MwALT 2006 Eighth Annual Conference Schedule At a Glance

Day 1 (Friday, September 29)

Time	Event	
08:30 - 11:30	Workshop, room 130	Facilitator: Vanessa Rouillon, UIUC
	Scott Walters [Queens College, New York] "Cheating and Plagiarism in Language Testing"	
11:30 - 01:00	Lunch, provided for participants in either workshop	
01:00 - 04:00	Workshop, room 130	Facilitator: Vanessa Rouillon, UIUC
	John Read [Univ. of Auckland, New Zealand] "Addressing Ethical Issues in Language Testing: The ILTA Response"	
04:00 - 04:30	Break	
04:30 - 06:00	Works-in-Progress, Atrium	Chair: Fred Davidson, UIUC
	04:30 - 05:00	Presentations (about 5 minutes per person)
	05:00 - 06:00	Works in Progress: discussion tables
		Okim Kang [Univ. of Georgia] "Impact of Rater Characteristics on Speech Ratings"
		Hongli Li [Univ. of Illinois at Urbana-Champaign] "A Study on the Writing Section of College English Test"
		Rebecca Present-Thomas [Monterey Inst. of International Studies] "Validation of an Automatically-Scored Spoken English Test"
		Wang, Xin (Diana) [Univ. of Illinois at Urbana-Champaign] "Effect of Rater Training for Integrated Writing Tasks"
		Xiaoju Zheng [Northwestern Univ.] "Comparison of the Speech Rhythm of L1 and L2 English Speakers"

Day 2 (Saturday, September 30)

Time	Event	
08:30 – 08:45	Announcements, Welcomes, and Seating	
08:45 - 10:00	Keynote Address, Atrium	Moderator: Rick Partin, UIUC
	Lyle Bachman [Univ. of California at Los Angeles] "Issues in Assessing the English Proficiency and Academic Achievement of English Language Learners"	
10:00 - 10:30	Break	
10:30 - 12:30	Parallel Paper Presentations: Set 1 (20 minutes for the presentation, 10 minutes for discussion)	
12:30 - 02:00	MwALT Business Meeting and Lunch, provided to conference registrants	
02:00 - 04:00	Parallel Paper Presentations: Set 2 (20 minutes for the presentation, 10 minutes for discussion)	
04:00 - 04:30	Break; (Student Presentation Committee Meeting)	
04:30 - ??	Announcement of Best Student Presentation and Conference Closing Event: The MwALT Pizza Party, Atrium Music provided by: The Painkillers (blues)	

Parallel Paper Presentation Sessions

Set 1 A Room: ATRIUM

Time	Event
	Chair: Melissa Bowles, Univ. of Illinois at Urbana-Champaign
10:30 - 11:00	Miguel Fernández Álvarez [Cicero (IL) Public Schools] & Inmaculada Sanz Sainz [Univ. of Granada, Spain] "A new English test for university entrance in Spain: From theory to practice"
11:00 - 11:30	Lia Plakans & Maureen Burke [Univ. of Iowa] "Studying test use in placement for an intensive English program"
11:30 - 12:00	April Ginther et al., [Purdue Univ.] "Local testing and the integration of scores, instruction, and research"
12:00 - 12:30	Maria McCormack [Columbia Univ.] "Quality control of a small-scale but high-stakes essay exam"

Set 1 B Room 130

Time	Event
	Chair: Kadeessa Abdul-Kadir, Univ. of Illinois at Urbana-Champaign
10:30 - 11:00	Alison Bailey et al. [Univ. of California Los Angeles / CRESST] "Determining the measurability of K-12 academic English for test development"
11:00 - 11:30	Charles Stansfield & Alexis A. López [Second Language Testing, Inc., Rockville, MD] "Examining the validity of an alternate assessment for ELLs"
11:30 - 12:00	Michael O'Duill, [Rosenheim Univ. of Applied Sciences, Germany] "Voluntary language assessment at upper-secondary public vocational schools in Germany"
12:00 - 12:30	Jiyoung Kim [Univ. of Illinois at Urbana-Champaign] "Mixed-methods research in language testing: An overview"

Set 2 A Room: ATRIUM

Time	Event
	Chair: Jiyoung Kim, Univ. of Illinois at Urbana-Champaign
02:00 - 02:30	Melissa Bowles [Univ. of Illinois at Urbana-Champaign] & Susan Bosher [College of St. Catherine] "Reducing content-irrelevant variance in nursing examinations: Could linguistic modification be a solution?"
02:30 - 03:00	F. Scott Walters [Queens College] "Normative stability in CA-informed testing of L2 pragmatics"
03:00 - 03:30	John Read [Univ. of Auckland] "Adjusting the scale of academic English assessment"
03:30 - 04:00	Jeanne Lee [Purdue Univ.] "English intonation as a descriptor for evaluating oral English proficiency"

Set 2 B Room 130

Time	Event
	Chair: Tammy Hsu, Univ. of Illinois at Urbana-Champaign
02:00 - 02:30	Shigetake Ushigusa [Purdue Univ.] "The relationship between oral fluency and multi-word units"
02:30 - 03:00	Rui Yang [Purdue Univ.] "Factor structure of the oral English proficiency test"
03:00 - 03:30	Hui-Jeong Woo [Univ. of Illinois at Urbana-Champaign] "Factors affecting NAEP reading scores"
03:30 - 04:00	Youngmi Yun [California State Univ.Bakersfield] "Pedagogical implications of the structural equation modeling approach for L2 writing for writing instruction"

Keynote Speech

Day 2: Saturday, September 30, 2006
8:45 - 10:00 Room: ATRIUM

ISSUES IN ASSESSING THE ENGLISH PROFICIENCY AND ACADEMIC ACHIEVEMENT OF ENGLISH LANGUAGE LEARNERS

Lyle F. Bachman
University of California, Los Angeles

The Elementary and Secondary Education Act ("No Child Left Behind", or NCLB) of 2001 mandates the assessment of the "academic English proficiency" of English language learners (ELLS, aka LEPs, or limited English proficiency students) in the areas of listening, speaking, reading and writing. This assessment is mandated at all levels, from beginning to advanced, K-12, is to be conducted annually, and needs to be sensitive to annual increases in proficiency.

