessment of ESL writing has employed think-aloud protocols (Ericsson & Simon 1993) to investigate a variety of issues related to rater behaviour, providing evidence that different groups of raters behave differently in various ways. However, what is lacking is a systematic investigation of the interpretations raters make of the criteria that are provided for them. The field of language testing is somewhat confused about how far it is reasonable to expect raters to actually agree with each other (Weigle 1998). Nevertheless, since one of the most commonly stated aims of training is to clarify interpretations of the assessment criteria, it seems reasonable to assume that a group of expert raters (trained, experienced and reliable), with relatively homogeneous backgrounds, should share common interpretations of the features included in a set of specified rating criteria. This study investigates this assumption, aiming to answer the question: Does a group of expert raters understand (rather than apply) given rating categories and accompanying level descriptions in ways that are a) similar to each other and b) consistent across performances?

The context for the study is the Special Test of English Proficiency, a large-scale test used to assist in immigration decisions in Australia. Each of 4 expert raters rated 24 samples of performance, and provided concurrent think-aloud protocols describing the rating process.

The findings have implications for:
- a) expectations of the training process;
- b) understanding the role of the rater in performance assessment;
- c) the validity of judgements and the use of scales to describe test performance.

References


Tom Lumley
Asian Centre for Language Assessment Research
Department of English
Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong
Fax: 852-2333-6569
eglmum@polyu.edu.hk

Tom Lumley is Assistant Professor at Hong Kong Polytechnic University, and an Associate of the recently established Asian Centre for Language Assessment Research there. Before that he was a Research Fellow at the Language Testing Research Centre, University of Melbourne, for a number of years. He has had extensive involvement in a wide range of language assessment research and development projects in numerous contexts in Australia and Asia. His research interests are in performance assessment, especially of writing, and in specific-purpose assessment, as well as in applications of Rasch measurement in language assessment. His doctoral studies combine several of these threads with analysis of verbal protocols to explore the process of rating from the perspective of the rater, in a large-scale test of English.
Multimedia Performance-Based Language Assessment:
The Computerized Oral Proficiency Instrument (COPI)
Valerie A. Malabonga and Dorry Kenyon
Center for Applied Linguistics

Poster Session

Abstract

The COPI is a multimedia, computer-administered adaptation of the tape-mediated Simulated Oral Proficiency Interview (SOPI). Both the SOPI and the COPI are oral proficiency tests based on common oral proficiency testing techniques used in the U.S. Oral proficiency tests like the SOPI and COPI use simulated real life tasks to elicit ratable speech.

Our goal is to make the COPI an improvement over the SOPI through harnessing the advantages offered by multimedia computer technology. We anticipate that these advantages will show in the following: more positive examinee and rater affect because of increased control over the testing and rating situation, improved testing efficiency with the availability of about a 100 tasks in a pool but the need for only 7-11 tasks per examinee, and increased rater efficiency through reduced scoring time. While procuring these advantages, we expect that examinees will obtain comparable scores across types of oral proficiency assessments.

Examinee and rater performance on the Spanish version will be compared with those of the SOPI and face-to-face interviews, whereas examinee and rater performance on the Arabic and Chinese versions will be compared with those of the SOPI. In the poster session, we will present the conceptual framework for the administration and scoring programs of the COPI and present the results of the above comparative studies. We will also have a demonstration of the two COPI programs available on a laptop computer.

Valerie A. Malabonga
4646 40th St. NW
Washington, DC 20016-1859
U.S.A.
Fax: (202) 362-3740
valerie@cal.org

Valerie A. Malabonga is a Research Associate at the Foreign Language Education and Testing Division of the Center for Applied Linguistics. She is the Coordinator for two technology projects: the Computerized Oral Proficiency Instrument and Web-Based Proficiency Tests.

Dorry Kenyon is director of Foreign Language Education and Testing at the Center for Applied Linguistics. His special research interests include the application of technology to performance-based language assessments.
Expanding the Construct of Reading Comprehension: Beyond Basic
Mary McGroarty and Latricia Trites
*Northern Arizona University*

**Abstract**

Most reading tests assess comprehension based on ability to answer multiple-choice questions, but this is one possible purpose for reading. Recent reading theory and research have begun to pay closer attention to the various purposes for reading and to development of assessment tasks that better reflect these purposes. Here we present the rationale for and operationalization of two additional constructs within reading comprehension in English: Reading to Learn (R-L) and Reading to Integrate (R-I). Based on research by Goldman (1997) and Perfetti (1997), we defined R-L as the ability to categorize information in a text and R-I as the ability to synthesize information from multiple texts.

To assess R-L, we asked students to read a 1,200-word text on air pollution and then complete a chart organizing the information according to macrostructures in the text. We expected that accurate categorization of textual information would demonstrate readers' comprehending of the logical structure of the text. Students' resulting charts were scored with a rubric based on accurate categorization of the macrostructures (problems, causes, effects, and solutions) and textual examples of each. Observed R-L scores ranged from 0-120. To assess R-I, we asked students to read two 600-word texts, both on water pollution, and write a prose synthesis based on the information in the texts. Current research suggests that such a task would demand the generation of a new representation, thus demonstrating ability to integrate new information with old. R-I synthesis essays were scored with a three part analytical rubric that awarded points for readers' ability to express integration, recognize and use macrostructures, and use relevant details, with observed scores ranging from 0-80. We tested these new tasks with 244 participants: 101 undergraduate native speakers of English (Group 1); 103 undergraduate nonnative speakers of English (Group 2); and 40 graduate nonnative speakers of English (Group 3). Order of testing was counterbalanced so that half of the participants received the R-L measure first, and half the R-I measure first.

