Eighteenth Annual Language Testing Research Colloquium

New Contexts, Goals, and Alternatives in Language Assessment

Hotel Rosendahl
Tampere, Finland
July 31 – August 3, 1996
ACKNOWLEDGEMENTS

LTRC 96 Steering Committee

Lyle Bachman, LTRC 95 Program Co-chair
Adrian Palmer, LTRC 95 Program Co-chair
Ari Huhta, LTRC 96 Program Co-chair
Antony Kunnan, LTRC 97 Program Chair
Mary Spaan, LTRC 97 Program Chair

Local Organising Committee

Viljo Kohonen, University of Tampere (Program Co-chair)
Ari Huhta, University of Jyväskylä (Program Co-chair)
Sari Luoma, University of Jyväskylä (Program Secretary)
Liisa Kurki-Suonio, University of Tampere (Local Secretary)

with the assistance of

Ms Anne Saarinen and Ms Leena Sulonen, Tampere Conference Service

Abstract Vettors

Lyle Bachman, UCLA/The Chinese University of Hong Kong
Gary Buck, Educational Testing Service
Micheline Chalhoub-Deville, University of Minnesota
Caroline Clapham, Lancaster University
Janna Fox, Carleton University
Christine Jensen, University of Kansas
Liz Hamp-Lyons, University of Colorado
Christa Hansen, University of Kansas
Dorry Kenyon, Center for Applied Linguistics
Brian Lynch, University of Melbourne
Timothy Pychyl, Carleton University
Nick Saville, University of Cambridge Local Examinations Syndicate
Elana Shohamy, Tel Aviv University
Bernard Spolsky, Bar-Ilan University
Sauli Takala, University of Jyväskylä
Carolyn Turner, McGill University
We would like to thank the following organizations for their financial and practical support:

Ministry of Education  
University of Tampere  
University of Jyväskylä  
Tampere University Foundation  
Language Centre, University of Tampere  
Department of Teacher Education, University of Tampere  
Centre for Applied Language Studies, University of Jyväskylä  
City of Tampere  
Teleste Educational  

and the following persons for their assistance:

Erika Eischer, Tampere Convention Bureau  
Mika Käki, University of Tampere  
Virginia Mattila, University of Tampere  
Anja Mattila-Tolvanen, University of Tampere  
Anne Pitkänen-Huhtta, University of Jyväskylä  
Timo Repo, University of Tampere  
Helena Tapaninaho, University of Tampere  
Timo Törmäkangas, University of Jyväskylä  
Outi Veivo, University of Turku
## CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conference program</td>
<td>1</td>
</tr>
<tr>
<td>Paper and symposium paper abstracts</td>
<td>5</td>
</tr>
<tr>
<td>Poster abstracts</td>
<td>29</td>
</tr>
<tr>
<td>Work-in-progress abstracts</td>
<td>38</td>
</tr>
<tr>
<td>Roundtable abstracts</td>
<td>45</td>
</tr>
<tr>
<td>Time</td>
<td>Event</td>
</tr>
<tr>
<td>-------</td>
<td>----------------------------------------------------------------------</td>
</tr>
</tbody>
</table>
| 8.00 - 10.00 | Registration at Hotel Rosendahl  
(registration office is also open on Tuesday evening) |
| 10.00 | Opening words by Prof. Matti Parjanen, Vice Rector of the University of Tampere, and Prof. Viljo Kohonen, Chair of the local organising committee |
| 10.20 - 12.30 | **COLLOQUIUM**  
Chair: Viljo Kohonen  
Discussants: Sauli Takala  
Presenters:  
John Trim: The Proposed Common European Framework for the Description of Language Learning, Teaching and Assessment  
Brian North: The Development of a Common Framework Scale of Descriptors of Language Proficiency Based on a Theory of Measurement  
Rolf Scharer: Developing Learner Portfolios  
Michael Milanovic, Nick Saville and Lynda Taylor: A user’s guide for examiners prepared by the members of ALTE with reference to the Common European Framework |
| 12.30 | Lunch break |
| 13.30 | **PAPER SESSION 1**  
Chair: J. Charles Alderson  
Alan Davies: Theory, Research and Practice in Language Testing: Another Case of Fear and Loathing? |
| 14.00 | Brian Lynch and Liz Hamp-Lyons: Positivistic versus Alternative Perspectives on Validity within the LTRC |
| 14.30 | Andrew Cohen: Towards Enhancing Verbal Reports as Source of Insights on Test-Taking Strategies |
| 15.00 | Coffee break |
| 15.30 | **PAPER SESSION 2**  
Chair: Charlene Rivera  
Antony Kunnan: Connecting Fairness and Validation in Assessment |
| 16.00 | Carol Lynn Moder and Gene B. Halleck: Resistance vs. Accommodation: Sociolinguistic and Sociopolitical Factors in Performance Testing |
| 16.30 | Bonny Norton Peirce and Gail Stewart: Culturally Fair Task-based Assessment: The Challenge of Diversity |
| 17.00 | Break |
| 17.30 - 19.15 | Tampere sight-seeing tour by bus hosted by the City of Tampere  
(buses leave: Rosendahl 17.30; tour ends at the University Main Building) |
| 19.30 - 21.30 | Reception and get-together hosted by the University of Tampere and the University of Jyväskylä. Welcome by Prof. Olof Borg and Prof. Krista Varantola, Vice Rectors of the University of Tampere  
(Kalevantie 4, University Main Building) |
### THURSDAY, AUGUST 1