In order to adequately implement this mandate, a number of issues in assessing the English proficiency of ELLs need to be addressed. These include:

1. defining academic language use
2. defining academic language proficiency
3. specifying language use activities or tasks as a basis for assessment tasks,
4. distinguishing the effects of English proficiency from those of content knowledge on test performance, and
5. setting performance standards

These issues are discussed, along with some general issues with currently used assessments of English proficiency for ELLS, and an agenda of research is suggested.

Workshops

Day 1: Friday, September 29, 2006
08:30 - 11:30
Room 130

CHEATING AND PLAGIARISM IN LANGUAGE TESTING

F. Scott Walters
Queens College, City University of New York

Seemingly endemic to mainstream as well as L2 education, test cheating and plagiarism are problems that continually threaten to disrupt assessment, affecting both small- and large-scale testing contexts. It is thus incumbent on aspiring or practicing language testers to learn how to recognize and deal with these recurring problems. However, the issues are not always clear-cut, as stakeholders can come to the L2-testing enterprise with differing motivations and understandings of what constitutes cheating and plagiarism. This workshop proposes to introduce language testers to various factors and issues related to L2 test-taker cheating and plagiarism, including problems of detection and cross-cultural understandings. Participants will then explore possible solutions to cheating and plagiarism in L2 testing with potential application to the participants' own L2 teaching and testing contexts.

Day 1: Friday, September 29, 2006
01:00 - 04:00
Room 130

ADDRESSING ETHICS IN LANGUAGE TESTING: The ILTA Response

John Read
University of Auckland

Ethical concerns are at the heart of any work in the field of language testing. Language testers are very conscious of the need to maintain high standards of professional conduct, especially when test scores are being used to make important decisions about individual test-takers. Right from its inception in 1991, the International Language Testing Association (ILTA) has given priority to the development of an ethical code that would promote good conduct among its members and help to address the ethical dilemmas that language testers can face in their work. It has proven to be a challenging task to achieve the right balance between a statement of general principles and guidelines for practice in the variety of situations in which ILTA members work worldwide.

This is one of a series of workshops sponsored by ILTA at language testing conferences to stimulate discussion of ethical issues and more particularly to review the work to date on the ILTA Code of Ethics (2000) and the draft Code of Practice (2005). Attendance at the workshop is not restricted to ILTA members and indeed all those at the conference are encouraged to participate. A large part of the workshop activity will involve small-group discussion in which participants will be able to share their experiences and express their views. There will also be the opportunity to provide feedback to ILTA on the future direction of the Association's work in this area.

Works-In-Progress

Friday 29 Sep
4:30 – 6:00 PM
ROOM: Atrium

IMPACT OF RATER CHARACTERISTICS ON SPEECH RATINGS

Okim Kang, Department of Language and Literacy Education, University of Georgia
125 Aderhold Hall, Athens, Georgia 30602-7123 (okang@uga.edu)

Ratings of speaking skills are extremely susceptible to rater expectation and stereotype, because listeners are prone to rendering social inferences about speakers on the basis of just a few seconds of speech. The study attempts to ascertain the proportion of variance in speaker ratings attributable to potentially biasing rater characteristics and the proportion attributable to measurable parameters of accentedness. The study further investigates the disparity of rater effects in different samples of English speaking proficiency. To be precise, the study investigates selected factors that may affect ratings of oral English proficiency of non-native speakers (NNSs) of English by using two different samples. In Study A, 40 iBT TOEFL speech samples are rated by 75 raters, and in Study B 40 raters assess 12 International Teaching Assistants' (ITA) in-class oral presentations from each ITA's field of study. The measures for the rater background characteristics include (a) composite index of exposure to non-native English speaking friends and acquaintances; (b) sojourns in non-Anglophone nations; (c) formal training in language studies; (d) score on a reverse linguistic stereotyping task; and (e) raters' educational attainment. The objective linguistic measures via instrumentation include (f) intonational (f_0) deviation from a standard American English model; and (g) intra-pausal speech rate. Then, the study additionally seeks to determine the degree to which rater training may reduce the impact of rater characteristics on speech ratings. The results of this research from Study A and B lead to additional studies of comparing rater bias effects in iBT TOEFL ratings and in ratings of classroom teaching of ITAs. This comparison study also takes account of the effects of two different training types: (1) a course of formal rater training administered for iBT TOEFL and (2) that of rater social-psychological training (i.e., one hour interaction between raters and ITAs at an informal meeting). The study represents an innovative approach to assessing rater characteristics that are likely biasing factors in speech evaluations and comparing the impact of those "nuisance" rater effects with the impact of features of pronunciation.

A STUDY ON THE WRITING SECTION OF COLLEGE ENGLISH TEST

Hongli Li, Division of English as an International Language (DEIL),University of Illinois at Urbana-Champaign ,707 South Mathews Ave, Urbana, IL 61801 (hli24@uiuc.edu)

The College English Test (CET) is a large-scale standardized test to measure college students' English ability with the purpose of serving college English teaching in China. Accounting for 15% of the total score, its writing section is to evaluate students' ability to express themselves in written form. However the average score of the writing section has been relatively low compared to that of the other sections; even students from key universities are unable to earn an average passing score, i.e. 9 out of 15. Therefore, many researchers are concerned with Chinese college students' poor writing competence and are attempting to change this situation. The purpose of this paper is to improve the existing writing section of the CET so as to increase its validity and beneficial washback and thus generate valuable implications in the teaching of English writing in Chinese colleges. The following questions are explored: How valid and reliable is the CET writing test? Does the CET writing test have positive or negative washback and how? What could be done to improve the test so as to maximize its positive washback? What are the other factors

that influence the teaching of writing in Chinese colleges?