Results were analyzed through a variety of statistical procedures, all aimed at distinguishing R-L and R-I from basic reading comprehension. We used analysis of variance to examine potential differences across groups and found that Groups 1 and 3 performed better than Group 2. Correlations indicated that, while all measures of basic comprehension (assessed through multiple choice questions) were strongly interrelated, correlation of basic comprehension with the new measures was only moderate, suggesting that the new tasks contributed some unique variance to the overall assessment of reading comprehension. Finally, we used regression to determine whether initial levels of basic comprehension would predict scores on the new measures differentially. Results thus far suggest that it is indeed possible to develop tasks that assess R-L and R-I reliably, and also that there is a threshold level of basic comprehension below which students are unable to perform either of the new tasks successfully. We conclude by discussing implications for the design and evaluation of new tasks to tap additional aspects of academic reading.

**References**


Assessment criteria for NNS and NS essays: Do uniform standards work?

Sally O'Hagan
The University of Melbourne

Abstract

This paper reports on the findings of a small-scale survey of the assessment criteria used by markers of essays written in English by NNS students in an Anglophone university. The literature on the experiences of NNS overseas students identifies several of the difficulties presented by studying in English as a second language (for example, Ballard 1993; Craswell 1992). Receiving less attention in educational research, are the experiences of the teachers of these students, as markers of essay assignments.

In the study reported here, a sample of 21 markers were surveyed to find out whether assessment criteria are applied in the same way to NNS essays, as they are to NS essays. The relative importance of these criteria— for NNS essays on the one hand, and NS essays on the other— was gleaned from descriptive statistics of markers' ratings of the importance of the criteria. From a subset of the sample, 3 markers' verbal reports of NNS essay marking protocols were collected using a 'think aloud' method. These were analysed qualitatively, comparing what was reported in the survey with what occurred in authentic marking situations.

The paper will discuss the assessment of NNS essays in relation to the question: in what ways are the assessment criteria for NNS essays different from those for NS essays? While the study found that, in general, the criteria are the same, it found a tendency for markers to show leniency on some criteria for NNS essays. Moreover, markers' reported feelings of uncertainty about the legitimacy of this response, suggest a conflict between markers' actual responses to NNS essays, and their feelings about how they should respond. The implications of such a conflict will be considered in terms of the growing imperative—in the context of an 'internationalised' higher education environment—for teachers to be able to engage confidently with a linguistically diverse student body.

c/o Language Testing Research Centre
Level 2, 138-146 Cardigan St
Carlton VIC 3052
AUSTRALIA
FAX: +61 3 9344 5163
s.ohagan@linguistics.unimelb.edu.au
Gender Bias in Oral Proficiency Testing
Kieran O'Loughlin University of Melbourne

Abstract
To date, the important issue of gender bias has received limited attention in language testing research. It is possible in oral interviews, for instance, that both interviewing and rating may be highly gendered processes. In tests like the IELTS (International English Language Testing System) interview, where the interviewer also acts as the rater, this poses the question of whether gender bias, if it exists, stems from the interview itself, the rating decision or a combination of both these 'events'.

The data collected for this study consisted of the audio-taped performances of 8 female and 8 male test takers who undertook a practice IELTS interview on two different occasions, once with a female interviewer and once with a male interviewer. The interviews were transcribed and analysed in relation to features of language use which have been identified in previous research as key markers of gendered communication. These included the use of overlaps, interruptions and minimal responses by both interviewers and test takers. The scores subsequently assigned by 4 raters (2 male and 2 female) to each of the 32 interviews were also examined in relation to the gender of both raters and test takers using multifaceted Rasch bias analyses.

The results of the discourse and test scores analyses will be reported in detail and the findings interpreted in relation to more recent theoretical understandings about gender in the field of gender studies and new empirical work being done in a sub-branch of applied linguistics, namely language & gender.

Kieran O'Loughlin
Department of LLAE
5th Floor, Doug McDonell Building
The University of Melbourne
Parkville
Victoria
Australia 3052
FAX: +61 3 9344-8612
k.oloughlin@edfac.unimelb.edu.au

Kieran O'Loughlin is Convenor of TESOL teacher education in the Department of Language, Literacy & Arts Education at the University of Melbourne. He is also an Associate of the Language Testing Research Centre at the University of Melbourne. His current research interests focus on social identity in language assessment as well as in language learning and teaching.
Using Observation Checklists to Validate Speaking-Test Tasks
Barry O'Sullivan  
University of Reading

and  
Nick Saville  
University of Cambridge Local Examinations Syndicate

Abstract

Milanovic and Saville's (1996) overview of the variables which interact in performance testing - influential in revisions to the Speaking components of Cambridge examinations - implies that particular concern be paid to the design of these tests, in terms of i) possible interaction patterns between speakers and ii) the tasks through which language is elicited. Increased interest in the testing of language ability using performance-based methodologies has led to a range of approaches to test validation, both quantitative and qualitative. This paper reports on the development and use of observation checklists in the validation of the face-to-face, multiple-task speaking tests within the Cambridge examination system, focusing on their application within the Cambridge Certificate of Proficiency English (CPE) revision project. The approach to task classification adapted here builds on previous studies (e.g. FCE 1996 revision) and on Weir's (1993) task categorisation framework, which sees tasks in terms of their operations/functions and conditions. Three Senior Examiners first used the checklists on a sample representative of all levels of the Cambridge system. Feedback led to revisions to both the checklist and procedures for its application. Speech samples from 20 candidates from 10 CPE speaking tests were selected to reflect a variety of language backgrounds and with varying language proficiency around the CPE level. These were analysed using the checklists by a group of 30 trained examiners. Results are reported and the implications of this study for task/test design and for the a priori and a posteriori validation of tests which involve an evaluation of performance on such tasks are discussed.

