LTRC 2005 Program Overview

MONDAY, JULY 18		
9:00 - 12:00	Workshop (by pre-registration only)	Simard 425
12:00- 1:00	LUNCH	
1:00 - 4:00	Workshop Continued	
TUESDAY, JULY 19		
9:00 - 12:00	Workshop (by pre-registration only)	Simard 425
12:00- 1:00	LUNCH	
1:00 - 4:00	Workshop Continued	
5:00-8:00	Registration	Quebec Suite, Chateau Laurier
6:00-8:00	Welcoming Reception	Quebec Suite, Chateau Laurier
WEDNESDAY, JULY 20		
8:15-8:35	Welcome	Arts 026
8:35-9:35	Keynote Address - Professor Bruno D. Zumbo	Arts 026
9:35-9:45	Presentation of the Messick Award	Arts 026
9:45-10:00	Presentation of Jacqueline A. Ross Dissertation Award	Arts 026
10:00-10:25	Break	Lobby outside Arts 026
10:25-12:15	Session 1 (3 papers +1 issue)	Arts 026
12:15-2:00	Lunch	
	Language Assessment Quarterly EAB Meeting	Salon L'Orangerie, Chateau Laurier
2:00-3:50	Session 2 (3 papers+ 4 posters)	Arts 026
3:50-4:15	Break	Lobby outside Arts 026
4:15-6:30	Session 3 (4 papers + 3 posters)	Arts 026
6:30-7:00	Group Photograph	
7:00-9:00	Student Organized Session	
7:30-9:00	ILTA Executive Board Meeting	
THURSDAY, JULY 21		
8:15-10:00	Session 4 (3 papers + 3 posters)	Arts 026
10:00-10:25	Break	Lobby outside Arts 026
10:25-11:25	Works in Progress (WIP) Session (13 WIPs)	Simard 125
11:30-12:30	Plenary – Professor Lyle Bachman	Arts 026
12:30-2:15	Lunch	
	Language Testing EAB Meeting	Salon L'Orangerie, Chateau Laurier
2:15-4:05	Session 5 (3 papers + 1 issue)	Arts 026
4:05-4:30	Break	Lobby outside Arts 026
4:30-6:00	Session 6 (3 papers)	Arts 026
FRIDAY, JULY 22		
8:15-10:25	Session 7 (4 papers + 2 posters)	Arts 026
10:25-10:50	Break	Lobby outside Arts 026
10:50-12:10	Plenary – Professor Charles Alderson (+1 issue)	Arts 026
12:10-1:55	Lunch	
	ILTA Business Meeting	Arts 026
1:55-3:25	Session 8 (3 papers)	Arts 026
3:25-3:50	Break	Lobby outside Arts 026
3:50-5:50	Symposium – Rethinking Language Testing	Arts 026
5:50	Closing Remarks	Arts 026
7:00-11:00	Banquet & Awards Presentation	Salon Laurier, Chateau Laurier

LTRC 2005 Program Details

MONDAY, JULY 18

LTRC Workshop 1

Quantitative Approaches to Language Testing Research and Development

Professor Bruno D. Zumbo, University of British Columbia

Location: Simard 425

9:00 - 12:00	Workshop (by pre-registration only)
12:00- 1:00	LUNCH
1:00 - 4:00	Workshop Continued

TUESDAY, JULY 19

LTRC Workshop 2

Qualitative Approaches to Language Testing Research and Development

Dr. Anne Lazaraton, University of Minnesota

Dr. Lynda Taylor, University of Cambridge ESOL Examinations

Location: Simard 425

9:00 - 12:00	Workshop (by pre-registration only)	
12:00- 1:00	LUNCH	
1:00 - 4:00	Workshop Continued	
5:00-8:00	Registration	Quebec Suite, Chateau Laurier
6:00-8:00	Welcoming Reception	Quebec Suite, Chateau Laurier
	Sponsored by Cambridge ESOL	

LTRC
ottawa · canada
2005

WEDNESDAY, JULY 20

Registration: 8:00 a.m. - 6:00 p.m. Lobby outside Arts 026

Publishers' Exhibits: 9:00 a.m. - 6:00 p.m. Simard 129

8:15-10:00	**Opening Session**	Arts 026
8:15-8:35	**Welcome:** Janna Fox, Carleton University; George Lang, Dean of Arts, University of Ottawa; Desmond Allison, Director, School of Linguistics and Applied Language Studies, Carleton University	
	Introduction of Keynote Speaker: Janna Fox, Carleton University	
8:35-9:35	**Keynote Address**	
	Reflections on validity at the intersection of psychometrics, scaling, philosophy of inquiry, and language testing	
	Professor Bruno D. Zumbo, University of British Columbia	
9:35-9:45	Presentation of the Messick Award by Mary Enright, Educational Testing Service (ETS)	
9:45-10:00	Presentation of Jacqueline A. Ross Dissertation Award for Outstanding Research on Second/Foreign Language Testing, 2005 by Mary Enright (ETS)	
10:00-10:25	**Break**	Lobby outside Arts 026

10:25-12:15 **Session 1**, Chair: Robert Edwards — Arts 026

Papers

1. Cognitive processes and use of knowledge in performing new TOEFL listening tasks

   Dan Douglas & Volker Hegelheimer

2. Methods and findings for differential functioning of language assessments

   Tracy Ferne, Hyeran Choi & André A. Rupp

3. Factors explaining EFL learners' performance in a timed-essay test: A structural equation modeling approach

   Youngmi Yun

Issue, Chair: Robert Edwards

1. Issues in testing listening comprehension in the multimedia web environment
   Gary Buck

12:15-2:00 **LUNCH**

Language Assessment Quarterly EAB Meeting

Salon L'Orangerie, Chateau Laurier

2:00-3:50 **Session 2,** Chair: Liying Cheng Arts 026

Papers

4. Validating the use of a standards-based classroom assessment of English proficiency: A multi-trait multi-method approach
   Lorena Llosa
5. Statutory testing and teaching in the primary EAL/ESL context: A case study
   Katie Scott
6. Technological, cultural, political and philosophical considerations: Validating language teachers' classroom assessment practices
   Ofra Inbar & Smadar Donitsa-Schmidt

Posters, Chair: Doreen Bayliss

1. Developing Arabic and Russian web-delivered listening and reading tests
   David MacGregor, Mohammed Louguit, Margeret E. Malone & Dorry M. Kenyon
2. Architects and engineers constructing an EAP placement test from grounded theory up
   Barbara K. Dobson, Mary C. Spaan & Amy D. Yamashiro
3. Introducing the modern language aptitude test for Spanish-speaking children
   Daniel J. Reed & Charles W. Stansfield
4. The development of a new proficiency test for Japanese EFL students
   Shoichi Ishikawa, Yuji Nakamura, Kahoka Matsumoto, Hiromi Kobayashi, Atsuko Okada & Dennis Schneider

3:50-4:15 **Break** Lobby outside Arts 026

4:15-6:30 **Session 3,** Chair: Carolyn Turner Arts 026

Papers

7. Conflicting demands in integrated reading/writing tasks
   Tom Lumley & Annie Brown
8. Individual feedback to enhance rater training: Does it work?
   Ute Knoch & Gary Barkhuizen
9. Towards a rater taxonomy: Examining rater types in language performance assessments
   Thomas Eckes
10. Dependability and separability of analytic rating dimensions for ESL essays
    Yong-Won Lee, Claudia Gentile & Robert Kantor

Posters, Chair: Doreen Bayliss

5. An online training programme for raters of academic writing
   Janet von Randow
6. Relationships between productive vocabulary knowledge and speaking test performance using multiple measures
   Rie Koizumi
7. The Treviso test
   Geraldine Ludbrook

6:30-7:00 **GROUP PHOTOGRAPH**

7:00-9:00 **Student Organized Session** Simard 125

7:30-9:00 **ILTA Executive Board meeting** 112 Daly Avenue

LTRC
ottawa • canada
2005

THURSDAY, JULY 21

Registration: 8:00 a.m. - 6:00 p.m. Lobby outside Arts 026

Publishers' Exhibits: 9:00 a.m. - 6:00 p.m. Simard 129

8:15-10:00 **Session 4**, Chair: Bob Courchène Arts 026

Papers

11. Rating writing performance: What do the judgments of test users tell us?

    David Qian & Tom Lumley

12. Setting the standard: What English language abilities do overseas trained doctors' need?

    Jayanti Banerjee & Lynda Taylor

13. The contribution of error-tagged learner corpora to the assessment of language proficiency

    Sylviane Granger & Jennifer Thewissen

Posters, Chair: Doreen Bayliss

8. Equivalence of two special purpose tests

   Amelia Kreitzer Hope & Doreen Bayliss

9. A simple framework for developing language tests on the web

   Alysse Weinberg

10. The effects of task types on listening test performance: A retrospective study

    Yo In'nami

10:00-10:25 **Break** Lobby outside Arts 026

10:25-11:25 **Works in Progress Session** Simard 125

---

1. A study of motivation and task selection for an exit e-portfolio

   Vivien Berry

2. The relevance and relativity of specific linguistic features in oral performance tests

   Sarah L. Briggs & India C. Plough

3. Cramming or teaching: The washback effect of the TOEFL in China
   Ho Fai Chau (Michael)
4. Analyzing examinees' cognitive processes in standardized reading tests: A tree-based modeling approach
   Lingyun Gao & Changjiang Wang
5. Native English speaking teachers' impact on classroom assessment abroad
   Deniz Gokcora
6. Chinese students' experiences of two English writing tests: TWE versus LPI
   Ling He & Ling Shi
7. Exploring difficulty in an L2 monologue test task
   Tomoko Horai
8. Issues in assessment of ESOL – Washback of the skills for life strategy
   Tania Horak
9. The development and validation of a corpus-based rating scale for writing
   Ute Knoch
10. The interplay of portfolio assessment and language testing practice
    Lucilla Lopriore & Guido Benvenuto
11. Unraveling second language speaking proficiency
    Rob Schoonen, Jan Hulstijn, Nivja de Jong, Margarita Steinel & Arjen Florijn
12. Measuring the knowledge of text structure in academic ESL reading
    Viphavee Vongpumivitch
13. Do test formats in reading comprehension influence ESL and non-ESL students differently?
    Ying Zheng

11:30-12:30 ***Plenary*** Arts 026

What are we assessing? The dialectic of constructs and contexts in language assessment

Lyle Bachman

12:30-2:15 **Lunch**

Language Testing EAB meeting Salon L'Orangerie, Chateau Laurier

---

2:15-4:05 **Session 5**, Chair: Desmond Allison Arts 026

Papers

14. Frequency and quantity of output in group oral tests

    Miyoko Kobayashi & Alistair Van Moere

15. The contribution of lexical analysis to speaking test validation

    John Read

16. Automatic grading of story retelling

    Anish Nair, Jennifer Balogh, Isabella Barbier, Jared Bernstein, Matt Lennig & Masanori Suzuki

Issue, Chair: Robert Edwards

2. Measuring content: Lessons from an undergraduate history course for language assessment

   Muhammad Usman Erdosy

4:05-4:30 **Break** Lobby outside Arts 026

4:30-6:00 **Session 6**, Chair: Randy Thrasher Arts 026

Papers

17. Safe to practice? Setting minimum language proficiency standards among nursing professionals

    Lynda Taylor & Thomas O'Neill

18. Language testing in the military: Problems, politics and progress

    Rita Green & Dianne Wall

19. Using national standards to develop profession-specific language assessment tools

    Lucy Epp

LTRC
ottawa · canada
2005

FRIDAY, JULY 22

Registration: 8:00 a.m. - 6:00 p.m. Lobby outside Arts 026

8:15-10:25 **Session 7**, Chair: Sari Luoma Arts 026

Papers

20. A comparison of production questionnaires and role plays assessing L2 pragmatic competence

    Miyuki Sasaki

21. Planning before speaking: What difference does it make to test performance?

    Catherine Elder & Gillian Wigglesworth

22. Examining rater biases affecting phonological assessment

    Hiroko Yoshida

23. Design patterns in language assessment

    Hua Wei, Diwakar Khanal & Robert Mislevy

Posters, Chair: Doreen Bayliss

11. Setting and monitoring professional standards: A QMS approach

    Nick Saville, Piet Van Avermaet & Henk Kuijper

12. Developing a web-delivered K-12 test of ESL listening and speaking

    Gary Buck & Cheryl Alcaya

10:25-10:50 **Break** Lobby outside Arts 026

10:50-12:10 Arts 026

Plenary

The challenge of (diagnostic) testing: Do we know what we are measuring?