First the presenter will do a validation study of the CET writing test by comparing it with other equivalent writing tests. Then the presenter will carry on questionnaire surveys and interviews with English teachers and undergraduates in a university in China. Based on this investigation and the previous studies by other researchers, the presenter will illustrate the existing washback of the CET writing test and offer suggestions on how to increase its beneficial washback. Finally, the other potential factors influencing the teaching of English writing in Chinese colleges will also be studied, e.g. teaching methods, teaching materials, national policy and curriculum, and faculty qualifications, etc.

VALIDATION OF AN AUTOMATICALLY-SCORED SPOKEN ENGLISH TEST

Rebecca Present-Thomas, Monterey Institute of International Studies
530 California St., Mountain View, CA 94041 (rebeccapt@gmail.com)

This presentation will discuss the planned and ongoing validation research for the new *Versant for English* spoken English test (formerly SET-10). This test is designed to measure a test-taker's facility in spoken English. The test is unique in that it is administered over the phone or computer and scored automatically with speech recognition and speech processing technologies. Previous studies have shown that this test is a valid and reliable measure of common English proficiency. However, the test has since been expanded and research studies are currently being planned and executed to determine the reliability and validity for a new set of items to be incorporated into the new version of the test. The majority of participants in these studies will include non-native speakers of English, varied by location, native language and level of English proficiency. A smaller number of native speakers of English, varied by region and dialect, will also be included as participants.

First, a fixed form of the test will be administered twice to a number of participants to measure the test-retest reliability. The standard Versant for English is randomly generated, so another set of participants will take two versions of the test to measure the test-retest reliability with different items. In addition, studies are being planned to measure the impact of a new computer-based administration and the use of different headsets in this type of administration on the test-taker's scores. These studies will likely involve a within-subject design where each participant takes the Versant for English test under two of the following conditions: phone administration, computer administration with headset type A, and computer administration with headset type B. The order in which these are administered will be counterbalanced across participants. Finally, concurrent validity studies are also being planned to measure how closely the construct of the Versant for English test is correlated with the construct measured in other tests of spoken English, such as the Oral Proficiency Interview and the TOEFL speaking test.

The presenter will address issues of research design, participant recruitment and selection, quality assurance, data and demographic tracking and data analysis.

EFFECT OF RATER TRAINING FOR INTEGRATED WRITING TASKS

Wang, Xin (Diana), Department of Educational Psychology, UIUC, 1310 S. 6th St.
Champaign, IL 61820 (xinwang2@uiuc.edu)

It is well accepted that rater training is essential for reliable essay ratings in both large scale writing assessment and small-scale educational measurement. Previous studies have explored the purpose and effectiveness of rater trainings in the context of holistic writing assessment (Huot, 1988, 1990; Cumming, 1990; Shohamy, Gordon and Kraemer, 1992). However, few if any studies have looked at the effect of rater training for integrated writing tasks.

Unlike traditional independent writing tasks, integrated tasks such as iBT TOEFL writing tasks require examinees to integrate multiple language skills in a substantial way to respond to writing prompt.

This study looks at the rater training design and examines its effect on analytic scoring in integrated writing tasks for ESL Placement test (EPT) at UIUC.

The EPT at UIUC is an integrated writing test used to examine international students' academic writing ability in order to place them into proper ESL courses. The purpose of the current study is to (a) develop rater training materials for EPT on analytic scoring dimensions including content accuracy, (b) investigate the effect of rater training on their scale-step choice (Davidson, 1991), and (c) examine the inter-rater and intra-rater reliabilities of analytic scoring in EPT writing tasks. Both inexperienced and experienced raters are recruited to grade sample student essays before and after the rater training of regular EPT and the day-long enhanced EPT. Rater training materials will be designed in June and July and the training data will be collected at the end of August.

COMPARISON OF THE SPEECH RHYTHM OF L1 AND L2 ENGLISH SPEAKERS

Xiaoju Zheng, Dept. of Linguistics, Weinberg College of Arts and Sciences, Northwestern University
2016 Sheridan Road, Evanston, IL 60208-4090 (zheng2@purdue.edu)

It is well known that adult learners are rarely, if ever, completely successful at mastering the sound system of an L2, and researches have been done from both perceptual and acoustic perspective to find out the reasons. However, most researches only address segmental features, such as vowel space difference or consonant contrast. This paper is trying to look at one of the suprasegmental features, speech rhythm, produced by L2 speakers of English**.** Speech data are from a corpus compiled from L2 English produced by L1 Mandarin speakers, with which L1 English speakers' utterances are compared. Traditionally, languages are grouped into two large categories: stress timed, e.g. English, and syllable timed, e.g. Mandarin, although the boundaries are rather blurred and further empirical evidences are yet to be found. However, this comparison may serve as a qualitative pilot study probing into the rhythmic patterns produced by L1 and L2 English speakers, and could provide hint on further studies as to what contribute to foreign accent in terms of prosodic features. Pairwise Variability Index (PVI) is the main factor compared, percentage of vocalic interval, standard deviation of vocalic intervals, and standard deviation of consonantal intervals are also investigated to provide supplementary analysis of the rhythmic patterns. The hypothesis is that L2 speakers of English may produce different rhythmic pattern from L1 speakers: variability in vocalic intervals produced by L2 speakers are less than that produced by L1 speakers. Evidences show that the rhythmic pattern of L2 English speakers with Mandarin as native language deviate from the native pattern significantly in two aspects significantly: V% (percentage of vocalic intervals) and nPVI (normalized PVI value). Further large-scale quantitative study can be conducted to obtain more thorough understanding.