References


Barry O'Sullivan  
Course Administrator  
Centre for Applied Language Studies  
The University of Reading  
PO Box 241  
Reading  
RG6 6WB  
United Kingdom  
Fax: **44 118 975 6506**  
B.E.0Sullivan@reading.ac.uk

Barry O'Sullivan is at present based at the Testing and Evaluation Unit, CALS, University of Reading, UK. He has previously worked in Ireland, the UK, South America and Japan in EFL and teacher education, and has been involved in language assessment since 1993, and has written and presented a number of papers in this area. His PhD studies are concerned with variables affecting oral performance in language tests. He worked as a Visiting Consultant with ACLAR at the Hong Kong Polytechnic University, working on the GSLPA project between February and August 1999 and is presently working on an EAP speaking test development project at CALS as well as his involved with the UCLES CPE revision project.
Nick Saville has over 20 years experience in the field of linguistics, English language teaching and language testing. After completing a degree in Linguistics, he worked at a university in Italy where he gained extensive practical experience in setting and administering language tests. He has a masters degree in TEFL from Reading University with a dissertation in the field of language testing. In 1987, he began working for the University of Cambridge Local Examinations Syndicate (UCLES) and since 1994 he has been Group Manager for Test Development and Validation and is an Assistant Director for UCLES EFL. His own research interests are in the field of performance testing and in particular the development and validation of rating scales for testing spoken language.
Assessment of Dimensionality of Language test data
Serena Pyo
University of Illinois at Urbana-Champaign

Poster Session

Abstract

The connection between constructs and test dimensionality is not clear-cut in language testing area. Unidimensionality has been interpreted as 'conceptual rather than factual, qualitative rather than quantitative' (Wright and Linacre, 1989), and 'a matter of degree' (Choi and Bachman, 1992). Several different approaches and criteria for checking dimensionality were used and inconsistent results on the number of dimensions were obtained in Blais and Laurier (1995) and Choi and Bachman (1992) studies.

In the past the dimensionality of a test has been assessed in a wide variety of ways in the general measurement field. Hattie's (1985) list of various techniques for checking dimensionality is impressive, but researchers found none of them fully satisfactory on different conditions of a test. Likewise, many simulation studies conducted in the general methodology area to evaluate the dependability of various indexes of dimensionality stress caution when generalizing results to other conditions (test length, sample size, latent correlation, and so on) (De Champlain & Gessaroli, 1998; Gessaroli & De Champlain, 1996; Hattie et al., 1996; Nandakumar, 1994).

The purpose of this study is to examine the behavior of three dimensionality procedures (nonparametric approach (Stout's procedure), Holland and Rosenbaum's approach, and nonlinear factor analysis (NOHARM)) when applied in the language testing area. The sensitivity and accuracy of the procedures (empirical type I error rates and rejection rates for uni- and bi-dimensionality) is examined with data sets simulated to reflect various language test characteristics and conditions (i.e., extremely high correlation between traits).

Each method is examined and compared with other methods first on simulated data sets from a variety of conditions to determine their validity and reliability as a tool for dimensionality check and then on real language data sets. Data sets with 24 different conditions are generated: six unidimensional and eighteen two-dimensional data sets. One hundred data sets are generated under each condition for the purpose of statistical comparisons among three approaches. Three levels of correlation between abilities are considered: 0.5, 0.65, and 0.8, three levels of number of examinees: 200, 500, and 1000, and two levels of number of items: 20, and 40. Four different real data sets are examined: FCE and TOEFL (1990), MELAB (1994), and the UTUCEPT (1999).

The significance of this study would be (1) to demonstrate the pitfalls in indiscriminately applying approaches to language testing data and (2) to direct language testers toward a possibly unified view on the dimensionality issue rather than a hasty conclusion on the structure of the tests without questioning the utilities of the methods utilized. Although the results of the present simulation analyses with various approaches are restricted to the conditions tested, a more clear-cut interpretation of dimension(s) questioned in the construct validation would be possible than ever before. Then, language testers would be equipped with an insight on how to deal with the inconsistent and opaque findings made in the assessment of dimensionality from different perspectives.

References


Serena Pyo
303 E.Clark St. #1
Champaign, IL 61820
USA
k-pyo@uiuc.edu

Serena Pyo is a 4th year student in the Quantitative and Evaluative Research Methodologies (QUERIES) Ph.D. program at the Dept. of Educational Psychology at UIUC.
Web-Based Language Testing: Opportunities and Challenges
Carsten Roever
University of Hawai'i at Manoa

Abstract

Web-based language testing is an extension of computer-based language testing to the world wide web. While traditional computer-based tests require extensive programming expertise and specially designed delivery platforms, web-based tests (WBTs) are written in HTML, which can be produced with an editor and only requires a standard web browser for display. Media, such as sound files and video files, can be easily integrated in the test, and scoring engines written in JavaScript can automatically score test taker responses on dichotomous items.

It is important to note that test security and item confidentiality will be extremely low if maximum use is to be made of the flexibility and "anytime, anyplace" accessibility of WBTs. They are therefore most appropriate as low or no-stakes instruments, e.g., for self-assessment, feedback to learners on degree of mastery of a curricular objective, or as control mechanisms in self-directed learning. Web technology can also be used for the design and delivery of medium or high stakes tests (placement test, admissions tests, certification tests), but must be accompanied by secure and standardized testing conditions, as well as secure data transfer and storage.

Examples of web-based item types will be used throughout this presentation and their programming code will be made available to participants.

Carsten Roever
Dept of ESL
University of Hawai'i at Manoa
1890 East-West Rd
Honolulu, HI 96822
Fax: (419) 821-3591
roever@hawaii.edu

Ph.D. student (ABD) in the Ph.D. Program in Second Language Acquisition at the University of Hawai'i, M.A. from Duisburg University in Germany. Research interests are web-based assessment, assessment of pragmatics, and assessment in the service of learning. Working on a dissertation on web-based assessment of second language pragmatics, and a monograph on foundations of web-based language testing.
The Journey toward Standardization:
The University of British Columbia, English Language Institute Experience
Joy Lin Salzberg, Pat Burn, Pegram Deupree, Lori Hargreaves,
Elaine Klein, Laurel Saba, and Joyce White
English Language Institute, University of British Columbia

Abstract
The English Language Institute at the University of British Columbia offers integrated content and skill based Intensive English Program (IEP) courses in 12 week sessions. It has been working over the last two years on student placement, in-class, and exit assessment. This poster session displays the process followed as teachers negotiated language assessment criteria and standards implemented in IEP courses. Placement testing is used initially, while exit testing is used only at the highest level for university entry and granting of level certificates. Movement between levels is determined by teacher-developed multiple assessment measures administered in classes. The journey toward standardization has involved teacher teams developing various language assessment measures addressing matters such as core tasks, course criteria and weighting, rating scales and checklists, all in light of curriculum objectives. This challenging process involves issues of reliability, validity and replicability.