Charles Alderson

Issue, Chair: Robert Edwards

3. What have we learned in three thousand years?

   Liz Hamp-Lyons

12:10-1:55 **Lunch**

ILTA Business Meeting Arts 026

1:55-3:25 **Session 8**, Chair: Antony Kunnan Arts 026

Papers

24. An empirical investigation of L2 reading comprehension skills

    Hossein Farhady & Gholmraza Hessamy

25. Exploring the relationship between item feature and students' performance on reading tests

    Changjiang Wang & Lingyun Gao

26. Strategies in responding to the next generation TOEFL reading tasks

    Andrew Cohen & Thomas A. Upton

3:25-3:50 **Break** Lobby outside Arts 026

3:50-5:50 **Symposium – Sponsored by ETS** Arts 026

Rethinking language testing: Voices from experience
Organizer: *Mari Wesche*

Presentations

1. On second thought…

   Bernard Spolsky

2. The coming of age of test-taking strategies

   Andrew Cohen

3. Academic language proficiency testing in the UK: Why IELTS?

   Alan Davies

4. Language testing: A question of context

   Tim McNamara

5. Tests as power tools: Looking backwards, looking forward

   Elana Shohamy

5:50	**Closing Remarks:** Janna Fox, Carleton University *Lado Award for Outstanding Student Paper at LTRC 2005*	Arts 026
7:00-11:00	**Banquet & Awards Presentation**	Salon Laurier, Chateau Laurier

Music by Jazz Apéritif

AWARD

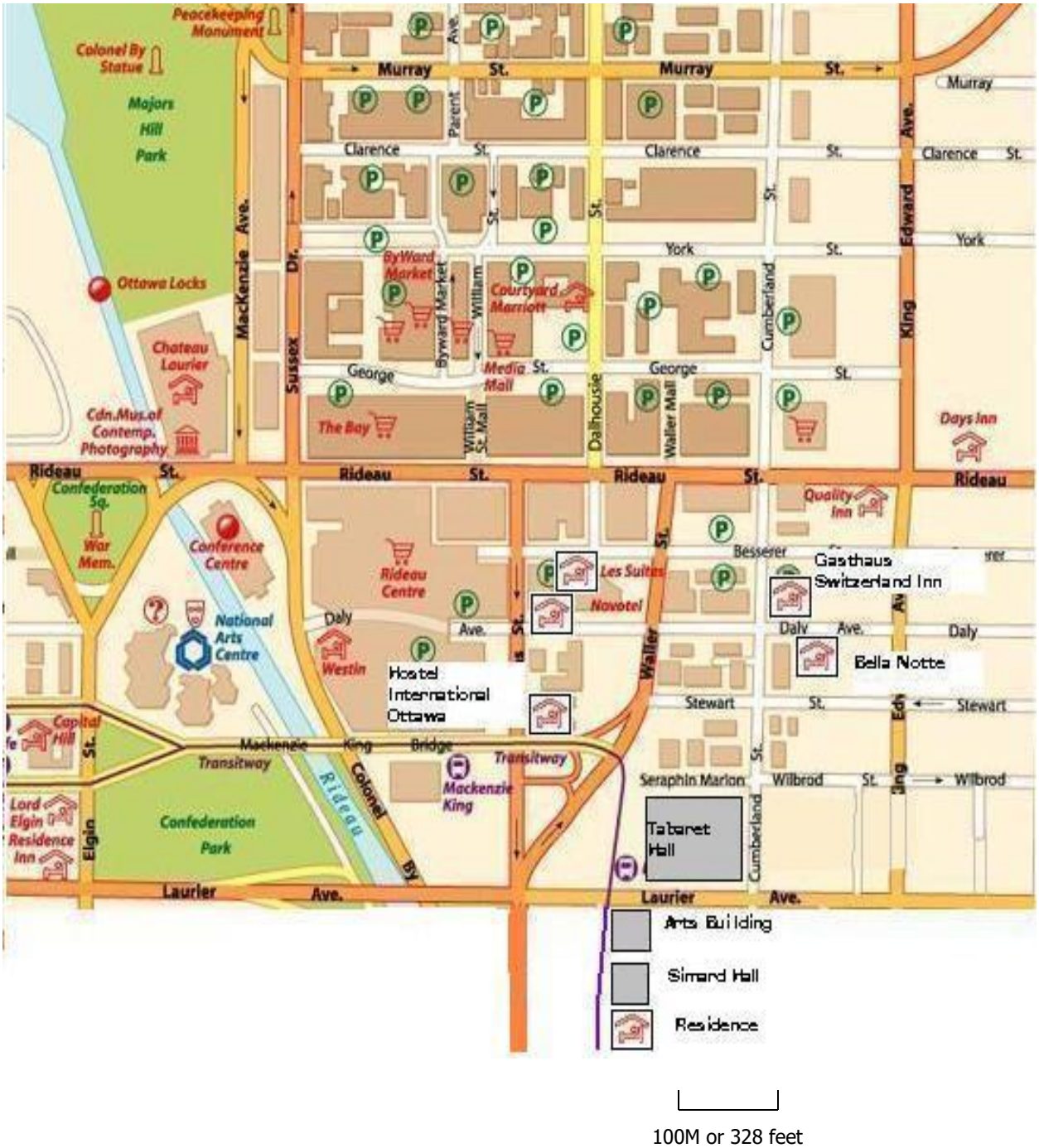
ILTA-UCLES Lifetime Achievement Award

presentation to

and acceptance by

Professor Bernard Spolsky

Site Map



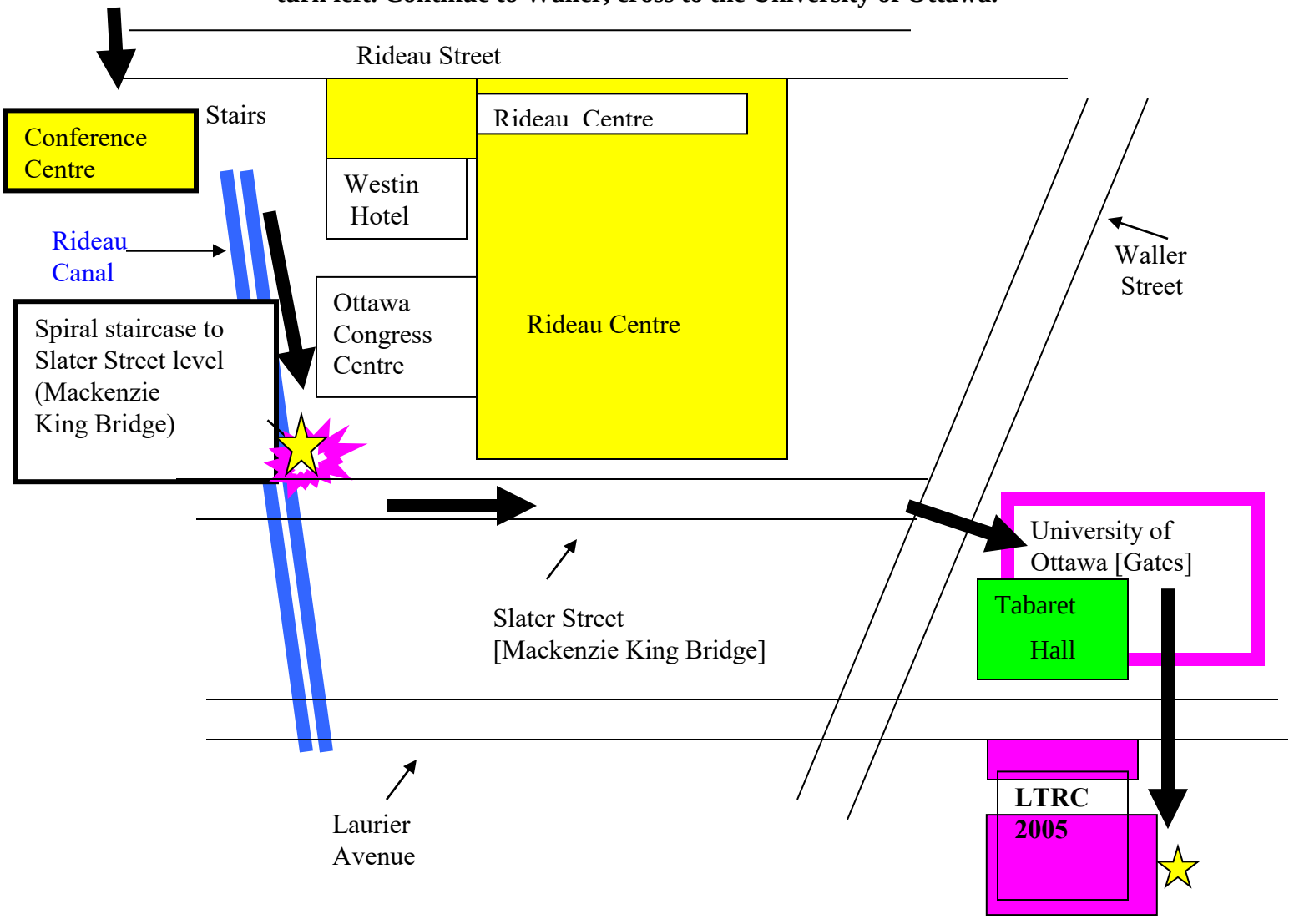
Without rushing, the walk from the Chateau Laurier to the Arts building should take less than 10 minutes.

Walking Directions

Cross Rideau Street directly in front of the Chateau Laurier hotel. Walk to the left down the stairs (beside the Conference Centre) and then up the ramp toward the Canal. Walk along the Canal (the Canal should be on your right; the Westin Hotel and the Ottawa Congress Centre should be across the street on your left). Just past Daly Avenue you will see a spiral staircase on your right. Walk up the spiral staircase to the Mackenzie King Bridge (on Slater Street) and continue walking across the bridge. Continue past the Rideau Centre to the next cross street (Waller). Cross Waller and enter the University of Ottawa campus through the campus gates. Walk to the right, in front of Tabaret Hall, cross Laurier. LTRC 2005 is in the Arts Building (you will see signs on the building on your right).

Chateau Laurier Hotel 1 Rideau St.

If it's raining (or you want a shortcut) you can walk through the Rideau Center. Enter at the corner of Rideau Street and Sussex (by the Elephant and Castle pub), go up one level on the escalator to your left. Turn right at the top of the escalator and walk in the direction of Sears. Leave the Rideau Centre through a Mackenzie King Bridge exit (there are three of them) and turn left. Continue to Waller, cross to the University of Ottawa.