#### PAPER SESSION 3  Chair: Tom Lumley

**9.00**  
Alan Davies: Migrant Gatekeeping Through English Proficiency: Is Efficiency Replacing Equity?

**9.30**  
John H.A.L. de Jong and Fellyanka Stoyanova: Discrimination Parameters and Test Dimensionality

**10.00**  
Tim McNamara and R.J. Adams: Item Bundling: New Approaches to the Analysis of Task-related Test Items in Performance Assessments

**10.30**  
Coffee break

**11.00**  
**CONCURRENT SYMPOSIUMS**

A) Portfolio Assessment  
Chair: Margo Gottlieb  
Discussants: Elana Shohamy and Viljo Kohonen  
Presentations:  
1) Liz Hamp-Lyons: Uncovering Problems with Portfolio Assessment in the ESL Context  
2) Margo Gottlieb: A Peek into Portfolio Practices  
4) Sue Anne Spath Hirschmann and Ana Traversa: When a Portfolio Program Adopts the Portfolio

B) Diagnostic Language Testing: The Use of Rule Space Methodology  
Chair: Lawrence T. Frase  
Discussants: J. Charles Alderson and Micheline Chalhoub-Deville  
Presentations:  
1) Kumi Tatsuoka: Introduction to Rule Space Methodology  
2) Gary Buck and Kumi Tatsuoka: Towards the Building of Theory: Developing a List of Attributes Underlying Performance on a Free Response Listening Comprehension Test  
3) Hisami Saito and Masahiro Kasai: An Application of the Rule Space Approach to Provide New Interpretations of Individual Performance Differences on TOEFL  
4) Gary Buck, Kumi Tatsuoka and Irene Kostin: The TOEIC Linguistic and Cognitive Diagnostic Assessment Project: An Application of the Rule Space Methodology to Large Scale Diagnostic Assessment

**13.00**  
Lunch break  (Exhibitions 13.00 – 18.00)

#### PAPER SESSION 4  Chair: Caroline Clapham

**14.00**  
Jo Lewkowicz: Authentic for Whom? Does Test Authenticity Really Matter?

**14.30**  
Janna Fox, Bruno Zumbo and Timothy A. Pychyl: An Investigation of Background Knowledge in the Assessment of Language Proficiency

**15.00**  
Lynda Taylor and Alastair Pollett: The Reading Process and Reading Assessment

**15.30**  
Coffee break
POSTER PREVIEW & PRESENTATIONS

1) Marijke Cascallar: Development and Validation of the Modified Oral Proficiency Interview (MOPI): Impact on Adult Language Training and Job Assignments
2) April Ginther and Leslie Grant: The Influence of Proficiency, Language Background, and Topic in the Production of Grammatical Form and Error on the TWE
3) Gene B. Halleck and Carol Lynn Moden: Performance Testing of International Teaching Assistants: The Role of the Assessor
4) Liz Hamp-Lyons: Ethical Test Preparation Practice: The Case of the TOEFL
7) Dorry Kenyon: Linking Performance on a Multiple-Choice Reading Test to Verbally-Defined Proficiency Levels
8) Pavlos Pavlou: Do Different Speech Interactions in an Oral Proficiency Test Yield Different Kinds of Language?
9) Mirja Tarnanen, Tommi Salmela, Anu Halvari, and Ari Huhta: The Comparison of Different Levels and Languages in a Criterion Referenced Testing System
11) Ingrid Wijgh: On How Efficient a Test of Oral Interaction Ability Can Be
12) Elaine Wylie: Test and Non-Test Assessment for the Professional Second Language User