This poster will be of benefit to teachers and administrators concerned with Intensive English Programs. The process has revealed the gains to both students and teachers in developing authentic assessment measures that fit course content and objectives.

Joy Lin Salzberg, Assessment Head Teacher
English Language Institute
University of British Columbia
2121 West Mall
Vancouver, B.C. Canada V6T 1Z4
Fax: (604) 822-1599
salzberg@mail.cstudies.ubc.ca

Joy Lin Salzberg is the Assessement Head Teacher at the English Language Institute, UBC. She hold a MA degree in TESOL from the University of British Columbia. Since the implementation of the Intensive English Program at the English Language Institute in 1998, Joy has worked extensively on the development and standardization of student placement, in-class and exit assessment.

Pat Burn is an instructor at the English Language Institute, UBC. In addition to having taught many IEP courses at different levels, Pat has been instrumental in the recent revision of the IEP curriculum at the English Language Institute. Among her other interests are writing novels and travelling in China and Tibet.

Pegram Deupree is a senior instructor at the English Language Institute, UBC. After 22 years of teaching, she has chosen writing and reading as favorite subjects. She has been instrumental in the teacher development of the IEP curriculum and assessment.

Lori Hargreaves - English Language Institute, University of British Columbia She holds a MA degree in TESOL and in Theology. She has done field research in South Africa and just recently returned from teaching in Singapore. She is interested the assessment not only of students but of teachers and programs as well.

Elaine Klein is an instructor at English Language Institute, UBC. She is particularly interested in reading and has been a leader in development of materials for intermediate levels of reading and writing.
Laurel Saba is an instructor at the English Language Institute, UBC. She loves to teach speaking and listening classes and has currently pioneered assessment techniques for beginning level writing students in IEP. She has developed assessment tools that use stories to measure fluency in speaking and listening.

Joyce White is an instructor at the English Language Institute. She holds her MA degree in TESOL with a specialty in writing. She is one of the initial team designers for the ELI’s IEP curriculum. She currently gives training in teaching writing, while teaching writing classes as well.
Test Takers’ Performance on and Affective Responses to Paper-and-Pencil and Computer-Administered Japanese Reading Tests
Yasuyo Sawaki
University of California, Los Angeles

Abstract
Interest in computer-administered tests of reading in a second/foreign language has been rapidly increasing in the past decade. Because of the potential differential effects of different modes of presentation, it is vital to investigate the comparability of paper-and-pencil and computer-administered reading test results. That is, the extent to which our test-based inferences of reading ability generalize to examinees’ performance on reading tasks in their target language use domains is an important aspect of construct validity in computer-administered reading tests. The present study will investigate examinees’ performance on and affective responses to reading comprehension tasks on a pilot version of the Japanese web-based placement test currently under development at the University of California, Los Angeles.

Pilot versions of the test—paper-and-pencil and web-based—will be administered to a group of university-level learners of Japanese as a foreign language in two sessions. The subjects will take a preliminary paper-and-pencil version in Session One in January, 2000, and another version of the test in both paper-and-pencil and web-based forms in Session Two later in the spring, 2000, when the web-based form becomes available to this study.

The subjects’ think-aloud verbal protocols about the reading comprehension tasks will be qualitatively and quantitatively analyzed and compared across modes of presentation, in terms of: (1) outcome variables (test completion time and reading comprehension); (2) task completion process (pattern of text and task processing, and reading and test taking strategy use); and (3) examinees’ affective responses to the reading test tasks.

In this session, I will discuss the following: (1) the results of the verbal protocol analysis of data collected for the paper-and-pencil version in Session One, (2) plans for data collection on the web-based version during Session Two and (3) plans for further data analysis procedures.

Yasuyo Sawaki
Dept. of Applied Linguistics & TESL
University of California, Los Angeles
P.O.Box 951531 3300 Rolfe Hall
405 Hilgard Avenue
Los Angeles, CA 90095-1531
U.S.A.
Fax: (310) 206-4118
ysawaki@ucla.edu

Yasuyo Sawaki is a Doctoral student at the Department of Applied Linguistics/TESL, University of California, Los Angeles.
Native and Nonnative Speaking EFL Teachers' Evaluation of Chinese Students' Writing

Ling Shi
University of British Columbia

Abstract
This study examines the differences between native and nonnative EFL (English as a Foreign Language) teachers' ratings of the English writing of Chinese university students. It aims to explore whether native and nonnative teachers 1) give similar scores to the same writing and 2) use the same criteria in their judgments. Forty-six teachers, 23 Chinese and 23 English, rated 10 expository essays using a 10-point scale, then wrote and ranked three reasons for their ratings. The reported reasons were coded as positive or negative criteria under five major categories: general, content, organization, language, and length. MANOVA showed no significant differences between the two groups in their scores for the ten essays. Chi-square tests, however, showed that the English teachers tended more frequently and positively in their criteria to the content and language whereas the Chinese teachers more frequently but negatively to the organization and length of the essays. The Chinese teachers were also found to be more concerned with content and organization in their first criteria whereas English teachers more on language in their third criteria. The results raise questions about the underlying differences between native and nonnative EFL teachers in their instructional goals of L2 writing and the importance of cooperation between the two groups.