Papers

(According to the sequence of the presentations)

#	Day/Time	Paper
1	July 20 10:25– 12:15 **Session 1**	***Cognitive processes and use of knowledge in performing new TOEFL listening tasks*** *Dan Douglas & Volker Hegelheimer* This presentation is an interim report on an investigation of the cognitive processes and use of language and content knowledge by test takers while completing the New TOEFL listening test tasks. We employ a three-pronged approach for data collection and analysis: verbal protocol analysis of participants' test-taking processes, screen capture of their online behavior, and analysis of their handwritten notes to investigate the following research questions: 1) What cognitive and metacognitive strategies do test takers use? 2) What linguistic and content knowledge do test takers use to respond to the tasks? We contend, following previous research that the processes of listening comprehension are accessible to researchers by means of the analysis of introspective verbal protocols, or think aloud procedure, and handwritten notes. The focus in this project is on integrating these three types of data to more fully understand the cognitive processes and use of language and content knowledge required for responding to the new TOEFL listening test tasks. Participants in the study, 12 undergraduate and graduate test takers enrolled in social sciences, science, technology, and the humanities, completed tasks based on three of the listening passages in the *LanguEdge* (ETS 2003) materials. Analysis will focus on strategies and content knowledge employed and how notes are integrated into the response process. Findings will provide support for interpreting test performance as evidence for language knowledge and communicative language use as defined in the test construct. The presentation will report on the results of the pilot study focusing on the research methodology.
2	July 20 10:25– 12:15 **Session 1**	***Methods and findings for differential functioning of language assessments*** *Tracy Ferne, Hyeran Choi & André A. Rupp* This paper reviews a variety of modern psychometric methods for the detection of a differential functioning of language assessments such as logistic regression (e.g., Zumbo, 1999), SIBTEST (e.g., Stout, 2003), confirmatory factor analysis (e.g., Bae & Bachman, 1996), item response theory (e.g., Raju, Lafitte, & Byrne, 2002), and hierarchical linear modeling (e.g., Swanson et al., 2002). Furthermore, it is shown how predictor variables can be incorporated into investigations with such models to investigate the reasons for a differential functioning within a framework of

explanatory measurement (see de Beock & Wilson, 2004) and the implied potential biases for certain examinee groups that take language tests.
To illustrate the utility of these models, a synthesis of peer-reviewed articles, theses, and reports from 1995-2004 will be presented. To supplement these analyses, an empirical analysis of a large-scale data set of a reading assessment will illustrate how the different methods can be used en concerto to provide diverging and converging pieces of evidence for a differential functioning of individual items and the assessment as a whole. Together, the components in this paper provide an integrative overview of a variety of methods for evaluating the inferential validity of language tests that allow for the development of data collection tools suited to the proper and efficient analysis of potential biases in language tests.

3 July 20 10:25–12:15 **Session 1**

Factors explaining EFL learners' performance in a timed-essay test: A structural equation modeling approach

Youngmi Yun

Many studies have suggested several explanatory factors for L2 writing performance: L1 writing ability, L2 proficiency, past L2 writing experience, past L2 reading experience, and test preparedness. In the context of adult Korean EFL learners (total 351), three central issues are addressed in this study.
First, employing a structural equation modeling (SEM) approach, the study formulates a comprehensive model regarding the structure of these variables and performance in an Adapted Test of Written English (ATWE) and investigates the relative influence of these variables on L2 performance. Specifically, it is examined whether the structural model is invariant across good and poor L2 writer groups using multi-group SEM. The results show that for both groups, L2 proficiency is a better predictor than L1 writing ability. Second, the study concerns the presence and the nature of "writing expertise," which has been argued to transcend the two languages, by examining the factor structure of the relationship between the two latent constructs (L1 and L2 writing ability). Three competing factor structures are tested using confirmatory factor analysis (CFA). The findings support for the presence of writing expertise. Third, the study examines the extent to which the relationship between L1 and L2 writing performance across differing L2 proficiency supports the linguistic threshold hypothesis: the linguistic threshold hypothesis claims that a certain minimum level of competence in L2 is required in order for the literacy knowledge acquired in L1 to transfer to L2. The latent interaction effect is tested. The findings support the presence of threshold level. The modeling of L2 writing performance will lead to a better understanding of the construct being measured in the ATWE. The findings will contribute to enriching our understanding of the multidimensional characteristics of L2 writing and provide pedagogical applications useful both for EFL writing instructors and EFL writing learners.

4 July 20 2:00 - 3:50

Validating the use of a standards-based classroom assessment of English proficiency: A multi-trait multi-method approach

Lorena Llosa

Session 2 The use of standards-based classroom assessments to test English learners' language proficiency is becoming increasingly prevalent. In a large urban school district in California, for example, a classroom assessment is used to make high-stakes decisions about English learners' progress from one level to the next, and it is also used as one of the criteria for reclassification as Fluent English Proficient. Yet many researchers have questioned the reliability and validity of using classroom assessments for making high-stakes decisions about students (Brindley, 1998, 2001; Rea-Dickins & Gardner, 2000).

One way to investigate the validity of the inferences drawn from these standards-based classroom assessments is to examine them in relation to other measures of the same ability. In this study, a multivariate analytic approach was used to examine the extent to which the standards-based classroom assessment used in a large district in California measures the same constructs as the CELDT (California English Language Development Test), the statewide standardized test of English proficiency also aligned to the California ESL Standards. Using confirmatory factor analysis of multitrait-multimethod data, this study investigates the convergent/discriminant validity of these two assessments as well as method effects. Findings will be discussed in term of their implications for the use of standards-based classroom assessments within a high-stakes accountability system.

5 July 20
2:00 - 3:50

Statutory testing and teaching in the primary EAL/ESL context: A case study

Catriona Scott

Session 2 This paper reports on an exploratory case study of the effects of high-stakes statutory testing on primary English as an Additional Language (EAL) learners in the UK. The study involved interviews with EAL learners, teachers and parents, classroom observation, and analysis of samples of learners' work and test papers.

The findings suggest that manifestations of washback may both be positive in terms of their congruence with guidance on the teaching of EAL learners and yet have concurrent negative effects. For example, the allocation of extra time for writing within the timetable may facilitate rehearsal and reworking of tasks, but may also contribute to the reduction of time for non-core subjects. Equally, streaming/ability grouping may aid differentiated teaching and provision of scaffolding by teachers and peers, but the prescribed time-frame for completion of the curriculum can potentially diminish the benefits.

In addition, there would seem to be a tension between the testing component of statutory assessment, requiring independent performance within time constraints, and the collaborative teaching and learning environment considered to be especially beneficial to EAL learners. Therefore, there are potential trade-offs between preparing EAL learners for tests to enable 'best performance', as explicitly advocated in official guidance, and maintaining classroom practices which promote and support language development.

This presentation considers the extent to which there is congruence and/or

dissonance between EAL-oriented teaching and washback from high-stakes testing. A conceptualization of washback in the EAL context is proposed, discussed in relation to Alderson and Wall's (1993) hypotheses and Bailey's (1996) model.

6 July 20 2:00 - 3:50 **Session 2**

Technological, cultural, political and philosophical considerations: Validating language teachers' classroom assessment practices

Ofra Inbar & Smadar Donitsa-Schmidt

The purpose of this research was to reflect on different perspectives involved in language teachers' implementation of classroom assessment practices, focusing specifically on the teachers' perceptions of alternative assessment and on the reasons for differential use of assessment measures. The framework for analysis was derived from the Hargreaves, Earl & Schmidt's model (2002) (based on House, 1981), previously used for examining teachers' views and use of alternative assessment according to 4 perspectives: Technological, Cultural, Political and Postmodern. Using a self-report questionnaire developed for this research (76 likert-scale items & 2 open-ended items), the study attempted to empirically validate the existence of the four perspectives in a language assessment context by examining teachers' views regarding the four perspectives, and by exploring the relative effect of each factor on the teachers' classroom assessment practices. The sample included 113 EFL teachers teaching in different school levels in Israel. Results of a two-stage factor analysis empirically validate the model, while highlighting the complexity of alternative assessment as not merely a testing approach but as a social and educational paradigm.

7 July 20 4:15-6:30 **Session 3**

Conflicting demands in integrated reading/writing tasks

Tom Lumley & Annie Brown

In order to replicate more closely the language demands placed on students in academic contexts, writing tasks in the new TOEFL require test takers to integrate input material into their writing, rather than simply to draw on their own background knowledge or ideas in response to a question or prompt. These materials consist of cognitively complex written texts or lectures, typically on subject matter with which the test taker cannot expect to be familiar. As a result these *integrated* tasks are more complex and more demanding than the more traditional stand-alone or *independent* tasks. As performance on integrated tasks involves cognitive skills (information selection and structuring) alongside more conventional language skills (comprehension and production), the question of the balance between the two becomes potentially problematic. This paper considers the fundamental issue of how test takers and raters deal with these potentially conflicting task demands.

A study of test taker behaviour and rater responses in integrated reading/writing tasks related to the new TOEFL explored this tension through examination of a) written texts produced by 60 Mandarin-, Cantonese- and Korean-speaking learners of English, b) retrospective verbal reports by these learners, and c) verbal reports produced by six raters of these learners' written texts. Findings show how for test takers, conflict may

arise between the demand to complete the task successfully and the requirement to produce a piece of accurate language. For raters, the tension between assessment of content and of language is equally difficult to resolve.

8 July 20 4:15-6:30 **Session 3**

Individual feedback to enhance rater training: Does it work?

Ute Knoch & Gary Barkhuizen

Research on the utility of feedback to raters in the form of performance reports has produced mixed findings (Lunt, Morton, & Wigglesworth, 1994; Wigglesworth, 1993) and has thus far been trialed only in oral assessment contexts. This paper reports on a study investigating raters' attitudes and responsiveness to feedback on their ratings of an analytically-scored writing task administered in the context of a large-scale university writing assessment program. After participating in an online refresher training program, 50 scripts were rated independently by eight raters. A multifaceted Rasch analysis (using FACETS) was used to generate individualized reports on each rater's pattern of bias with respect to particular categories of the rating scale as well as their overall consistency and relative severity in relation to others in the group. Reports were explicated at a group briefing session and raters were asked to attend to the feedback when scoring a further batch of 60 scripts. Qualitative data on rater attitudes to the feedback were also elicited via questionnaire. Results showed that most raters found the feedback useful and showed enhanced awareness of their rating behaviour. A comparison of ratings before and after feedback (again using Rasch analysis) revealed that many were able to modify their scoring, resulting in greater intra-group consistency and a reduced incidence of item bias. There was nevertheless variation in receptivity to feedback and questionnaire responses are used to suggest reasons for this. Gains in rater reliability moreover had the effect of reducing the test's discriminatory power suggesting that the costs of implementing this rather elaborate approach to training may outweigh the benefits in this instance. Further research is advocated, in particular on high stakes tests where reliability of scoring is at a premium.

9 July 20 4:15-6:30 **Session 3**

Towards a rater taxonomy: Examining rater types in language performance assessments

Thomas Eckes

Research on rater effects in language performance assessments has provided evidence for a considerable degree of variability among raters. In particular, raters have been shown to differ substantially in the severity/leniency they exercise when scoring examinee writing or speaking performance. Building on this research, the present study examined two questions: (a) Can rater subgroups or rater types be identified that show distinctive and coherent patterns of scoring tendencies, and, if so, (b) what are the characteristic features of these rater types? It was hypothesized that rater types can be distinguished by the way raters make use of well-practiced criteria for scoring examinee performance. Specifically, to the extent that rater types systematically differ in the importance attached to

particular criteria, differences in the level of severity/leniency should ensue. To address these issues, raters actively involved in scoring examinee performance on the TestDaF (Test of German as a Foreign Language) writing section were asked to indicate how much importance they would attach to each of nine routinely used criteria. The rating criteria covered various performance aspects, such as grammatical and lexical correctness or degree of structure. Use of a two-mode clustering approach allowing construction of a joint classification of raters and criteria yielded a clear-cut picture of rater types and their interrelationships with criteria. The findings have implications for the quality of large-scale rater-mediated language assessment, rater monitoring, and rater training.