Parallel Sessions Set 1A

Saturday 30 Sep
10:30 AM to 12:30 PM
ROOM: Atrium

A NEW ENGLISH TEST FOR UNIVERSITY ENTRANCE IN SPAIN: FROM THEORY TO PRACTICE

Miguel Fernández Álvarez, Cicero (IL) Public Schools &
Inmaculada Sanz Sainz, University of Granada, Spain
869 West Buena Apt. 523**,** Chicago, IL 60613 (migfdez@correo.ugr.es)

The Spanish University Entrance Exam, the Selectividad, started being used 25 years ago. Since then, only a few changes have been made to the English component. This test is designed as school-leaving examination, but, on the other hand, is serving as a selection criterion to enter Spanish University. This test must be redesigned, new specifications should be developed, including a clear definition of contents and aims, and a new test construction protocol should be set to guarantee the quality of the test.

During the last two years I have been working on a project which objective is the development of a new test for University Entrance, based on the theories proposed by the experts. It is the desire of the project to improve the current testing system by proposing a new model and a new methodology to follow in the design of a new Selectividad.

After collecting empirical evidence of the vagueness of Selectividad results, I wrote the specifications for the new test that I have recently piloted with a group of students who have just entered the University system (those students who took the Selectividad test either in June or September). Apart from that, a second piloting took place with a group of High School students who are currently preparing for the test.

This presentation will focus on the first results of item analysis and internal reliability, at the same time that I will provide details on how the test is being validated.

STUDYING TEST USE IN PLACEMENT FOR AN INTENSIVE ENGLISH PROGRAM

Lia Plakans & Maureen Burke, ESL Programs, University of Iowa
1112 University Capitol Centre, 220 S. Clinton St., Iowa City, IA 52242-5500
(lia-plakans@uiowa.edu, maureen-burke@uiowa.edu)

One major difference between large-scale and small-scale assessment may be issues related to test use. Large-scale assessment developers experience difficulty knowing the broad uses of their tests and can only study some of the many varied contexts of use. On the other hand, in small scale assessment, the context is more readily apparent and constrained. Despite this controlled setting for use, little exists on the procedure of placement in intensive English programs (IEP), even though this type of assessment occurs in every program. This paper will describe a study on small scale assessment use in the context of a university IEP. Using practitioner research, test users in an IEP conducted this study to learn about the use of tests and other assessment sources in placement decisions. The following research questions guided the study: What factors affect test use and placement? How are tests used to make placement decisions? At the beginning of

four terms, data was collected by audio-recording placement sessions in which the program coordinator and a teacher made decisions based on a three-part placement exam, TOEFL scores, course grades, instructor evaluations, and other information. The transcribed sessions were read for patterns, which developed into subsequent coding. These codes revealed four major areas impacting test use and placement: (a) test factors (b) student factors (c) test user factors and (d) program factors (i.e. number and size of levels, curriculum, textbooks). Each area will be detailed in the paper and substantiated with data from the placement sessions. The results suggest that, in the small scale assessment context of IEP placement testing, test may be used/ interpreted in several ways during placement--as normative, criterion-referenced, and diagnostic. Furthermore, test users' experience and trust of tests can impact use. Lastly, placement decisions require nearly as much consideration of program factors as they do test factors. The paper will conclude with some discussion on recommendations for further study in small scale assessment and distinctions between this work and study of large scale assessment.

LOCAL TESTING AND THE INTEGRATION OF SCORES, INSTRUCTION, AND RESEARCH

April Ginther, Jeanne (Yu-Chen) Lee, & Rui Yang, Purdue University
302 Wood Street, 816 Young Hall, West Lafayette, IN 47907
(aginther@purdue.edu, ylee@purdue.edu, yang69@purdue.edu)

A critical gap for score users-—especially for test takers and researchers—is the one that exists between a test score and its meaning. Local testing systems in which a test is linked to instruction and test data are accessible for research can be used to bridge this gap by allowing data to be leveraged for uses beyond the provision of a score.

The Oral English Proficiency Program at Purdue University developed a local, computer-based, semi-direct test of oral English proficiency that became operational in 2001. Last year, the test was placed into a web-based network system that links registration, administration, and rating. The information gathered from all input into the system is saved as a database that can be used for score reporting, research, and instructional purposes.

In the rating component of the system, raters have access to the prompt, the scale, and sample benchmark performances. In addition, raters not only assign scores but also provide a justification for the scores they assign. These justifications serve as extra interpretative information to the score users (examinees, departments, and researchers).

The raters are also instructors in the program. For those examinees who are placed into an OEPP class, instruction begins with a review of each item response by the student with the instructor/rater. Access to the raw performance data, the review of the item responses, scores, and rater comments allows the students to understand the reasons/meaning for their scores. These negotiated interpretations of actual performance serve as the foundation for each student's instructional plan, midterm evaluation, and final evaluation. For many of the students, it's the first time that they have had the opportunity to examine their actual performance in relation to their assigned score. All data generated by the test and subsequent evaluations are accessible to researchers. Studies generated by these data will be reviewed.

This presentation will demonstrate the components of the system and how they are accessed by its users. The local system has many advantages that are not possible for large scale assessments and tests that focus primarily on placement instead of the extended and system-integrated opportunities for test score use.

QUALITY CONTROL OF A SMALL-SCALE BUT HIGH-STAKES ESSAY EXAM

Maria McCormack, Columbia University, American Language Program
504 Lewisohn Hall, Columbia University, 2970 Broadway, New York, NY 10027
(mm624@columbia.edu)

This study centers on an assessment procedure at an intensive English program (IEP) at a large, private, northeastern university. At this program, a longstanding tradition of faculty autonomy regarding the evaluation of students at semester's end was found to be problematic with one population: students from a nearby prestigious college of music, who are found to require ESL instruction before registering for courses. The college requires such students to enroll in the IEP and reach "level 7" (low-advanced) within one year. Failing to do this, students are asked to leave the college. Given the high-stakes nature of these students' evaluation, a special essay examination was introduced in 2001. The purpose of this study was to provide quality control for the examination, focusing in particular on how the raters differed in severity, and the prompts in difficulty, and to what extent this variation affected examinee scores.