Ling Shi
Department of Language Education
2034 Lower Mall Road
UBC, Vancouver, B.C.
Canada V6T 1Z2
Fax: 822-3154
ling.shi@ubc.ca

Ling Shi is an Assistant Professor at the University of British Columbia. She holds a Ph.D in Curriculum Studies from Ontario Institute for Studies in Education. She has also worked in the field of tertiary level EAP and ESP in Hong Kong and Mainland China.
Software Demonstration for Creating Computer-Delivered/Human-Rated Speaking Exams
Diane Strong-Krause and Troy Cox
Brigham Young University

Poster Session

Abstract

Software for creating computer-delivered/human rated speaking exams will be demonstrated. The software is composed of three modules: The Test Preparation Module, the Test Administration Module, and the Test Assessment Module. A demonstration test will include use of audio, pictures, and video for use in speaking prompts.

Summary

Most language teachers are concerned with assessing speaking proficiency. However, face-to-face interviews are not always practical in terms of testing and training time. Some have opted for tape-mediated simulated oral proficiency tests (such as the SOPI). This delivery method, however, is limited because of the linear nature of the test and the type of media that can be used in the test prompts.

A natural outgrowth of simulated oral proficiency exams is the use of computer technology to deliver the exams. This technology allows more flexibility in presenting speaking tasks including using video, having multiple test forms in one exam, and building tests that more closely match the language proficiency of the examinee.

In this poster session, a computer-delivered speaking test will be demonstrated. The testing software is composed of three modules: 1) the Test Preparation Module is a template that allows the teacher to build a computer-administered oral test; 2) the Test Administration Module administers the test to the student; and 3) the Test Assessment Module facilitates assessing the student responses.

The software allows great flexibility in creating tasks and designing exams that target particular language levels. In addition, since responses are stored as sound files, raters are able judge responses more efficiently.

Diane Strong-Krause
Brigham Young University Linguistics Department
2144 JKB
Provo, UT 84602
FAX: (801) 378-8295
diane_strong-krause@byu.edu

Diane Strong-Krause, an assistant lecturer in the Linguistics Department at Brigham Young University, works with ESL testing, teaching, and teacher training.

Troy Cox received his M.A. in TESL from Brigham Young University and is currently teaching and working on projects related to teaching ESL and technology at the BYU English Language Center.
From Rating to Assessment: The Rater Perspective on Assessing Writing in Finnish as a Second Language

Mirja Tarnanen
Center for Applied Language Studies, University of Jyväskylä

Abstract

In the 1990s, several studies have appeared on rater behaviour. The need to know more about the way that the people making the assessments relate to the process and outcomes of assessment has been noted e.g. by Vaughan 1991, Connor-Linton 1995, Upshur & Tuner 1999. It is obvious that the assessment process is complex, but it is less clear how well the different variables involved in the assessment process can, or should, be explained and interpreted as measurement error.

This poster presents an attempt to describe judges' behaviour by combining a quantitative and a qualitative research method. The purpose of the study is to investigate the nature of intra- and inter-rater consistency. The assessors in the study used two kinds of scales, one without level descriptors and one with level descriptors, in assessing writing performances in Finnish as a second language. The quantitative approach investigates the consistency numerically. The qualitative approach makes use of the arguments, beliefs and ideas reported by the judges. The rater comments relate to aspects in the rating process such as the type of scale they are using, the task type that they are rating, and their feelings during the rating process.

In the present study, the writing performances were collected from learners of Finnish as a 2nd language (N=258) representing different linguistic and cultural backgrounds. Each of them completed three tasks from a set of nine tasks in a matrix design, such that one task was common to all writers. The raters (N=17) were teachers of Finnish as a second language, all of whom rated performances from at least 100 writers. There were three rating rounds. In the first round, they used a numerical scale from one to six, without any level descriptors. In the second rating round, they used a six-level scale with descriptors (based on the Council of Europe scale). The third rating round was a replication of the second round. After the first and the second rating round, the raters were interviewed using a semi-structured interview. They also completed a questionnaire on the rating process. After the third rating round, the raters participated in a repertory grid interview. This procedure was completed by all raters.

This poster explores the links between the quantitative and the qualitative results. It discusses views on consistency and inconsistency, especially the way these are accounted for by the interview data. The raters' perspectives on the rating process and the target skill are discussed in the context of the numerical results on how well the raters' ratings agree.

Mirja Tarnanen
Center for Applied Language Studies
University of Jyväskylä
PO BOX 35
40351 Jyväskylä
Finland
Fax: +358 14 2603521
tarnanen@cc.jyu.fi

I am a Ph.D. student at the University of Jyväskylä in Finland. My current research interests in this field are rater behaviour and more generally writing skill assessment. I am also interested in the possibilities of the use of information technology in language learning. During my working history I have been involved in the development of the Finnish national testing system (National Certificate) and the joint European language testing project (DIALANG). I also work as a teacher of Finnish as a 2nd language.
A Cognitive-Psychometric Approach to Construct Validation of Web-Based Language Assessment

Lihshing Wang, Lyle Bachman, Nathan Carr, Greg Kamei, Mikyung Kim, Lorena Llosa, Yasuyo Sawaki, Sang Keun Shin, Viphavee Vongpumvitch, Xiaoming Xi and David Yessis
University of California, Los Angeles

Work in Progress

Abstract

As part of the Web-Based Language Placement Examination Project at a large university in the US, this study examines some of the critical issues involved in measuring language ability with the latest advances in cognitive-psychometric theories. Among the issues being explored are: (1) Behavioral Scaling for categorizing competency levels in accordance with the ACTFL guidelines; (2) Diagnostic Profiling for identifying differential sub-skill levels for instructional purposes; (3) Latent-Class Modeling for classifying examinees into one of several ordinal categories for placement purposes; (4) Response Generative Modeling for automated item generation from a specified construct universe; and (5) Optimization algorithms for maximizing measurement efficiency during automated test assembly.