10 July 20 4:15-6:30 **Session 3**

Dependability and separability of analytic rating dimensions for ESL essays

Yong-Won Lee, Claudia Gentile & Robert Kantor

Analytic scoring rubrics are useful in generating diagnostic feedback for writers. However, such analytic rubrics have not been widely used for large-scale writing assessments due to cost factors in rating. Another argument against analytic scoring might be that analytic scores for different rating dimensions often turned out to be highly correlated. The main purpose of the study is to investigate the separability and dependability of analytic rating dimensions and their relationships to holistic scores in the context of TOEFL CBT (computer-based TOEFL) writing assessment.

A total of 930 essays were double-rated by trained raters according to six analytic scoring rubrics (development, organization, vocabulary, language use (variety/error), mechanics). Multivariate G-theory analyses were conducted on the analytic scores. Pearson product moment correlations were computed for these six analytic scores and holistic scores. Partial correlations were also computed for these variables after controlling for essay length on these scores.

Results of G-theory and correlation analyses showed that (1) score reliability was the highest for the vocabulary dimension but the lowest for the mechanics dimension; (2) the universe scores of the six dimensions were rather highly correlated; and (3) the development and vocabulary scores were most highly correlated to the holistic scores. One interesting finding was that the essay length seemed to be responsible for such high correlations among the analytic and holistic scores. The results of these and further analyses (e.g., Facets, cluster, score profile analyses) and their implications for ESL writing assessment will be discussed in the presentation.

11 July 21 8:15-10:00 **Session 4**

Rating writing performance: What do the judgments of test users tell us?

David Qian & Tom Lumley

Tests for university students have typically focused on selection, as part of the entry process. This paper reports on an *exit* test for university students, the Graduating Students' Language Proficiency Assessment (GSLPA), which

is designed to provide information to prospective employers about the English language proficiency of newly graduating job applicants. The written component of this test includes tasks relevant to the domain of business, the destination of most graduates in Hong Kong. Since the test performances are rated operationally by language-trained raters rather than the users of the results, the question arises of how well their judgements of candidates' ability relate to those of the main users of the test results, business employers. A further question concerns whether employers in different occupations make similar judgements to each other.
The paper examines data collected from interviews with 17 business employers or human resources personnel from two occupational groups in Hong Kong, engineering and hospitality. They each provided ratings of 10 sample GSLPA writing performances, indicating 1) how likely they would be to employ the writer on the basis of demonstrated English proficiency and 2) either what kind of position they would offer, or why they wouldn't employ the writer. The data were compared with the operational ratings produced by trained GSLPA raters. The paper discusses the similarities and differences observed, including the weight apparently accorded to particular criteria, and the extent of agreement amongst raters from different occupations, providing implications for rater training and refinement of the rating scale.

12 July 21 8:15-10:00 **Session 4**

Setting the standard: What English language abilities do overseas trained doctors' need?

Jayanti Banerjee & Lynda Taylor

This paper will describe a review of the General Medical Council's (GMC) minimum language proficiency criteria for International Medical Graduates (IMGs), taking into account the judgments of 3 stakeholder groups: patients, doctors, and other health workers (such as nurses, physiotherapists and phlebotomists). These represent the main groups that come into contact with the IMGs.
The study design drew on the standard-setting methodology presented in ETS (1995), Jones & Hunter (1996), Impara & Plake (1997), Lumley (1998) and Council of Europe (2003) and focused on two skills – writing and speaking. As all IMGs hoping to practice in the United Kingdom are required to provide an International English Language Testing System (IELTS) test score as proof of their English language proficiency, the sample performances were taken from the IELTS writing and speaking tests and covered a range of performance levels from IELTS band 5 to IELTS band 8. The nationalities of the test-takers reflected the nationalities and English varieties of the IMGs who typically seek accreditation.
This paper will describe the study design and present the results, including the minimum IELTS writing and speaking scores that the GMC has now set. In particular it will discuss the differences between the stakeholder groups in their perceptions of what is adequate language proficiency. This study demonstrates the value of conducting standard setting studies with all stakeholder groups and also raises important questions about how language testers should reconcile differences in stakeholder opinion.

13 July 21 8:15-10:00 **Session 4**

The contribution of error-tagged learner corpora to the assessment of language proficiency

Sylviane Granger & Jennifer Thewissen

Learner corpora (LC) are electronic collections of spoken or written texts produced by foreign or second language learners. Despite its relative youth, this new resource displays considerable potential for interlanguage studies. It is, among others, currently being used to inform pedagogical tools (learner dictionaries, grammars, textbooks, CALL programs) and NLP tools such as spell- and grammar checkers. Another possible field which would most certainly gain from the use of LC is that of language testing, but as highlighted by Barker (2003), "they are not yet exploited fully by the language testing community". Our presentation will consider how a quantitative and qualitative analysis of the number and type of errors found in LC can contribute to the assessment of language proficiency levels. It focuses on learner texts extracted from the *International Corpus of Learner English* (Granger 2003) and which we error-tagged on the basis of the error tagging system designed at Louvain (Dagneaux et al 1998). Our aim was to establish clear links between certain error types and the different levels of proficiency found in the *Common European Framework of Reference for Languages* (CEF). As each of the error-tagged texts had already been assessed independently by a professional rater on the basis of the CEF guidelines, we were able to compare them with the level of proficiency each had been assigned independently. Our aim in doing so was to flesh out the descriptors found in the CEF which are very explicit on the learners' 'Can Dos' but only implicitly refer to their grammatical and lexical 'Can't Dos'.

14 July 21 2:15-4:05 **Session 5**

Frequency and quantity of output in group oral tests

Miyoko Kobayashi & Alistair Van Moere

The peer group discussion format for oral performance testing is becoming more widespread in schools and universities. In this resource-economical test arrangement, candidates are free to direct their conversation and so can be assessed in natural language use on both the mechanical features of oral proficiency (i.e. pronunciation, grammar) as well as the interactional features. However, without an interlocutor-examiner to guide them, some candidates may dominate floor time and others may not participate enough. This may affect how much other candidates in the group are allowed to contribute. The effects could be exacerbated when group sizes and test time vary.

The current study addresses this potential problem in group oral tests administered at a university in Japan. This study builds on existing research investigating variables related to this test format, including task and rater variables, gender, peer-interlocutor proficiency levels, and shyness/ outgoingness. We investigate whether the amount of output has any impact on students' scores, both at group and individual levels. The performances of approximately 180 individuals within about 50 separate small-group tests were video-recorded and transcribed for analysis. The data shows that, higher scores are associated with the larger number of words spoken,

rather than with frequency of turns. This is true at all levels of English proficiency. These findings have implications for test design, such as the optimum amount of speech for a ratable sample, how raters respond to students who speak far more than required, and how behaviour of one candidate
affects the performance of the others.

15 July 21 2:15-4:05 **Session 5**

The contribution of lexical analysis to speaking test validation

John Read

An analytic approach to the rating of a speaking test assumes that raters can make a valid assessment of separate components of the test-takers' performance. One such component is vocabulary range and use. In applied linguistic research, comparatively little is known about the nature of oral vocabulary since most research until now has been based on written texts. This means that there is scope for studies of vocabulary use by learners in oral assessment contexts.
The present study draws on a corpus of 88 transcripts of audiotaped performances under operational conditions by candidates in the IELTS Speaking Test, which has Lexical Resource as one of its four rating criteria. The first phase of the analysis involved the calculation of a number of lexical statistics to measure lexical output, lexical variation, lexical sophistication and the occurrence of key words related to the test topic. This quantitative approach, which is based on the counting of individual word forms, was complemented by a qualitative analysis of the transcripts to investigate the role of multiword formulaic expressions, particularly in distinguishing the performance of candidates at three distinct proficiency levels: Bands 8, 6 and 4. The paper reports selected results of the analysis, which suggest that the lexical features in the speech of candidates at the different levels are not easy to separate. The implications of the findings for the training of raters and the definition of the rating scale for Lexical Resource will be discussed.

16 July 21 2:15-4:05 **Session 5**

Automatic grading of story retelling

Anish Nair, Jennifer Balogh, Isabella Barbier, Jared Bernstein, Matt Lennig & Masanori Suzuki

Recent advances in speech recognition and grading technology have led to the development of automatically graded tests of spoken language. One such example is Ordinate Corporation's Spoken Spanish Test (SST) which generates vocabulary, sentence mastery, pronunciation and fluency scores based on a variety of test items. Among the test items is the task of story retelling, which is not currently scored, but is used to elicit spontaneous speech samples that can be rated by human experts for validation purposes.

These samples contain important information about the test-taker's ability to produce meaningful and coherent spoken language for an extended period of time. This motivated us to undertake an effort to develop technology that automatically and accurately grades these responses.
We adopted an information theoretic approach to assess the relative quality of spoken responses to the story retelling task based on the presence and absence of words and word sequences taken from automatically generated transcriptions. Simultaneously, we undertook a rating experiment to collect human grades for these responses based on human transcriptions, to be used for development and validation of the grading methodology. We found that our approach produced scores, which were highly correlated with the human ratings and clearly distinguished native and non-native speakers. On integration with the SST, we noticed a significant improvement in the reliability of the vocabulary sub-score. This paper and talk describes the motivation behind this work, the experimental methodology and the results obtained. It discusses further research to improve the other sub-scores using different feature sets.

17 July 21 4:40-6:30 ***Safe to practice? Setting minimum language proficiency standards among nursing professionals***

Lynda Taylor & Thomas O'Neill

Session 6

The opening up of international borders and the growth in global employment opportunities has led to considerable expansion in the number of overseas health professionals (e.g. doctors, nurses) entering the UK and the USA to work in these nations' health services. Both countries are becoming increasingly dependent on recruiting suitably qualified overseas personnel to meet their employment needs; this raises questions about the level of English language proficiency needed by health professionals for whom English is a second language – especially those seeking official registration or license to practice.
This paper will report on a recent study conducted in the USA to identify the minimum level of English language proficiency needed for entry-level nursing professionals to be able to work safely and effectively, and to determine the associated cut scores on an internationally recognized test of English language proficiency (IELTS).
The presentation will begin with a brief explanation of the background to the study and go on to describe the different methodologies adopted for cut score estimation consistent with the characteristics of each language subtest: a) a modified Angoff (1971) method for the Reading and Listening subtests, and b) a modified Analytical Judgement Method (Plake and Hambleton, 2000) for the Writing and Speaking subtests. The selection and qualifications of the panellists will be outlined, along with the training given to them. The paper will illustrate how such studies can assist policy makers in the setting of reasonable and defensible cut scores.

18 July 21 4:40-6:30 ***Language testing in the military: Problems, politics and progress***

Rita Green & Dianne Wall

Session 6

There appears to be little literature available - either descriptive or research-related - on language testing in the military. This form of specific purposes assessment affects both military personnel and civilians working within the military structure in terms of posting, promotion and remuneration, and it could be argued that it has serious social consequences if not carried out professionally and to the highest standard. This presentation provides a general overview of the language testing that is carried out by military teams in Central and Eastern Europe, using the findings of three surveys conducted with teams in this region. It explores the design problems the teams have had to deal with, the political issues that influence the work they do, and the progress they have been able to make over recent years. The presentation concludes with an attempt to link these findings to broader issues in the assessment of English for Specific Purposes.