The data for the study were 128 essays from four test administrations. Each essay was read by two out of six raters, with a third rater called in for adjudication. The data were analyzed using the FACETS computer program, which performs multi-faceted Rasch measurement.

Overall, the results reveal that this exam is functioning reasonably well for the great majority of examinees, and this study contributes evidence of its reliability and validity, even though variation in rater severity and prompt difficulty caused problematic differences between examinees' observed scores and "fair" scores in approximately 9% of the cases.

However, small-scale but high-stakes performance assessments of this nature are rarely expected by IEP faculty or administrators to be submitted to rigorous quality control measures. This implies that experts in language testing need to do a better job at explicating the necessity of continual questioning of any tests, large or small, which have a significant impact on examinees' lives.

Parallel Sessions Set 1B

Saturday 30 Sep
10:30 AM to 12:30 PM
ROOM 130

DETERMINING THE MEASURABILITY OF K-12 ACADEMIC ENGLISH FOR TEST DEVELOPMENT

Alison Bailey, Christine A. Ong, Becky H. Huang & Frances A. Butler Dept of Education, (CRESST) Box 951521 UCLA, Los Angeles, CA 90095-1521 (abailey@gseis.ucla.edu)

In this paper we address the central question of the conference "Small- and large-scale language assessment: Is there a gap?" with illustration from K-12 academic language proficiency prototypes developed at the National Center for Research on Evaluation, Standards and Student Testing (CRESST). Our approach to test development allows us to investigate the measurability of the academic English (AE) construct empirically. Within an evidentiary framework for operationally defining AE, linguistic analyses of ELD standards, content standards, classroom discourse, and textbooks have led to specifications for large-scale assessment of AE.

We outline the test development process and report the results of studies of prototypes for 4-6 grades. In addition, we make suggestions on the basis of these studies for small-scale (i.e., classroom-based) assessments that we argue will be needed to provide full coverage of the AE construct.

The iterative process of test development included: 1) determining task format by frequency of assessment types in textbooks; 2) specifying the construct by synthesizing evidence from linguistic analyses of ELD and content standards, textbooks (mathematics, science and social studies), and teacher talk, resulting in language demand profiles; and 3) creating prototypes aligned with profiles. Following Davidson et al. (in press), the process also incorporated an audit trail to document task modifications during three tryout phases: 1) initial tryout of prototypes to estimate duration and clarity of directions; 2) administration to 77 students in whole-class settings, and 3) verbal protocols with 19 additional students to provide greater detail of performance. The rationale for retaining or rejecting prototypes will be presented.

This process revealed which language demands were not readily measurable in a large-scale assessment framework. For example, standards that captured the complexities of discourse-level demands are not easily measured due to such constraints as whole-group administration. Collectively, these findings suggest a need for development and use of small-scale assessment tasks to augment the large-scale AE assessments currently used to measure ELD progress, and pose an important extension of current K-12 test development (e.g., provide teachers with tools to identify AE language demands and create their own formative assessments in advance of and separate from content-area assessments).

EXAMINING THE VALIDITY OF AN ALTERNATE ASSESSMENT FOR ELLS

Charles Stansfield & Alexis A. López, Second Language Testing, Inc (Rockville, MD)
6135 Executive Blvd. Rockville, MD 20852 (cstansfield@2lti.com)

The NCLB requirement that schools must demonstrate that students are making annual yearly progress (AYP) in mastering the content standards places demands on tests to be sensitive to gains in

achievement. When the regular assessment is not appropriate, an alternate assessment may be used. However, it is incumbent on the test developer to demonstrate that the alternate assessment is more appropriate than the regular assessment, if the alternate assessment is to be considered valid. Criteria for judging the validity of alternate assessments of academic achievement should include the traditional criteria of content validity (alignment to standards) and psychometric appropriateness.

This study examines the validity of the DELLA (Delaware's English Language Learners Assessment), a small-scale alternate assessment that was developed to meet the needs of ELLs whose English language proficiency precludes them from meaningful participation in the regular Delaware Student Testing Program (DSTP). The DELLA is intended to provide evidence of student progress toward achievement of the Delaware English Language Arts Standards.

The prevailing validity paradigm in educational testing calls for a comprehensive argument based on many sources of evidence (Messick, 1989). Hence, we gathered quantitative and qualitative evidence based on intercorrelations between parts and section scores on the test, correlations with other tests (LAS and DSTP), alignment to standards, alignment to the DSTP performance levels, and teacher feedback.

We suggest that judgments of the validity of alternate assessments of academic achievement for ELLs should be based on evidence that there is alignment between the assessment and the content. That is, that the alternate assessment and the regular assessment should assess the same grade-level content and that both assessments should be aligned to the same content standards. Of course the breadth and depth of coverage may be different, and it is necessary to understand the differences in order to meaningfully interpret test scores and to place students into state performance levels. We also suggest that teachers' perceptions of the validity, utility, and appropriateness of the alternate assessment for the subgroup for which it is intended be considered when evaluating the validity of an alternate assessment. Such teacher perceptions must support the inferences made with test scores, i.e. that the test measures the relevant content and that it is psychometrically appropriate for the intended population.