With an aim at developing a methodological framework generically applicable to different languages, this study proposes to attain the following objectives:

1. Based on a critical review of the psychometric components of existing computerized language assessment systems, and in consultation with the existing guidelines for computerized testing, this study proposes a recommended procedure for developing a Web-based language assessment program.

2. Integrating recent advancements in cognitive-psychometric theories, this study outlines promising research directions for implementing intelligent language assessment systems on the Web.

3. To provide empirical evidence for the feasibility of cognitive-psychometric approach to Web-based language assessment, this study develops an assessment prototype which implements at least one of the five modeling procedures as outlined above.

This innovative paradigm of cognitive-psychometric modeling provides language testing researchers with unprecedented access to language assessment integrating both quantitative and qualitative information in the assessment process. An important implication of such integration is that, by demonstrating statistically optimal fit between model and data, the researcher has obtained evidence of construct validity.

Lihshing Wang
3300 Rolfe Hall
University of California Los Angeles
Los Angeles, CA 90095-1531
Fax: (310) 206-3432
lswang@ucla.edu

Lihshing Wang is Professor of Psychology at National Chung Cheng University in Taiwan, and Visiting Professor of Applied Linguistics & TESL and Education at University of California Los Angeles, USA. Her research interests are Item Response Theory, cognitive-psychometric modeling, and intelligent assessment.

Lyle F. Bachman is Professor of Applied Linguistics & TESL at University of California Los Angeles and Editor of Language Testing Journal. His research interests are language assessment, second language acquisition, and language program evaluation.

Nathan Carr, Mikyung Kim, Greg Kamei, Lorena Llosa, Yasuyo Sawaki, Sang Keun Shin, Viphavee Vongpumvitch, Xiaoming Xi and David B. Yessis are graduate students of Applied Linguistics & TESL at the University of California, Los Angeles.
Investigation of Size of Vocabulary and Lexical Richness in Candidate Output on a Semi-Direct Oral Proficiency Test
Jessica Wu
The Language Training & Testing Center, Taipei

Work in Progress

Abstract
Vocabulary is often recognized as necessary in effective communications in a foreign language. Vocabulary knowledge is therefore considered a critical factor in the assessment of learners' language proficiency. This research aims to investigate lexical richness in candidates' output with special regard to vocabulary size on a semi-direct oral test. The investigation was conducted by using two measurement instruments – 1. VocabProfile program, aiming to quantify total words and word types used by candidates in the oral test, and 2. Collins Cobuild's five frequency band word lists, aiming to find out candidates' lexical richness by observing how the words used were distributed in the word lists. The findings of the research provide reports on the size of vocabulary in candidates at different proficiency levels. In addition, the size of vocabulary is interpreted by viewing the vocabulary as consisting of a series of levels based on frequency of occurrence (Laufer & Nation, 1999).

Jessica Wu
The Language Training & Testing Center
170 Hsin-Hai Road Section 2
Taipei, Taiwan
Republic of China 106
Fax: 886-2-2367-1944
jw@ltc.ntu.edu.tw

Jessica Wu is currently Testing Program Officer at The Language Training & Testing Center, Taipei, Taiwan. She has actively involved in a five-level English proficiency test development project in her country. She has an MA in education from The Ohio State University.
A Sociological Dimension in Tests of English for Immigrant Teachers
Elaine Wylie
Griffith University

Work in Progress

Abstract

For overseas-trained teachers who are not native speakers of English to practise in Australian schools, they must reach certain levels of English proficiency. Since 1983 the Queensland Board of Teacher Registration (BTR) has required a writing level of 3+ on the International Second Language Proficiency Ratings (ISLPR). In this system the yardstick is the performance of the candidate's native-speaking sociocultural peers, and testing for the BTR has assumed that candidates' peers are highly literate in a standard variety of English. However, as public criticism of education focuses increasingly on declining literacy standards of (native-speaking) teachers, issues of validity and equity in tests for non-native speaking teachers must be revisited.

In a pilot study, typical BTR tests were administered to five volunteer native speakers of English about to graduate as teachers. Their scripts and those of seven actual test candidates (two rated as 4, three as 3+ and two as 3), were typed up and other distinguishing features removed. They were then presented in random order to eight informants (BTR officers, practising teachers, teacher educators, and parents), who were told simply that the writers were trained teachers. Informants were asked to consider the scripts in terms of their expectations of a teacher's writing, responding in an audiotaped interview.

Responses analysed in terms of theories of L2 development and of educational sociology (e.g. Bourdieu's work on cultural reproduction and Bernstein's 1996 writings on pedagogic codes) will be presented. The presenter will welcome suggestions about an appropriate quantitative complement for the qualitative data.

Elaine Wylie
Centre for Applied Linguistics and Languages
Griffith University
Nathan
Queensland
Australia 4111
Fax: 61 7 3875 7090
E.Wylie@mailbox.gu.edu.au

Elaine Wylie is a Senior Lecturer in the Centre for Applied Linguistics and Languages at Griffith University, Australia. Her main research interest is the development of proficiency scales and she is the co-author, with David Ingram, of a number of specified and non-specified purpose versions of the International Second Language Proficiency Ratings (ISLPR). She has recently been awarded a major grant from the Australian Government through the Committee for University Teaching and Staff Development (CUTSD) to improve proficiency outcomes in languages other than English programs, using the ISLPR as a curriculum framework. She has been involved in examiner training, research, and supervision of test writing for IELTS and a number of other test development and evaluation projects. She teaches postgraduate courses in language assessment and runs a language testing service.
The National College English Test

Organizer: Yang Huizhong National College English Testing Committee of China and Shanghai Jiao Tong University

Paper 1: The Validation Study of the National College English Test
Yang Huizhong
National College English Testing Committee of China and Shanghai Jiao Tong University

Abstract
The College English Test (CET) is a national test administered on behalf of the Ministry of Education of China, with 3 million students taking it every year. The CET is a large-scale standardized test, and is a criterion-related norm-referenced test by nature. Series of measures have been taken to ensure that the CET meets the quality requirements of a large-scale standardized test, including the objectivity of scoring, quality control of item development, score equating and score interpretability, computer adjustment to filter out subjectivity in essay marking, and positive backwash effect on teaching, etc. A three year "CET Validation Study Project", a Sino-British joint project, was carried out. The aims of the study include:

- to validate the CET test: whether it measures the students' English language proficiency as it claims;
- to further improve the backwash effect of the CET on teaching;
- to improve the test format and adopt new item forms, if necessary;
- to check the central statistical package developed for data processing of the CET;
- to promote the theoretical development of large-scale standardized testing in China.