19 July 21 4:40-6:30

Using national standards to develop profession-specific language assessment tools

Lucy Epp

Session 6

A feasibility study of nursing stakeholders across Canada indicated general dissatisfaction with language assessment tools used to measure the English language proficiency of internationally educated nurses. Stakeholders indicated the need for a language proficiency assessment tool that reflected language used in the nursing context. Based on this feedback, the Canadian English Language Benchmarks Assessment for Nurses (CELBAN) was developed. Test developers from Red River College collaborated with a team of experts from across Canada to develop the CELBAN. The Canadian Language Benchmarks (CLB), the national standards for English as a Second Language in Canada, were used as a framework. Using the CLB, an in-depth analysis of the language demands of the nursing profession across Canada was carried out. CLB Levels were assigned, and the CELBAN was then developed, using the data collected during this analysis. Throughout the process, feedback from nursing stakeholders was solicited.

Most Canadian nursing licensing bodies have now endorsed CELBAN as an option for providing evidence of English language proficiency for nursing licensure; a growing number of CELBAN administration sites have now been established. This project is the first in Canada to use national standards in the development of a profession-specific language assessment tool. As such, it provides a model for occupation-specific language test development in other fields.

The presentation will include the following topics:

1. The analysis of the language demands of the nursing profession
2. The development of CELBAN
3. The ongoing implementation of CELBAN

20 July 22 8:15-10:25

A comparison of production questionnaires and role plays assessing L2 pragmatic competence

Miyuki Sasaki

Session 7 Production questionnaires and role plays have been two popular measures of L2 pragmatic competence (Kasper, 2000). In the exploratory study (Sasaki, 1998), I found that these two measures elicited notably different performance from Japanese EFL learners in terms of response length, range/content of expressions, and appropriateness scores given to responses. In the present study, I tested the hypotheses formulated on the basis of the results of Sasaki (1998), and further explored the participants' thought processes underlying the intra-participant performance variations observed in the previous study. Forty Japanese college students with low-to-intermediate English proficiency responded to the two measures for the same four request and four refusal situations. Immediately after responding to each situation, the participants answered a questionnaire asking about their thought processes during their speech act formulation. The questionnaire included questions such as how much planning the participants had done beforehand and how much attention they paid to grammar and pronunciation while performing the speech act (Cohen & Olshtain, 1993). Sixteen participants provided additional retrospective protocol data, which supplemented the questionnaire data. The results revealed that (1) the performance elicited by the two measures was significantly different in response length and strategy variations; (2) the method factor was so large that we cannot claim that the two methods measured exactly the same trait; and (3) the method factor seems to have been partly caused by affective factors such as the participants' attitudes toward different measures.

21 July 22 8:15-10:25 ***Planning before speaking: What difference does it make to test performance?***

Catherine Elder & Gillian Wigglesworth

Session 7 Research on the effects of pre-task planning on speaking test performance has thus far yielded conflicting findings (Wigglesworth 1997 & 2000, Elder & Iwashita in press) and therefore warrants further investigation.
This paper reports on a study exploring a) whether performance on Task 2 of the IELTS academic speaking module changes according to the amount of planning time provided and b) whether the range and types of planning strategies used by candidates impacts on subsequent task performance.
90 English (L2) learners took part in the study, each performing 3 tasks with zero, 1, and 2 minutes of pre-task planning time respectively. A counterbalanced design was adopted with controls for language proficiency (intermediate and advanced), task (3 versions) and order of presentation. A post-test questionnaire and focus-group interview were used to elicit feedback on planning strategies and affective reactions. Performances were taped and double-rated by IELTS-certified raters using standard IELTS criteria. Multifaceted-Rasch analyses (FACETS) and discourse-analytic measures were used to compare speech quality under the different planning conditions. Data on strategy use under the planned condition were also correlated with test scores. Results showed only moderate effects for planning time, which had more impact on test-taker discourse than test scores. Some planning strategies were more influential than others on the quality of test performance, although there were complex interactions

between proficiency and strategy use. The validity implications of these results and their links to previous research are discussed and recommendations are put forward for both EAP test design and test-taker preparation.

22 July 22 8:15-10:25 **Session 7**

Examining rater biases affecting phonological assessment

Hiroko Yoshida

This study examined rater biases in assessing pronunciation performance. A total of 60 Japanese EFL college students read aloud two different texts (a prose type reading and a dialog type reading) that were designed to diagnose learners' pronunciation problems. Their audiotaped performances were rated by three L1 English instructors and three L1 Japanese instructors. The assessment items consisted of 15 items that examined three aspects of the sound system of General American English (segmentals, suprasegmentals, and paralinguistic features). Segmental features included vowel, diphthong, consonant, consonant cluster, and aspiration. Suprasegmentals consisted of word stress, sentence stress, rhythm, intonation, and weak form. Paralinguistic features included loudness, rate/tempo, smoothness, energy, and clarity. The data were analyzed using Multifaceted Rasch Analysis. The results of the study were: (1) Raters showed unique significant bias interactions with person abilities, tasks, and items. These bias interactions indicated that individual raters had unique patterns in their ratings; they rated certain persons, tasks, or items leniently or harshly. (2) Significant differences in severity were found among the six raters, but they were not peculiar to the raters' L1 backgrounds. (3) Rater-by-person bias interactions revealed that a higher percentage of significant biases was found for the persons with higher abilities. These rater-related biases suggest that employing multiple raters is important in pronunciation assessment because it helps to offset biases peculiar to specific raters. Furthermore, the findings of this study suggested that highly proficient L2 English speakers are as suitable as native speakers of English when assessing pronunciation ability.

23 July 22 8:15-10:25 **Session 7**

Design patterns in language assessment

Hua Wei, Diwakar Khanal & Robert Mislevy

Advances in research on language learning have led to improved understanding of the communicative nature of language use. Insights have been gained as to how people use language in different situations, what are salient features of these situations, what are important aspects of knowledge or skills required for successful communication, how features of situations interact with knowledge differences in individuals, and other aspects of language use that all have implications for assessment. Thus, how to organize this mass of information in representational forms that can guide the assessment design process is of special interest to language testers. Assessment design patterns were originally proposed for the assessment of science inquiry (Mislevy et al., 2003), but they can be applied in the context

of language assessment as well. A design pattern lays out considerations for building assessment tasks that address targeted aspects or situations of language use. Both characteristic and variable features of potential tasks, and alternative ways to score them, are addressed. This allows for systematic thinking about the potential observations we are interested in making and their connections with the targeted student knowledge, skill, and abilities, on the one hand, and connections with the task situations, on the other. Therefore, the use of design patterns in language assessment helps the designers structure a coherent assessment argument by making explicit each of the elements, which can be assembled into an operational assessment. In addition to an overview of design patterns, examples of design patterns for language assessment and tasks generated from these patterns are given in this presentation.

24 July 22 1:55-3:25 **Session 8**

An empirical investigation of the L2 reading comprehension skills

Hossein Farhady & Gholmraza Hessamy

Research on the componential nature of reading comprehension ability is still far from settled. While the divisibility of the construct has been questioned in a good number of studies (Rost, 1993), it has been supported in some other studies (Weir et al., 2000). More recently, research has focused on validation and relative contribution of separately identified linguistic contributory factors to the totality of reading comprehension ability (Shiotsu, 2004). Nevertheless, there is a great discrepancy between the theoretical taxonomies of reading skills and strategies found in the literature and the existing empirical findings. Besides, ample evidence supports the effect of test method on performance. Therefore, this study investigates the nature of the underlying reading skills, their relationships with and relative contributions to L2 reading ability measured by two methods. Based on a comprehensive taxonomy of 28 reading sub-skills (CTRS) derived through a thorough analysis of the components documented in the literature, a test of reading sub-skills (TRS) was developed in two methods of the expected response format, namely, 'selected response' and 'constructed-response'. This provided a context for studying the relationship between the response formats and reading test performance of the participants. After trialing the TRS, data was collected from 1606 Iranian EFL learners each answering one of the eight forms of the TRS. The data was analyzed using Exploratory Factor Analysis and Structural Equation Modeling to offer the most parsimonious model for reading comprehension.

25 July 22 1:55-3:25 **Session 8**

Exploring the relationship between item feature and students' performance on reading tests

Changjiang Wang & Lingyun Gao

A well-designed reading test should be able to diagnose students' weaknesses and help improve their reading performance. However, until in-depth relationship between item features and students' reading performance is established, this goal can not be achieved.

The objective of this study is to investigate how item features (linguistic characteristics and cognitive demands) affect students' performance using a

cognitive-psychometric approach. Test materials and item-level response data from two passage-based multiple-choice reading tests were used in this study. First, the item features of each test item were identified and coded by a group of five content experts. Then, student response data from the two tests were submitted to the DETECT analyses (Kim, 1994; Zhang & Stout, 1999). DETECT is a nonparametric statistical program designed to divide the items into clusters sharing similar psychometric features (e.g., conditional covariance). Finally, the item clusters produced by the DETECT analyses were mapped upon the item features defined by the content experts. Results demonstrated that items requiring lower-level cognitive skills, such as spotting specific information from the text, tended to cluster together. However, items requiring higher-level skills, such as drawing inferences, tended to cluster based on the reading passage. Further, results showed that cognitive demands of the items had differential effects on students' performance, but the effect of linguistic characteristics of items on students' performance was equivocal. The results of this study have implications for developing more diagnostic reading tests and providing more detailed feedback to help students improve their reading performance.

26 July 22 1:55-3:25 ***Strategies in responding to the next generation TOEFL reading tasks***

Andrew Cohen & Thomas A. Upton

Session 8

This ETS-funded qualitative study sought to describe the reading and test-taking strategies that respondents used on a prototype version of the Reading section of the *Next Generation TOEFL*. The study focused specifically on responses to the three item formats, including (1) five types of basic comprehension items, (2) three types of inferencing items, and (3) two types of reading-to-learn items. Test formats included both more traditional multiple-choice and new selected-response items that were designed to assess an examinee's ability to relate to the entire passage rather than focusing only on discrete points. The study involved the collection of verbal report data from 32 non-native subjects (Chinese, Japanese, Korean, and others). Their responses to the 13 items for each passage were digitally recorded. A sub-sample of each item type was then transcribed and coded for type of strategy use. The reading section as a whole was seen to be challenging because the passages were lengthy and the tasks required understanding of text meaning, typically at the paragraph and passage levels. Key strategies included the reading strategy of paraphrasing text meaning and the test-taking strategy of using paragraph meaning to select and discard options. A brief description of strategies used for each of the ten item types and a comparison of the three broad categories of item formats will be provided. Our findings would tend to lend validity to TOEFL's claims that this new test does in fact assess the comprehension of academic texts.

LTRC
ottawa • canada
2005

Posters

(According to the sequence of the presentations)

**Day/Time** *Poster*

1 July 20 2:00–3:50 ***Developing Arabic and Russian web-delivered listening and reading tests***

David MacGregor, Mohammed Louguit, Margeret E. Malone & Dorry M. Kenyon

Session 2 Development of tests for students of less commonly taught languages (LCTLs) and delivery of tests to such students are hampered because LCTL programs are small, scattered throughout the nation and under-resourced. Because of the dearth of appropriate tests, LCTL students often take the same proficiency tests repeatedly as they are learning a language. In addition, many students studying more difficult LCTLs may never reach very high levels of proficiency. If such students must take a long test with many items beyond their ability level, test affect can be negatively impacted. This poster describes the development and validation of web-delivered, listening and reading proficiency tests in Arabic and Russian. The tests are semi-adaptive, or adaptive not at the level of individual items but by sections (i.e., testlets). An examinee's performance on a placer test determines the level of the next testlet presented. Thus, an examinee can take the same language test several times without encountering the same items. Scores are reported as scale-scores, and equated to the ACTFL Proficiency scale. We use the software program Questionmark Perception for Web to develop, deliver and score the test. This program provides sufficient security for the test, and offers the scoring and multimedia capabilities needed. Although the poster will describe the entire test development process, including issues of test security, Web delivery, item development and adaptivity through the use of testlets, the focus will be on the logistic and statistical challenges of programming a semi-adaptive test of listening and reading.