VOLUNTARY LANGUAGE ASSESSMENT AT UPPER-SECONDARY PUBLIC VOCATIONAL SCHOOLS IN GERMANY

Micheal O'Duill, Rosenheim University of Applied Sciences (Germany)
Fachhochschule Rosenheim Hochschulstr, 1D-83024 Rosenheim, Germany (oduill@fh-rosenheim.de)

What is the relationship between high-stake and low-stake testing and large-scale and small-scale assessment (cf. Zieky,2001) &(Purpura 2004)? What significance does the voluntary or compulsory nature of tests have (Ecclestone,2005) & (Tognolini et al., 2001) ? Do small-scale and large-scale assessment procedures cause different backwash effects on teaching (Savignon 2004) ? What relevance do findings on complementary qualitative and quantitative research paradigms have for the choice of small-scale or large-scale assessment forms (Fulcher 2003)?

This paper will review the results of a three-year evaluation of voluntary language assessment at upper-secondary vocational schools in Germany. A nationally-funded workshop-based research design was conceived in analogy to the project which led to the development of descriptors of competence levels which form the basis of assessment procedures recommended by the Common European Framework of Reference for Languages (North 2000; Council of Europe 2001).

I shall report on the results of this nation-wide research project based on the interactive participation of teachers, school management, test developers and state administers. In 1998 The Standing Conference of German Ministers of Education passed a Framework Agreement permitting the introduction of voluntary tests to certify foreign language skills at vocational schools (Standing Conference 1998). In 2001 an independent consultant delivered his scientific evaluation of the tests (de Jong 2001). In 2002 Amendments were made to the Framework Agreement (Standing Conference 2002). By the end of the academic year 2004-2005 15 of the 16 German states were involved in this certification process and 27,295 students in the

whole country had taken this voluntary examination. In 2003 the Federal Ministry of Education approved an application (Zöller 2003) for an in-depth three-year review of the test processes.

Linking our concerns to those of the conference as addressed at the start of this abstract, I shall report on the three phases of the review process of which I have been the scientific director:

Phase I: Written tests: Task type
Phase II: Written tests: Assessment procedures;
Phase III: Oral tests: Task type and assessment procedures.

MIXED-METHODS RESEARCH IN LANGUAGE TESTING: AN OVERVIEW

Jiyoung Kim, Department of Educational Psychology, UIUC, 1310 S. Sixth Street Room 226, Champaign IL 61820 (jykim@uiuc.edu)

A gap between people who are working on small and large scale assessment might be caused by the lack of multiple perspectives on assessment practices. As a way to close the gap, the current paper focuses on the mixed-method use in research. There has been a long-standing debate on the relative superiority of quantitative and qualitative research methods; however, recently the debate has shifted to questions about the complementarity of the two methods and the possible degree of cross-perspective integration (Green & McClintock, 1985). In language testing, quantitative approaches have predominantly been used; however, influenced by the emergence of the interpretive research paradigm, the communicative approach to language use and ability, and Messick (1989)'s unitary view of test validity, qualitative approaches have received attention. In spite of the increased use of qualitative approaches in language testing, little discussion has been made about how the newly introduced qualitative approaches can be effectively mixed with predominant quantitative approaches so as to increase our understanding of complex testing practices. Observing the lack of the systematic use of mixed-method approaches in language testing, Bachman (2000) commented that it is necessary to develop ways in which to utilize quantitative and qualitative approaches in a complementary fashion. With the purpose of facilitating our conversation on mixed methods, the paper consists of three parts. In the first part, the use of mixed methods in language testing research will be reviewed. Then, my study on EFL students' task representation which used a mixed-method approach will be introduced. In order to analyze quantitative and qualitative data in a mixed-method framework, the study used the analytic strategies illustrated in Li and others (2000). The mixed-method approach of my study has provided the three benefits in terms of interpreting results: 1) expansion, 2) complementarity, and 3) explanation. Based on the literature review and my experience in conducting mixed-method research, finally I present three suggestions on the use of mixed methods in language testing: 1) transparent and systematic use of mixed methods, 2) expansion of research areas where mixed methods are employed, and 3) professional training in mixed-methods.

Parallel Sessions Set 2A

Saturday 30 Sep
2:00 PM TO 4:00 PM
ROOM: Atrium

REDUCING CONTENT-IRRELEVANT VARIANCE IN NURSING EXAMINATIONS: COULD LINGUISTIC MODIFICATION BE A SOLUTION?

Melissa Bowles, Department of Spanish, Italian and Portuguese, University of Illinois at Urbana-Champaign, Urbana, IL 61801; & Susan Bosher, College of St. Catherine (bowlesm@uiuc.edu)

Recent research in both psychometrics (Sireci, Li, & Scarpati, 2003) and nursing (Johnston, 2001) has indicated that for non-native speakers (NNS) of English taking nursing exams, language may be a source of construct-irrelevant variance. Therefore, exams may not accurately measure NNS' knowledge of nursing content. One accommodation often used to level the playing field for these examinees is linguistic modification, a process by which the reading load of test items is reduced while the content and integrity of the item are maintained. Research on the effects of linguistic modification has been conducted on examinees in the K-12 population (e.g., Abedi & Lord, 2001) but is just beginning in other areas. This study describes the collaborative process by which a pathophysiology course exam was linguistically modified and provides a qualitative evaluation of NNS nursing students' comprehension of original and modified test items. Findings indicate that in a majority of cases modification improved examinees' comprehension of test items. Implications for test item writing guidelines and future research involving a score comparability study with the NCLEX exam are discussed.