The scope of the study covers construct validity, content validity, concurrent validity, predictive validity, and face validity. The means of the study include questionnaire survey, introspection study, experiments, parallel tests, computer analysis of testing materials, statistical analysis of huge piles of data, etc. Findings, which fully support the high validity of the CET test, are presented and discussed in this paper. New measures based on the research findings, including new item forms, have been taken in the national CET test.

Paper 2: An Introspection Study of the Validity of the CET Reading Comprehension Test
Jin Yan
National College English Testing Committee of China and Shanghai Jiao Tong University

Abstract
As part of the Sino-British cooperative College English Test (CET) Validation Project, this study set about to investigate CET test takers' reading behavior using the process-oriented introspection technique. In the study, data were collected through students’ think-aloud verbal reports produced in the process of completing the tasks of the reading component of the Prototype College English Test. The transcription and analysis of the data, though extremely time consuming, enabled us to say that introspection is a very useful research tool in investigating the nature of the reading processes involved in taking a reading test. The analysis of the data also enabled us to tell the skills and strategies employed by the students in reaching answers to test items. Therefore, an introspection technique can be considered as a genuine process measure that provides an interesting insight into how students approach the test tasks and the reading materials. The study therefore supports the view that the introspection technique has advantages over the traditional statistical methods of validation which are largely product-oriented and usually termed as the quantitative analysis. More importantly, the study also proves that CET reading comprehension test has high validity in the sense that it is really testing reading comprehension not test-taking strategies.
Paper 3: Towards Improving Marking Reliability in the CET Test
Yang Huizhong and Xia Guozuo
National College English Testing Committee of China and Shanghai Jiao Tong University
National College English Testing Committee of China and Fu Dan University

Abstract
This paper describes the measures taken to improve marking reliability for China's National College English Test. In a large-scale test such as the CET test, which averages 3,000,000 testees every year, MC items are unavoidable for machine reading. However, the backwash effect of MC items on teaching is questionable. To improve the backwash effect, writing as a means of measuring the testees' productive ability is purposefully included in the CET. While it does help the teacher to focus his/her attention on improving the students' English ability rather than on test taking strategies, the marking of writing is subject to the marker's personal judgment, which reduces the reliability of the writing score. Measures have been taken to ensure intra-marker, inter-marker, and inter-centre consistency. A holistic approach is adopted, supplemented by a computer adjustment procedure to filter out the marker's subjective bias, based on the hypothesis of high correlation of the total testee population's performance on objective items and on subjective items. There are altogether six marking centres in the country to mark the large number of the writing papers. An explicit marking scheme is provided to each marker, together with five range-finders, selected carefully from among the testees' actual scripts, for marker training on the first day. The intra-marker consistency is supervised by the chief marker in each group. The results of three recent independent experiments show that this approach has a fairly high reliability. The CET will continue to use the present approach before any other more satisfying approach in marking is known.

Paper 4: The CET-SET Test --- Testing English Speaking Abilities ---
Yang Huizhong and Guo Jieke
National College English Testing Committee of China and Shanghai Jiao Tong University
National College English Testing Committee of China and South-China University of Science and Technology

Abstract
The College English Test (CET) is a national test administered on behalf of the Ministry of Education of China, with 3 million students taking it every year. The CET is a large-scale standardized test, and is a criterion-related norm-referenced test by nature. With 12 years of development, the CET is now well established. As a result of the open-to-the-outside-world policy, there has been a growing need of spoken English for the rapidly increasing international communication. The CET-SET has therefore been included as a component part of the CET test. It will be administered only on applicants already qualified in the CET-6 written test with a score of better than 80, or in the CET-4 written test with a score of better than 85, on a 100-point score scale, to limit the number of testees to a manageable size. This paper reports the results of a four year research project which completed all the designing aspects of the CET Spoken English Test, including the test format, the marking scheme, the test specifications, etc. The results of the pilot test show that the reliability and validity of the CET-SET can be ensured. A questionnaire survey shows that the CET-SET will provide very strong positive backwash on classroom teaching.
Professor Yang Huizhong
Chairman
National College English Testing Committee of China
Shanghai Jiao Tong University
1954, Hua Shan Road
2203 HaoRan High Tech Building
Shanghai P.O. Box 30-14
Shanghai 200030
China, P. R.
Fax: 0086-21-6495 6421
hzyang@mail.sjtu.edu.cn

Dr. Jin Yan
Director, Administration Office
National College English Testing Committee of China
Shanghai Jiao Tong University
1954, Hua Shan Road
2203 HaoRan High Tech Building
Shanghai P.O. Box 30-14
Shanghai 200030
China, P. R.
Fax: 0086-21-6282 6622

Professor Xia Guozuo
Vice Chairman
National College English Testing Committee of China
Fu Dan University
Department of Foreign Languages
Shanghai 200433
China, P. R.
Fax: 0086-21-6282 6622

Professor Guo Jieke
Vice Chairman
National College English Testing Committee of China
South-China University of Science and Technology
Department of Foreign Languages
Guangzhou 510 641
China, P. R.
Fax: 0086-20-8551 1446
e-mail: fjkguo@scut.edu.cn
The Test of English as a Foreign Language (TOEFL)
of
Educational Testing Service (ETS)

Congratulations

Dr. Angela Hasselgren
Department of English
University of Bergen
Bergen, Norway

Winner of the
TOEFL Doctoral Dissertation Research Award
for 1999-2000

This award recognizes doctoral dissertation research that makes
a significant and original contribution to knowledge about and/or
use of second or foreign language tests and testing. It brings the
recipient $2,500 and fully supports the cost of attendance at the
Language Testing Research Colloquium.