2 July 20 2:00–3:50 ***Architects and engineers constructing an EAP placement test from grounded theory up***

Barbara K. Dobson, Mary C. Spaan & Amy D. Yamashiro

Session 2 This poster reports on both the process and product of a collaborative project between teachers and testers to redesign an advanced-level EAP structure/reading test for matriculated university students. This test is part of a placement battery for in-sessional EAP courses for advanced NNS students (TOEFL CBT 220-250). The existing structure/reading component of the battery did not adequately reflect the university's EAP curriculum, lacked face validity to test takers, and did not sufficiently discriminate among this advanced-proficiency population. The goal of the project was to design a test that incorporated the expertise of the classroom instructors, who in their role as architects, envisioned a beautiful, customized, state-of-the-art instrument, and that of the testers, who in their role as engineers, wanted a

structurally sound, reliable, efficient, and functional instrument. The poster will describe the challenging iterative process of test development. The instructors conceptualized the design of the test by specifying the language features and tasks that target this advanced EAP population, while the language testers implemented the design by translating the classroom activities into functioning test tasks. Samples of successful and unsuccessful item types will be shown, along with results of piloting and the process taken to interpret new test scores.

3 July 20 2:00–3:50 ***Introducing the modern language aptitude test for Spanish-speaking children***

Daniel J. Reed & Charles W. Stansfield

Session 2 This poster introduces the MLAT-Elementary for Spanish speakers (MLAT-ES). The MLAT-ES is intended to serve Spanish-speaking children in the United States as well as students in Spanish-speaking countries. Stansfield and Reed (2003; 2004) and Reed and Stansfield (2004) reported on the developmental stages of the MLAT-ES, field testing in Costa Rica, subsequent revisions, and the decision to norm the test on 1000+ examinees in Hispanic countries. Students in these countries were chosen for the norming study rather than students in the U.S. in order to ensure that examinees had a level of Spanish proficiency and literacy that is sufficient for the valid measurement of the construct, "language learning aptitude." Criterion-related evidence of the validity of the MLAT-ES will be reported on the poster based on correlations between the MLAT-ES and grades in language classes and teacher estimates of students' foreign language learning abilities. Potential uses of the MLAT-ES with Spanish-speaking children in the U.S. will be discussed including selection, placement and guidance. In conjunction with information from psycho-educational test batteries, background information, interviews, etc. the MLAT-ES can also be used to diagnose a foreign language learning disability. In all cases, U.S. students who do poorly on this instrument will require further evaluation in order to determine whether their low score was due to something other than poor aptitude, such as low-proficiency in Spanish, weak reading skills in Spanish, a lack of motivation, or dyslexia.

4 July 20 2:00–3:50 ***The development of a new proficiency test for Japanese EFL students***

Shoichi Ishikawa, Yuji Nakamura, Kahoka Matsumoto, Hiromi Kobayashi, Atsuko Okada & Dennis Schneider

Session 2 This poster session illustrates the development and initial validation of a proficiency test specifically designed to measure average Japanese college students' English proficiency. The test has been developed by a group of seven researchers, and it is an attempt to create a 4-skill based proficiency test that not only reflects local EFL levels and requirements but also responds to actual needs in the IT-driven, global world. The search for

appropriate and effective ways to measure students' productive skills posed a major challenge.
The first part of this exhibit shows the process of test development based on detailed specifications. In order to obtain valid constructs that represent Japanese college students' target English proficiency, three reliable sources were used: 1) the results of a large-scale questionnaire survey conducted nationwide, 2) the results of a large-scale listening comprehension test conducted in Japan and 3) an updated review of relevant literature in second language acquisition and language testing. The second part of the exhibit is a step-by-step presentation of the piloting and revision process involved in producing the test. A preliminary test with revised versions which were later developed will be shown as a testimony to our continued efforts to improve the final product. Finally, the results of initial validation studies conducted by the group will be presented, which include reliability studies and a factor analysis of the components of the test. This work has been supported by JSPS (KAKENHI), Japan.

5 July 20 4:15-6:30 **Session 3**

An online training programme for raters of academic writing

Janet von Randow

The use of online rater self-training is growing in popularity and has obvious practical benefits. This poster displays an online training programme, which focuses on the analytically-scored writing component of a large-scale academic language proficiency assessment used to diagnose learners' English learning needs. The program was designed to give raters an opportunity to restandardize their performance after a break from rating. It was not intended to replace face-to-face rater training but to reduce the number of such sessions, thus saving both time and money. Eight raters trialled the programme and individually scored a number of pre-rated benchmark scripts receiving immediate feedback in the form of a score indicating the discrepancy between their rating and that of the benchmark and a brief comment on the relevant feature of the benchmark script. Raters' views about online training were canvassed prior to their using the programme and their comments about its effectiveness were recorded after the experience. There was considerable individual variation in terms of receptiveness to the training input, with reliability gains from training more evident in those who rated larger numbers of scripts and who were positively disposed to this form of training. The poster shows the online rating process, raters' comments and the design modifications made as a result of their feedback.

6 July 20 4:15-6:30 **Session 3**

Relationships between productive vocabulary knowledge and speaking test performance using multiple measures

Rie Koizumi

Vocabulary plays an integral role in speaking a language (e.g., Higgs & Clifford, 1982; Levelt, 1993). However, there have been few studies that examine the degree to which vocabulary affects speaking performance (Noro & Shimamoto, 2003). Although some studies deal with vocabulary used in speaking performance (e.g., Adolphs & Schmitt, 2003; Read, 2004), there

are very few studies that investigate relationships between productive vocabulary knowledge (i.e., knowledge to produce a word when one speaks and writes; Schmitt, 2000) and speaking test performance, which are measured independently. Therefore, this poster aims to explore how productive vocabulary knowledge and speaking test performance are related. Information on the extent of the relationships can help understand the importance of vocabulary knowledge in speaking performance, as well as provide the basis for an empirical model of speaking test performance and for rationales of language teaching and assessment.
The participants in this study were approximately 250 Japanese secondary school students who had studied English for about three to six years. They took two tests: (a) a written test of productive vocabulary knowledge of size and of depth (i.e., derivation and association), and (b) a tape-mediated speaking test of monologues. Speaking performance derived from five tasks was assessed from four perspectives using multiple measures: fluency, accuracy, syntactic complexity, and lexical complexity (e.g., Koizumi, 2004; Wolfe-Quintero, Inagaki, & Kim, 1998). After the validity evidence of inferences and uses of each test is presented, the results and implications for language teaching and assessment are discussed.

7 July 20 ***The Treviso test***
4:15-6:30 *Geraldine Ludbrook*
Session 3 In response to the pan-European request for certification of foreign language proficiency, a project at the University of Venice has developed a test to measure the English language skills needed by graduates in Foreign Trade to obtain professional information from the Internet.
Multimedia test texts are taken from Internet trade sites; test tasks involve making or justifying decisions on the basis of information retrieved; and test administration is by computer to maximise authenticity, validity, and efficiency. The specialist lexical content of the test has been determined by the analysis of a corpus of Internet trade texts. Naïve Bayesian statistical analysis is used to determine mastery and to permit adaptive administration of test items.
The test currently focuses on the receptive skills. Future plans are to measure productive competence, remaining within the Internet domain and the field of Foreign Trade: writing e-mails and telephone interaction. Preliminary research will include an analysis of a corpus of business e-mails and a survey of the authentic tasks performed by operators in the field of Foreign Trade. In determining real-world assessment criteria for production skills, attention will be paid not only to the linguistic features of new means of communication, but also to aspects characteristic of exchanges between non-native speakers using English as a lingua franca in a cross-country business context. The degree course of Foreign Trade, for whose graduates the new test has been designed, has its campus in the town of Treviso, creating a curious link to the medieval tradition of pioneering competency tests.

8 July 21 ***Equivalence of two special purpose tests***

8:15-10:00 *Amelia Kreitzer Hope & Doreen Bayliss*

Session 4 The University of Ottawa, the self-titled "World's Largest Bilingual University," has developed a Second Language Certification Test (SLCT) to certify its students as proficient in their second official language (French or English). To facilitate score interpretation for test takers and score users (potential employers), its band levels were designed as roughly equivalent to those of the federal government's Public Service Exam (PSE). The PSE is a high-stakes test of the second language skills of employees in jobs with bilingual classifications.

Both the PSE and the SLCT are tests of language for special purposes. The federal test reflects the language of the workplace; the tasks on the SLCT assess language in academic contexts. The formats of the two tests differ somewhat, most notably in that the PSE uses an indirect method of assessing writing. A pilot study has been designed to examine empirically the relationships between the components of the two tests. 100 federal public servants (50 in each language), with recent scores on the PSE, will take the SLCT in January 2005. The subjects will also answer a questionnaire concerning their demographics, experience with the second language, attitudes towards learning and using the second language, and opinions about the different testing formats. In the poster session, the results obtained will be presented, focusing on 1) the equivalence of the test components, 2) the relationship of test results and questionnaire responses, 3) implications for policy, and 4) implications for an expanded equivalence study in 2006.

9 July 21 8:15-10:00 ***A simple framework for developing language tests on the web***

Alysse Weinberg

Session 4 We have developed an html-based framework, using Javascript and the alysse7.cgi script that easily permits creation of web-based language tests. Each test has its own html files, which are modelled on a template. This template uses a frameset main html file that includes the separate working test html file. This test file has Javascript controls to enforce the number of listenings of sound files.

This template has been used to develop placement tests, mid-term and final exams, along with proficiency tests for reading, listening and writing. Texts being worked on during the test are displayed in a frame on the page beside the questions. Different texts can be displayed during different sections of the test and a button exists to do this. The test items include multiple choice questions, true or false, drop-down selections, and fill in the blanks activities. Parameters within the test html pass information to the alysse7.cgi script so that many different tests can all use the same alysse7.cgi script. The cgi processing script scores the test and returns the number of correct answers and percentage correct score to the students. The script also writes a summary line each time students submit their test to a daily log file. A debug mode allows the test developer to make sure that the correct answers are being properly evaluated. The correct answers to the test cannot be

determined by a student looking through the html code. Both English and French versions are available although the scripts could be adapted to almost any language.

10 July 21 8:15-10:00 **Session 4**

The effects of task types on listening test performance: A retrospective study

Yo In'nami

The increasing interest in outcome-based approaches to assessment in language testing (e.g., Brindley, 1998; McKay, 2000) has heightened the need for more research on fair testing by which more valid inferences can be drawn. Among many variables related to test performance, of particular interest are the effects of task types (Chapelle, 1998; Kunnan, 2000). Despite determined efforts to investigate how task types affect test performance, past studies are limited in two ways. First, most studies (Berne, 1992; Kobayashi, 1995; Rodriguez, 2003; Shohamy, 1984; Wolf, 1993) take only quantitative approaches to this issue. Although a few studies use qualitative methods (i.e., verbal protocol analysis), they focus on independent task types (open-ended tasks, Buck, 1991; matching tasks, Ross, 1997; cloze tasks, Storey, 1997; multiple-choice tasks, Wu, 1998; gap-filling tasks, Yamashita, 2003) and do not compare multiple task types directly. Second, although Riley and Lee (1996) compare recall and summary tasks, idea units (Carrell, 1985) are used as text comprehension units.