NORMATIVE STABILITY IN CA-INFORMED TESTING OF L2 PRAGMATICS

F. Scott Walters, 349 Kissena Hall, Dept. of Linguistics & Communication Disorders, Queens College
65-30 Kissena Blvd., Flushing, NY 11367 (francis.walters@qc.cuny.edu)

It has been argued (Levinson 1983) that the speech-act theoretic view of language (Austin 1962) underpinning second-language pragmatics testing (SLPT) (Hudson et al. 1995; Roever 2002) is of dubious validity. Thus, it has been suggested (Walters 2003, 2004) that conversation analysis (CA), with its own data-driven approach to language use, provides a more empirically valid basis for SLPT. However, employing CA methodology in SLPT development presents challenges to test validation, e.g., CA and Language Testing each have differing approaches to the concept of "norms." Unknowns include: Using CA, can operational norms be achieved from which test tasks and rating protocols can be designed, and valid test-inferences made (Messick 1989)? How might CA-trained raters assess examinee responses via a predetermined rating scale? How would rater feedback impact operational pragmatic norms as well as pedagogical norms (Gass et al. 2002)? As an attempt to answer such questions, this paper reports on the development of a CA-informed test (CAIT) of L2 oral pragmatics, the test method synthesizing CA research findings (Heritage 1984; Schegloff 1995; Kasper 2005) with a rating proposal by Bachman and Savignon (1986). The experimental CAIT method employed two CA-trained, independent raters, who evaluated oral pragmatic responses made by non-native speakers of English. This paper offers a model of SLPT test development chiefly applicable to small-scale assessment contexts, highlighting interrelationships among operational norms, test method, and basic CA research. Quantitative and qualitative analyses of response and rater data, and suggestions for further research, are offered.

ADJUSTING THE SCALE OF ACADEMIC ENGLISH ASSESSMENT

John Read, Department of Applied Language Studies and Linguistics, University of Auckland
Private Bag 92019, Auckland, New Zealand (ja.read@auckland.ac.nz)

A key purpose for language assessment in English-speaking countries is to ensure that students for whom English is an additional language, have adequate proficiency to meet the linguistic demands of their academic studies. The standard instruments for this purpose are the major international proficiency tests, TOEFL and IELTS, which are inherently large-scale in nature. However, many institutions find the need to complement such tests with smaller-scale, local assessments of incoming students, in order to evaluate the students' needs for further ESL instruction or other forms of language support. In the New Zealand case, an additional reason for local assessment is that, by law, students from migrant backgrounds who have permanent residence in the country must be treated as domestic students and cannot be required to take an English proficiency test as a condition of university admission unless citizens must also meet this requirement.

At the University of Auckland, the response has been to establish the Diagnostic English Language Needs Assessment (DELNA) program. DELNA consists of two components: a 30-minute screening measure, involving tests of lexical and grammatical knowledge, followed where necessary by a two-hour diagnosis, including tests of listening, reading and writing. The results are reported to the students and to academic program administrators in the form of not only scores but also recommendations for appropriate forms of language support and enhancement.

This presentation will explore the relative scale of the DELNA program from various perspectives, first by highlighting ways in which the program has deliberately been designed, implemented and presented as being different from a major proficiency test like IELTS. It will also discuss issues of scale within DELNA itself through a comparison of the screening and diagnosis components. A third area for discussion is the implementation of follow-up procedures in keeping with the program ethos, which favors individual counseling of at-risk students and their voluntary participation in language enhancement activities. One further issue is how to maintain the quality of the DELNA program if current approaches lead to its establishment in other New Zealand universities, thus requiring a scaling up of the operation beyond its present institutional base.

ENGLISH INTONATION AS A DESCRIPTOR FOR EVALUATING ORAL ENGLISH PROFICIENCY

Jeanne Lee, Purdue University, 148-5 Arnold Drive, West Lafayette, IN 47906 (ylee@purdue.edu)

Intonation is often listed as a descriptor in oral English proficiency scales, which demonstrates that it is a part of a performance that influences people's perception of proficiency. However, there is no explanation or guideline of how to evaluate intonation in oral English performance. This study proposes an acoustic approach to reliably measure English intonation for the analysis of pitch patterns by native speakers (NS) of English and English as a second language (ESL) speakers who speak Mandarin as their first language (L1). Speech samples of the five NS of English from the Midwest, USA, and five Chinese ESL speakers on their performance in reading aloud and leaving a telephone message are analyzed. Intonation is measured instrumentally with PRAAT, a speech analysis computer program, on sentence nonfinal and final positions, where sentence units are reliably determined by syntax.

The preliminary findings indicate that native English speakers may use different intonation patterns for different discourse situations: the five English NS of this study prefer to use level or rising contours in sentence final positions for leaving voicemails, whereas only falling contours are used to mark sentence endings for reading aloud. The Chinese ESL speakers, on the other hand, do not make use of intonation for different discourse functions: there is a prominent use of the level and falling contours for both reading aloud and leaving a telephone message. It is argued that Chinese learners of English tend to impose the

prosodic patterns of their L1 on their second language (L2) that results in the typical Chinese accent that can be characterized with a monosyllabic rhythm with constant high level and falling tones. The results of this study provide potential guidelines on how intonation may be systematically used as a descriptor for oral English proficiency.

Parallel Sessions Set 2B

Saturday 30 Sep
2:00 PM TO 4:00 PM
ROOM 130

THE RELATIONSHIP BETWEEN ORAL FLUENCY AND MULTI-WORD UNITS

Shigetake Ushigusa, English as a Second Language, English Department, Purdue University
500 Oval Drive, West Lafayette, IN 47907-2038 (sushigu1@purdue.edu)

The present descriptive study examines the relationships between temporal measures of oral fluency, the use of multi-word units, and examinee test score on the Oral English Proficiency Test. The test is used as a screening device for international graduate students at Purdue University. The study found significant and moderately strong relationships between the development of fluency across score levels and the use of particular subsets of multi-word units.

Many researchers say building a mental store of prefabricated sequences of words (prefabs) may help L2 learners increase fluency in L2. However, there have not been studies of prefabs as a quantitative variable with respect to their relationship to ESL fluency. The quantitative study has examined the relationship between fluency and multi-word units, which are a subset of prefabs. The use of prefabs, which are stored and retrieved from memory, is associated with automaticity or proceduralization and their use is argued to reduce speech-planning time.