For more information about the award,
visit the TOEFL display in the exhibition
area or the TOEFL web site at
www.toefl.org
Which of the following is "THE Michigan Test"?
A. The MELAB  
B. The MTELP  
C. The ECPE  
D. None of the above

The answer is D. None of the above!

In fact, the English Language Institute at the University of Michigan offers a variety of ESL/EFL tests and testing programs . . .

**MELAB Program**
*Michigan English Language Assessment Battery*
✓ for applicants to English-medium universities
✓ for professional certification or for immigration
✓ individual & group exams in U.S. & Canada
✓ a secure test battery with measures of writing, listening, reading, optional speaking test
✓ intermediate to advanced proficiency level

**EAP EVALUATION**
*AEE (GTASE, UAWA), IGSI*
✓ specialized tests for UM graduate and undergraduate students
✓ oral performance testing of prospective graduate student instructors

**TEST PUBLICATIONS**
*EPT, LCT, MTELP*
✓ commercially available to educational institutions for in-house use
✓ for placement in intensive programs and for academic readiness

**CERTIFICATE EXAMINATION PROGRAMS**
✓ for English language students outside the USA
✓ documentary evidence of language proficiency for education, employment, or job promotion
✓ group exams administered once annually in approx. 20 countries
✓ secure test batteries with measures of writing, listening, reading, speaking

**ECPE**
*Examination for the Certificate of Proficiency in English*
✓ advanced proficiency level
✓ emphasis on range & accuracy

**ECCE**
*Examination for the Certificate of Competency in English*
✓ upper intermediate proficiency level
✓ focus on communicative effectiveness

For further information about all ELI-UM Testing Programs, see the website or contact:

English Language Institute, University of Michigan
Testing & Certification Division
North University Building
Ann Arbor, MI 48109-1057 USA

http://wwwlsa.umich.edu/eli/testing/
Grammar for the 21st century

“For the foreseeable future, anyone with a serious interest in English grammar will have to take into account the information this book contains.”

David Crystal

- An entirely corpus-based grammar of English
- 6,000 authentic examples are used from the Longman Corpus Network
- Reveals the differences between spoken and written English
- Based on a ground-breaking six year research project carried out by an internationally renowned author team

The Longman Grammar of Spoken and Written English
Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan
Cased 0-582-23725-4 1216 pages

Pearson Education ESL
10 Bank Street  Suite 900
White Plains, NY 10606-1951
1-800-375-2375
www.longman-elt.com/dictionaries

For more information please contact your local ESL Sales Specialist
In support of the European Year of Languages 2001

ALTE

The Association of Language Testers in Europe

is organising a conference

‘European testing issues in a global context’

12 to 14 July
Universitat Pompeu Fabra, Barcelona

For further information on making a presentation at or participating in the conference, please contact:

ALTE Secretariat, 1 Hills Road, Cambridge, CB2 1EU, UK
Tel: +44 1223 553925, Fax: +44 1223 553036
e-mail: hirtzel.m@ucl.es.org.uk
Assessing Second and Foreign Language Skills

This series of six short videos presents a rich and accessible introduction to the theory and practice of language assessment for postgraduate students in Applied Linguistics and TESOL, for trainee second or foreign language teachers, and for experienced language teachers who wish to improve their practical skills and theoretical knowledge of assessment.

Viewers are introduced to the key concepts and terms associated with language assessment through interviews with leading professionals in the field illustrated with examples of assessment practice in a variety of contexts.

Each of the videos is between 20 and 25 minutes in length. They are:

- Video 1  Language Proficiency Assessment
- Video 2  Principles of Test Development
- Video 3  Objective and Subjective Assessment
- Video 4  Stages of Test Analysis
- Video 5  Performance Assessment
- Video 6  Classroom-based Assessment

The video series will be on display within the book exhibition. Staff of the Language Testing Research Centre will be on hand to demonstrate the series on Thursday 9th and Friday 10th March at the following times:

Thursday 12.30 & 4.30
Friday 1.30 & 6.00
CANADIAN ACADEMIC ENGLISH
LANGUAGE ASSESSMENT

The Canadian alternative.

Students applying for university may need to take an English language test as part of their application process. The CAEL Assessment offers a unique alternative to other standardized tests of English because it is designed specifically to test English in use for academic purposes within college and university settings. It is available in a number of locations around the world, including many where other standardized tests are not available (e.g., Iran, Cuba, Greenland, etc.). It is also offered at test sites throughout Europe, Asia, and in Canada in Vancouver, Edmonton, Toronto, Peterborough and Ottawa. CAEL Assessment test results are accepted by many Canadian, American and European universities.

For more information about the CAEL Assessment contact:

The Language Assessment and Testing Research Unit
Carleton University
SLALS/ 215 Paterson Hall
1125 Colonel By Drive
Ottawa, Ontario
CANADA K1S 5B6
Tel. +1 613 520-6612 Fax: +1 613 520-6641
E-mail: CAEL@carleton.ca

Or, visit our website at: www.carleton.ca/slals/cael.htm
Cambridge Language Assessment — a brand new series

A new and authoritative series that is essential reading for anyone involved in language testing.

Assessing Reading
J. Charles Alderson

Assessing Languages for Specific Purposes

Assessing Vocabulary

Paperback 400 pp 0 521 59999 7
Hardback 0 521 59000 0

Paperback 311 pp 0 521 58543 0
Hardback 0 521 58495 7

Paperback 279 pp 0 521 62741 9
Hardback 0 521 62182 8

www.cup.cam.ac.uk/elt/clas/