The purpose of the present research is to examine qualitatively the effects of three task types (i.e., multiple-choice, open-ended, and summary gap-filling tasks) on listening test performance, taking the two limitations of the previous studies into account. The present study specifically addresses how task types affect the cognitive processes of taking listening tests and how their effects vary across listening proficiency levels. Participants listened to texts and immediately reported their test-taking processes. Their verbal reports were propositionized based on Bovair and Kieras (1985). Results and implications for language assessment are discussed.

11 July 22 8:15-10:25 **Session 7**

Setting and monitoring professional standards: A QMS approach

Nick Saville, Piet Van Avermaet & Henk Kuijper

This paper presents the work of ALTE in setting and monitoring the standards of their examinations. ALTE published its Code of Practice in 1994 and set out the standards that members should aspire to. In 1999 the membership began addressing the question of how the standards could be monitored and established a working group to take this forward.

The outcome has led to the conceptualization of a *Quality Management System* (QMS) and the development of self-evaluation checklists to be used by members. Recently the working group has piloted a peer monitoring system based on the notion of "auditing", derived from other QM systems (e.g. ISO 9001:2000). While seeking to set minimum standards the aim is to identify where improvements can be made and to encourage continual professional development.

In the pilot system, a member organization which is being audited has to satisfy the auditor by meeting a set of minimum standards in terms of a "quality profile" appropriate for a particular examination system (purpose, context of use, stakes, etc.). For each of the minimum standards a satisfactory argument has to be provided supported by evidence of good practice. The speakers summarise this approach and discuss implications for improving the professionalism of ALTE stakeholders, taking into account the wide linguistic, educational and cultural diversity that exists within the ALTE membership.

12 July 22 8:15-10:25 ***Developing a web-delivered K-12 test of ESL listening and speaking***

Gary Buck & Cheryl Alcaya

Session 7 This poster session will describe the development of the *Test of Emerging Academic English: Listening and Speaking.* This test was developed for a consortium of states headed by Minnesota. The construct targeted is the English ability to function successfully in a US K-12 academic environment: so called 'academic English'. The test is administered over the World Wide Web to four different grade bands: K-2, 3-5, 6-8 and 9-12. Part of the speaking section of the test is automatically scored using speech-recognition technology.

The test design raises a number of challenging and interesting issues:

- operationalizing the construct of 'academic English'
- designing computer-based tests for very young children
- designing multi-media items with audio and complex visuals in the face of considerable bandwidth restrictions
- designing items to control the interaction between visual information and information provided by the language input
- automating the scoring of speaking over the world wide web

The session will discuss the solutions chosen to these challenging issues. Specifications will be available, as will sample items, and the test can be taken online, in real time, if an internet connection is available. Data regarding student reactions to the test is being collected, and that data will be shared with participants.

Works in Progress

(In alphabetical order by presenters' surname)

#	Day/Time	Work in Progress
1	July 21 10:25-11:25	***A study of motivation and task selection for an exit e-portfolio*** *Vivien Berry*

It has recently been mandated that all students should take the IELTS test on graduation from university in Hong Kong. However, since the mean overall band scores achieved by the eight universities in 2004 only ranged from 6.06 to 6.78, IELTS test scores on their own offer a very blunt instrument for discriminating between graduating students' language proficiency. It has been suggested that, in addition to a single test score, a standardized e-portfolio system should be developed, to document students' language achievements in several areas and motivate them to continue language enhancement over time. Motivation, both extrinsic and intrinsic, concerns "wanting to learn" (Race 1995:61). A portfolio compiled for assessment purposes is extrinsically motivating (Ashcroft and Palacio 1996:31); the major challenge for educators is to develop intrinsically motivating assessments that are also valid and reliable indicators of the ability in question.

In this session I will discuss a study designed to determine how:

- differences in motivation and English language proficiency affect task selection and completion in an e-portfolio context
- tasks previously identified as desirable for inclusion in an exit portfolio motivate students
- students can present themselves in the best possible light, relative to their initial motivational orientation
- individual differences can be incorporated into a mechanism that allows both fair and valid assessment of relevant abilities required in specific contexts

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

2	July 21 10:25-11:25	***The relevance and relativity of specific linguistic features in oral performance tests*** *Sarah L. Briggs & India C. Plough*

This Work in Progress explores the dynamic relationship that exists between the different components of the speaking construct as indicated in descriptors of levels of a holistic rating scale of an oral performance test currently used to assess the language competence and effectiveness of prospective university instructors. This research is motivated by observations that the influence of these components on an evaluator's final rating varies based, in part, on the task, the rater, and the presence or absence of other components. The current work builds on research utilizing empirical methods of scale development (Fulcher, 1996; Chalhoub-Deville, 1995; Upshur and Turner, 1995) and focuses on the relative effects of specific pragmatic and

textual features of language realized during tasks that require both interactional and transactional language use. The transcribed tests of eight candidates, four of whom received ratings of "approved for teaching" and four of whom received ratings "not approved for teaching" are initially being coded for those features--syntactic downgraders and cohesive devices-noted in the aforementioned observations. Linguistic features discovered during the analysis of the transcriptions will be added to the data under review as significantly impacting the construct component in question. Discriminant analysis will be used to examine the extent to which these features contribute to the final ratings.

3 July 21 10:25-11:25 ***Cramming or teaching: The washback effect of the TOEFL in China***

Ho Fai Chau (Michael)

It has long been noted that tests, especially high-stakes language tests, exert a powerful influence on language learners and teachers. Typical examples are the negative effect of "teaching to the test" of the TOEFL and the "ceiling effect" of the consistently high TOEFL scores from Asian students. However, there is little empirical evidence to support or reject these assertions. For example, we do not know what actually happens in those well-established TOEFL preparation courses.

The study is situated in the context of the literature of test washback, specifically focusing on claims that the TOEFL preparation courses only have negative effect on students' English language learning. It adopts the ethnographic approach to investigate the washback effect of TOEFL in the Chinese context through a field study at a very well-known test-preparation school in China. By observing classroom teaching activities and interviewing TOEFL teachers and students, a detailed and vivid picture of how teachers and students prepare for the TOEFL in China will be presented. Moreover, the above issues will be discussed in connection with the pedagogic and social implications arising from the findings of the study. Arguments will be presented regarding the conflicts between the Western standard of ethical test-preparation, and the deep-rooted traditional Chinese philosophies and practices in test-preparation. Because no empirical research findings about the washback effects of the TOEFL have been disseminated since the study by Alderson and Hamp-Lyons (1998), it is expected that our knowledge of test washback can be extended through this study.

4 July 21 10:25-11:25 ***Analyzing examinees' cognitive processes in standardized reading tests: A tree-based modeling approach***

Lingyun Gao & Changjiang Wang

There is a call in the language testing and measurement communities to integrate cognitive psychology with test theory to inform test design and score validation (Mislevy, Steinberg, & Almond, 2002). One approach to achieving this goal is to model item difficulty with the cognitive processes required to answer the item (Skehan & Foster, 2001). To date, a number of models have been developed linking item difficulty to text characteristics for some reading tests (e.g., Carr, 2003). However, few models have been developed that link item difficulty to examinees' cognitive processes

required to answer reading test items (Gorin, 2002). Limited by the concepts and methods employed, the research results have been equivocal and inconsistent (Bachman, 2002). The objective of this study was to develop a cognitive model underlying the performance on the reading test items included in the College English Test (CET), a large-scale high-stakes EFL proficiency test in China. The reading test materials and response data from previously administered CET were used in this study. The tree-based regression approach (TBR; Sheehan, 1997) was used to model the nonlinear relationship between the IRT item difficulty estimates and the cognitive processes that had been identified and coded for each item. Using a recursive partitioning algorithm (Breiman, Friedman, Olshen, & Stone, 1984), TBR is powerful at classifying test items into clusters that require similar cognitive processes. Findings from this study may contribute to the construct validation and item construction of standardized reading tests, and provide meaningful feedback to EFL reading instruction and learning.

5 July 21 10:25-11:25 ***Native English speaking teachers' impact on classroom assessment abroad***

Deniz Gokcora

While there has been continuing enthusiasm by graduate students in MA TESOL programs to teach English abroad, the impact of these native-speaker (NS) instructors on curriculum and assessment in intensive English programs of overseas universities has not received enough attention. There are a growing number of universities abroad that prefer to hire NS instructors; Turkey is one of the many countries where these instructors want to teach. Depending on the needs of the host institution, these instructors perform a variety of duties, such as classroom teaching, administration, and creating tests for the program. However, in most Turkish state institutions, classroom assessment of English remains traditional, including only vocabulary, reading, and grammar knowledge. Another concern is the lack of proficiency-based testing in these skill areas. This study investigates the difference in response styles between NS EFL instructors and Turkish EFL instructors to a written essay to determine how the NS teachers reconcile the structure-based approach to EFL with communicative approaches to assessment promoted in their graduate training. EFL instructors teaching either at private or state Turkish universities in a major metropolitan area were surveyed and interviewed about their opinions of assessment in their institution, and then asked to respond to and make judgements on a short student essay. These measures specifically address how these instructors perceive their pedagogical role in the Turkish context and the impact of their duties on curriculum and assessment. This study provides insights into differing perceptions of assessment held by NS and NNS EFL instructors and suggests a necessary emphasis on proficiency-based testing in EFL teacher-development workshops.

6 July 21 10:25- ***Chinese students' experiences of two English writing tests: TWE versus LPI***

11:25 *Ling He & Ling Shi*

The present study explores the experiences of undergraduates from Mainland China with two English writing tests, a 30-minute essay test of TWE (Test of Written English) and a 300-to-400-word essay test as part of a two-and-half-hour English test of LPI (Language Proficiency Index). Both are essay tests with a maximum score of 6.0, the former serves as an entrance test for international students who speak English as a second language, whereas the latter is required for both international students and native-English-speaking students (whose final English marks from high schools are below 80 percent) if they need to take first-year English courses at universities in western Canada. Among those international students who have failed to score 5.0 on Language Proficiency Index, many are from Mainland China with a high Test of Written English score of 5.0. In order to explain how these students passed one test but failed the other, the present study analyzes interviews and writing samples of about 15 Chinese students at a Canadian university to compare their experiences of the two tests. We are now in the process of analyzing the data. The findings are expected to identify how students' successes and failures can be traced to the differences between the two writing tests and between the trainings and preparations they had for Test of Written English in China and those for Language Proficiency Index in Canada. Cross-cultural comparisons of students' experiences will generate theoretical and pedagogical implications for second language writing instruction and testing in both China and Canada.

7 July 21 10:25-11:25 ***Exploring difficulty in an L2 monologue test task***

Tomoko Horai

Monologues are an established format in high stakes oral testing globally and in classroom assessment in the EFL context. However, very little research exists on what contributes to the degree of task difficulty within monologic tasks in a speaking test. This presentation reports on an on-going project to investigate factors that contribute to task difficulty in monologic speaking tasks.

The presentation reports on score comparisons and linguistic analysis of EFL/ESL university students' oral performance under different conditions, where monologic tasks are systematically manipulated in relation to a number of variables. One condition was manipulated in each of three tasks to create three experimental tasks (planning time, planning condition, and response time). In addition to responding to these three tasks, all students performed an anchor task. A feedback questionnaire on each task was given to the participants to investigate how candidates behaved in responding to language elicitation tasks in the test of speaking. 100 EFL/ ESL students' speeches were marked by two qualified examiners. The resulting score data were analysed, and a linguistic analysis of the actual written work was carried out. The responses on the questionnaires were also investigated resulting in a triangulated set of data. The results provide findings on how intra-task variables affect students' performance and help us define the level

of task difficulty. Further research, implications for L2 pedagogy, language testers, and researchers will also be discussed.