The study focuses on and discusses the relationships between several fluency measures and the use of multi-word units consisting of two main types: idiomatic lexical chunks and literal multi-word units. Idiomatic lexical chunks are lexicalized and non-compositional sequences of words, such as come up with, of course. Literal multi-word units are compositional sequences of words, such as talk about, before I forget. The study examined observed patterns of multi-word units and fluency as measured by temporal variables.

Four dictionaries of idioms and phrasal verbs were the primary guideline to identify idiomatic lexical chunks and literal multi-word units. Examinee response data were analyzed for temporal variables using PRAAT. Fifty examinees' responses to the test item, Advice to Student of the Oral English Proficiency Test were analyzed to address the question: What are the correlations between fluency measures and the use of multi-word units? Significant correlations that represent an interesting range of strength were found between the number and type of multi-word units, temporal measures of fluency, and test score. The correlations among these variables and the implications of these findings with respect to the development of fluency and the use of multi-word units will be discussed.

FACTOR STRUCTURE OF THE ORAL ENGLISH PROFICIENCY TEST

Rui Yang, Purdue University, 813 Young Hall, 302 Wood Street,
West Lafayette, IN 47907(yang69@purdue.edu)

The purpose of this preliminary study is to examine the factor structure of the Oral English Proficiency Test (OEPT). The OEPT is used at Purdue University to assess the oral English proficiency of prospective international teaching assistants (ITAs). Only ITAs who pass the test are allowed to conduct direct classroom teaching at the university. A Confirmatory Factor Analysis (CFA) was used to examine

the structure of the test since a one-factor model is hypothesized underlying the test structure. The sample for the factor structure study includes 641 subjects from 40 different countries and regions. For the analysis of factor invariance across test-takers from different native language backgrounds, 180 Chinese subjects and 110 Korean subjects were chosen from the overall sample of 641 subjects.

The result of the CFA confirms the hypothesized one-factor model (chi-square=87.45, GFI=0.973, CFI=0.996, RESEA=0.0523). However, errors of three items were found to covary. Further analysis using the Modification Indices provided in the LISREL program suggests an effect of test prompts. Compared to the other item prompts, the item prompts of the three items with correlated errors are less text-based. Thus the prompts provide minimal information that examinees could take advantage of in forming their responses. The one-factor model with covariance among three items was invariant across the Chinese and Korean samples chosen for the study. The covariance among the three items in the one-factor test structure implies that prompt revision is necessary.

FACTORS AFFECTING NAEP READING SCORES

Hui-Jeong Woo, Department of Educational Psychology, UIUC, 1310 S. Sixth Street Room 226, Champaign IL 61820 (hwoo@uiuc.edu)

The purpose of this study is to gain knowledge and insight into how student-and school-level factors impact English language learner (ELL) students' performance in the NAEP 4th grade reading assessment. To overcome the limitations of the multiple regression approach in modeling, this study uses the hierarchical linear modeling (HLM) approach to examine the interrelationships between the explanatory factors and the reading achievement scores in school and student levels. Study conclusions will include the production of relative explanatory results to account for ELL students' literacy as well as their reading performance in large-scaled standardized tests.

This study uses data from the NAEP 4th grade reading assessment 2005 to explore the numerous factors within the school and student levels that can have significant impacts on the scores obtained on ELL students' reading assessments. The hierarchical linear model examines the influence of students' background characteristics, school resources, and teacher quality on reading performance. By using a large-scaled, national study, this study provides insights into how the contextual factors are related to ELL students' literacy. Through hierarchical linear modeling (HLM), this study examines the interrelationships between the explanatory factors and the reading achievement scores in school and student levels. HLM offers several benefits including provision of the relationships among factors within each given level and specification of how factors at one level influence factors occurring at another.

The model is hierarchical with student and school levels and posits that students and school characteristics and their interaction shape students' literacy as well as reading attainment. Student characteristics are limited English status, SES, race/ethnicity, home literacy environment, etc. School characteristics are percent of ELL students, mean SES, class size, school location, etc. In turn, it is expected that these factors will influence the nature of students' interaction within the academic environment. ELL students' reading achievement scores are hypothesized as dependent upon the preceding factors in the model.

PEDAGOGICAL IMPLICATIONS OF THE STRUCTURAL EQUATION MODELING APPROACH FOR L2 WRITING FOR WRITING INSTRUCTION

Youngmi Yun, California State University at Bakersfield, 11014 Myers Ranch Court, Bakersfield, CA 93311 (yyun@csub.edu)

This study investigated the relationships among L2 writing performance and its explanatory variables in the context of Korean adult EFL learners. The data were collected from 351 students enrolled in Korean universities. Five variables most frequently addressed in the literature were selected: L1 writing ability, L2 language knowledge, L2 writing experience, L2 reading experience, and test preparedness. The results of the comprehensive modeling using the structural equation modeling revealed L2 language knowledge as the most important predictor for L2 writing performance with all five variables explaining 70.4% of the variance in L2 writing performance. Further, a simultaneous multi-group SEM with high and low L2 writing ability group indicated that L2 writing performance was explained by the full model to a different degree for each group as well as to a different manner across the groups.

Based on these findings, practical implications for ESL or EFL writing instruction were proposed. First of all, given the fundamental effect of L2 language knowledge and L1 writing ability on the quality of L2 writing, both of these factors should be addressed in a significant way in writing pedagogy. Since each learner's writing problems come from different sources, either from lack of L2 knowledge or from underdeveloped writing expertise, or both, writing instruction should be built on a recognition of the learners' level of L2 knowledge as well as their literacy ability in L1. Second, an extensive, guided reading program as a supplement to the writing instruction can effectively contribute to the development of writing competence including the improvement of L2 language proficiency. Third, instruction in English academic writing conventions coupled with training would be necessary to write properly according to the expectation of the target readers. Last, learners need to be given enough opportunities to use their writing expertise and linguistic knowledge in actual writing by means of frequent writing practice together with feedback given by peers or instructors.

Notes