8 July 21 10:25-11:25 ***Issues in assessment of ESOL: Washback of the skills for life strategy***

Tania Horak

The 2001 'Skills For Life' Strategy in Britain has had a profound effect on Adult Basic Education provision (covering Literacy, Numeracy and ESOL-English for Speakers of Other Languages) but there are few studies, which have looked at the assessment practices associated with the Strategy or at the effects of these practices on the ESOL classrooms.

The aims of my study are

- to clarify the range and nature of assessment practices currently used in the UK by ESOL providers
- to investigate what dictates the choice of these assessment practices
- to investigate the washback of these practices, with special reference to new ESOL exams introduced in September 2004, and through this
- to consider current treatment of the role of 'stakes' in washback studies.

I am structuring the study on the basis of Henrichsen's hybrid model of the diffusion of innovation (1989) because it provides a useful framework for understanding and tracking the progress of an innovation (in this case the Strategy) by comprehensively proposing the key elements to be considered.

Two interviews with two teachers, in three contrasting institutions, will be complemented by lesson observation and interviews with Directors of Study (DOS). The second interview will allow exploration of issues arising from the observations, the DOS data, and the preliminary analysis of the first interview. This in-depth investigation will be followed by a survey of institutions nationwide in order to identify common emerging themes and thereby determine the generalisability of my findings.

9 July 21 10:25-11:25 ***The development and validation of a corpus-based rating scale for writing***

Ute Knoch

Fulcher (2003) distinguishes between intuitive and empirically-based methods in the design of rating scales. The former method is largely based on intuition of experts, while the latter makes use of actual student performances. In recent years there has been a call for more empirically-based rating scales (Fulcher 1987, 2003; North 1995; North & Schneider 1998; Upshur & Turner 1995, 1999) as intuitive methods are often perceived to be arbitrary and inconsistent and reflect what theorists think happens in communication situations, and not what actually happens. The planned study reported on here intends to develop a corpus of student writing performances with sub-corpora of different levels of performance

using 2000 writing scripts from a large-scale diagnostic assessment procedure administered on entry to the university. A confirmatory factor analysis using the current descriptors has failed to produce evidence that the various criteria are distinct from one another and also shows that there is little justification for grouping the current subcategories together. The aim of the study is therefore to develop an improved rating scale based on discourse analytic measures derived from current writing theory. The new descriptors will be developed from the results of the corpus-based analysis of actual student writing, with an independent measure of proficiency used to divide the scripts into high, medium and low proficiency bands. The resultant descriptors will be used by raters to re-rate a body of the sample scripts and factor analytic techniques will be applied to the data in order to validate the new scale.

10 July 21 10:25-11:25

The interplay of portfolio assessment and language testing practice

Lucilla Lopriore & Guido Benvenuto

The European Language Portfolio – a tool to support the development of plurilingualism in Europe - was developed and piloted by the Language Policy Division of the Council of Europe. Different models of ELP can be devised for different contexts and learner needs, but models should be developed in conformity with the aims and principles described by the Council of Europe and in the Common European Framework of Reference. All ELP models should undergo a process of validation by the Council of Europe Language Division. Several ELP models are currently being used by a growing number of Italian foreign language students. Its use has determined a shift in language testing practice as most teachers' assessment is the result of traditional language testing and students' self-assessment. A university based research project aimed at monitoring and analysing the shift in current language testing practices after the introduction of portfolio assessment was developed and is currently being carried out in a number of Italian middle schools. The presentation is aimed at producing the first results of this research project.

11 July 21 10:25-11:25

Unraveling second language speaking proficiency

Rob Schoonen, Jan Hulstijn, Nivja de Jong, Margarita Steinel & Arjen Florijn

To provide a transparent system for language instruction and assessment in Europe, the Council of Europe produced the *Common European Framework of Reference for Languages* (2001). However, little research has been done on the validity of the framework. The aim of a project we recently started, is to develop an empirically-founded theory of speaking proficiency. Based on an 'interactionalist' view of second language assessment (Chapelle, 1998), this project aims to investigate the compositional nature of the construct of second language speaking proficiency by singling out components associated with linguistic 'knowledge', language processing 'control', and communicative setting in a cross-sectional design (language development). The project consists of three studies. In study 1, 200 learners of Dutch as a

second language at two proficiency levels and a control group of 50 native speakers will perform speaking tasks in the personal and public domain (setting), as well as a number of off-line 'knowledge' tests and on-line processing 'control' tests. With the use of structural equation modeling, the relative weight of (aspects of) linguistic knowledge and speed of processing will be assessed in speaking performance in personal and public communicative settings. In two small-scale studies, involving 30 Turkish and 30 English learners of Dutch and 15 native controls, in-depth investigations will be conducted to link assessment of speaking performance with what is known about stages of interlanguage development and about the influence of the first language on the acquisition of a second language (study 2) and with notions of fluency and automaticity (study 3).

12 July 21 10:25-11:25 ***Measuring the knowledge of text structure in academic ESL reading***

Viphavee Vongpumivitch

Although knowledge of text structure, i.e., the ability to understand hierarchy of ideas in the text, has been prominent in ESL reading research, it is an under-investigated area in ESL reading assessment (Grabe, 1999). My doctoral dissertation (Vongpumivitch, 2004) responded to this need by creating an ESL reading test with one reading passage featuring Meyer's (1985) *collection of description* text structure type, and four performance-based test tasks. The dissertation presented a structural equation model, which showed that all four tasks, namely, an incomplete outline, a graphic organizer, a summary, and a set of open-ended questions, were valid measures of this knowledge. Students' think-aloud and interview data further revealed that graphic organizer and summary tasks provided stronger evidence of validity as measures of knowledge of text structure than incomplete outline and open-ended question tasks.

The limitation of my dissertation was that, due to time constraints, only one reading passage with one type of text structure was used. Thus, I propose a research program to validate the results obtained in my dissertation, and to extend the research to include other types of text structure as indicated by Meyer (1985), namely, problem-solution, causation, and comparison. Four versions of an academic ESL test will be created, each representing different types of text structure. The study will investigate validity evidence to see if there are common answers to research questions across the four test versions. The focus of this work-in-progress session will be on the design of the testing instruments and data collection procedures.

13 July 21 10:25-11:25 ***Do test formats in reading comprehension influence ESL and non-ESL students differently?***

Ying Zheng

Bachman (1990) emphasized the importance of research into the effects on test performance of personal attributes and test method facets to ensure better and fairer measures of students. Various test items formats are used to assess reading comprehension (Anderson et al., 1991; Bachman & Palmer 1982; Shohamy 1983, 1984; Hancock, 1994; Bennett, Rock and Wang,

1991). Much attention is given to the multiple-choice item format (e.g. Freedle & Kostin, 1993; Katz, Lautenschlager, Blackburn, & Harris, 1990; Royer, 1990), comparatively less research focuses on other test item formats (e.g., constructed response) or the relationship between students' performance and test item formats.

The present study intends to looks into the different test formats in the reading comprehension component of the Ontario Secondary School Literacy Test (OSSLT). They are multiple choice, constructed response questions and constructed response questions with explanations. The purpose is to find out whether English as second language (ESL) students exhibit different performance patterns compared with other non-ESL students with regard to the different test formats used. A random sampling from each group is formed, and subscale level data will be used to investigate the discrepancy of ESL students' performance in these three test formats compared with their non-ESL counterparts. Further multiple comparisons in post hoc tests will be performed to examine the relational statistics of the data. It is hoped that this work in progress could contribute to the debate of multiple choice and constructed response questions, as well as the research in ESL teaching and learning.

Issues

(According to the sequence of the presentations)

**Day/Time**

1 July 20 10:25-12:15 ***Issues in testing listening comprehension in the multimedia web Environment***

Gary Buck

Session 1

The development of the World Wide Web has provided a new platform for the delivery of language tests, and this provide some new challenges for test developers. Traditionally, listening tests were tape administered and the test-taker was presented with a dis-embodied voice. The listener was thus deprived of the visual information that is normally available to aid comprehension. Although computer-delivered tests did use technology capable of providing a rich visual environment, generally they did not make full use of the possibilities available; the TOEFL, for example, presented just a small number of simple, still images. The World Wide Web, however, has become a visually rich multi-media environment: with audio, animations, video and so forth. This puts considerable pressure on test developers to conform to the standards of this new environment.

This paper will discuss the implications of testing listening comprehension in a visually-rich, multi-media environment. Clearly, language tests are intended to test the ability to comprehend language—i.e. oral linguistic information—not graphical information, and in order to ensure construct validity, test developers need to take into account the complex interaction between the information provided by the graphics, the information provided by the language and the information targeted by the test question.

This paper will present a discussion of the theoretical and practical issues involved in this interaction. This will be illustrated with examples of actual test items. The paper will conclude with a set of theoretically-motivated, general guidelines for creating construct-valid listening tests for the world wide web.

2 July 21 2:15-4:05 ***Measuring content: lessons from an undergraduate history course for language assessment***

Muhammad Usman Erdősy

Session 5

The study underlying my paper examined how a professor presented 'content,' communicated task requirements for written assignments, and graded and commented on students' work in a third-year course on modern Chinese history. Analyses (both qualitative and quantitative) of interviews, students' written work, and the professor's grades and responses showed that the quantity, relevance, and sequencing of students' arguments exerted the greatest influence on their grades (although syntactic complexity, command of academic vocabulary and accuracy exerted significant, if secondary influences.) Additionally, in-class observations have shown that the imparting of critical reasoning skills and critical writing skills – the staples of English for Academic Purposes programs, and two of the constructs language proficiency tests aim to assess – was entirely subsumed in

substantive, 'content'-oriented discussions.
In light of this, my presentation will discuss how such insights can be incorporated into ongoing development of the Canadian Academic English Language (CAEL) Assessment, concerning, particularly, the development of

1. reading and listening tasks that extract specific information test takers will need to utilize in their written response;
2. writing tasks that elicit the kinds of complex responses expected of test takers in their target environment; and
3. scoring rubrics that judge 'content' and 'language' as they would be judged in the target environment

Because CAEL specifically targets the measurement of performance in academic settings, such issues are central to its credibility; however, the issues raised by studies of writing assessment in university courses have implications for the development of any measures of academic English proficiency.

3 July 22 11:50-12:10

What have we learned in three thousand years?

Liz Hamp-Lyons

High stakes testing has been documented since the Chou period (1111-771 B.C.), through the establishment of a national university during the Han period (206 B.C.-220 A.D.), to a formalised system of written examinations in the national school system of the Sung period (960-1280 A.D.). In imperial China, education, notably memorization, rote repetition and written analysis, was the mark of the successful man: education led to money, power and everything else. In principle, impartiality in the examination process was ensured through a rigorous sequence of increasingly-demanding exams in which candidates and examiners were locked away together. However, in practice these ideals were marred by bribery, cheating and sometimes extreme measures such as tunneling below exam cubicles to bring in books! (Cleverley, 1985; Miyazaki 1976)
In this paper I will explore the relationship between the ancient Chinese imperial examination system and test preparation practices, and the approach to test preparation (e.g., for the TOEFL) currently storming through China. I will ask whether language testing has learned anything from the history of examination pressure and educational competition, and whether it has anything to learn from current test preparation practices that capitalize on rote memorization, content subject study strategies, and that (seemingly) short-circuit the values of 'real learning', 'critical thinking' and 'problem solving' that we wish our tests to promulgate. As very large numbers of ambitious young Chinese compete for access to Western educational opportunities, it is timely to consider how our testing programmes will respond to these aggressive and effective test preparation practices.