FRIDAY, AUGUST 2

PAPER SESSION 5  Chair: Bernard Spolsky

9.00  Erna van Hest: Self-repair as a Measure of L2 Proficiency

9.30  Carolyn Turner and John A. Upshur: Scale Development Factors as Facets of Method

10.00  Vivien Berry: Ethical Considerations When Assessing Oral Proficiency in Pairs

10.30  Coffee break

WORK-IN-PROGRESS PREVIEW & PRESENTATIONS

1) Doug Absalom and Alejandro Brice: Sharpening our Pragmatic Instruments: Additive Assessment in ESL
2) Micheline Chalhoub-Deville: Proficiency-based and Diagnostic Computer Adaptive Testing: Development Issues
3) Carol Chapelle and John Read: Toward a Framework for Vocabulary Assessment
4) Patricia Dunkel: Computer-Adaptive Testing of Second/Foreign Language Listening Comprehension: Issues to Consider When Constructing and Calibrating Multi-Media Item Banks
5) Mirja Tarnanen, Tommi Salmela and Sauli Takala: Testing Vocabulary: How Much Does the Method Affect the Scores and Grades?
6) Angela Hasselgren: The Development and Trialling of CR Tests of Spoken English in Norwegian Secondary Schools
7) Ari Huhta: Variability vs. Generalizability of Speaking Proficiency across Different Test Tasks
8) Valmar Kokkota, Hele Pärn and Leeni Simm: Development of the Test of Estonian Language for Naturalization Purposes
9) Jose Lai: Self-Assessment as a Meta-cognitive Strategy in Managing Self-directed Language Learning
10) Sari Luoma: Analysing the Language Samples Elicited in a Tape-mediated and a Face-to-face Test of Speaking
11) Charlene Rivera and Margaret E. Malone: Assessing Native Speaker Language Development in Bilingual Education Programs: The Spanish Language Assessment - Oral
<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.00</td>
<td>Lunch break</td>
</tr>
<tr>
<td>14.00</td>
<td><strong>PAPER SESSION 6</strong> Chair: Gary Buck</td>
</tr>
<tr>
<td>14.00</td>
<td>Constant Leung and Alex Teasdale: What Do Teachers Mean by Speaking and Listening? A Contextualised Study of Assessment in Multicultural Classrooms in the English National Curriculum</td>
</tr>
<tr>
<td>15.00</td>
<td>Steven Ross and John Langille: Negotiated Discourse and Interlanguage Accent Effects on a Second Language Listening Test</td>
</tr>
<tr>
<td>15.30</td>
<td>Coffee break</td>
</tr>
<tr>
<td>16.00</td>
<td><strong>PAPER SESSION 7</strong> Chair: Elaine Wylie</td>
</tr>
<tr>
<td>16.00</td>
<td>Charles Stansfield: Description and Analysis of Tests for the Certification of Court Interpreters in the USA</td>
</tr>
<tr>
<td>16.30</td>
<td>Daniel Reed and Gene B. Halleck: Probing Above the Ceiling in Oral Interviews: What's Up There?</td>
</tr>
<tr>
<td>17.00</td>
<td>Break</td>
</tr>
<tr>
<td>18.30 - 22.00</td>
<td>Short cruise and Banquet (Viikinsaari Island)</td>
</tr>
</tbody>
</table>

**SATURDAY, AUGUST 3**

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.00 - 10.00</td>
<td>LTRC Business Meeting</td>
</tr>
<tr>
<td>10.00</td>
<td>Coffee break</td>
</tr>
<tr>
<td>10.00</td>
<td><strong>PAPER SESSION 8</strong> Chair: Charles Stansfield</td>
</tr>
<tr>
<td>10.30</td>
<td>Kathryn Hill: Who Should Be the Judge? The Use of NNS Raters in the Testing of English as an International Language</td>
</tr>
<tr>
<td>11.00</td>
<td>Tom Lumley and Annie Brown: Interlocutor Variability in Specific-Purpose Language Performance Tests</td>
</tr>
<tr>
<td><strong>CONCURRENT ROUNDTABLES</strong></td>
<td></td>
</tr>
<tr>
<td>11.30 - 12.30</td>
<td>A) Technology for Language Assessment and Learning</td>
</tr>
<tr>
<td></td>
<td>Chair: Lawrence T. Frase</td>
</tr>
<tr>
<td></td>
<td>Participants: Jared Bernstein, Jill Burstein, Carol Chapelle, April Ginther, and Judith Klavans</td>
</tr>
<tr>
<td>12.30 - 13.00</td>
<td>Summary and Closing words</td>
</tr>
<tr>
<td>15.00 - 22.00</td>
<td>Post conference tour</td>
</tr>
</tbody>
</table>
Vivien Berry

Ethical considerations when assessing oral proficiency in pairs

Criticisms that traditional oral proficiency interviews do not accurately reflect realistic features of natural communication or conversation, and attempts to elicit more authentic language by designing tasks containing features of non-test discourse, have led to experimentation with alternative oral test formats such as paired or group discussion tests. However, results of psycholinguistic research focusing on individual learner differences indicate that caution should be applied in the interpretation of scores awarded to learners participating in group discussion tests since the stability of the scores can be susceptible to extraneous factors such as group affective tone or the amount of extraversion present within a group. Aspects of rater behaviour such as degrees of rater severity have also been extensively studied and reported as have investigations into the extent to which learners’ personalities and raters’ preferences towards fluency or accuracy interact and influence scores awarded on oral tests.

Building on previous research, this paper will describe a study designed to investigate the hypothesis that learners whose personalities can be characterized as either extreme extravert (E) or extreme introvert (I) will score differently on a paired test of oral ability depending on: i) the dominant characteristics of selected test tasks; ii) whether the personalities of pairs of learners are homogeneous or heterogeneous.

54 first-year students from an English medium university in Hong Kong, each of whom scored ± 1 standard deviation from the mean on a personality questionnaire (EPQ) and were therefore characterized as being either extreme E or extreme I, participated in the study. Results of a pre-course test of general language proficiency were examined and no significant differences were observed between extraverts and introverts. Same gender pairs were formed on the basis of personality yielding 9 pairs each of E + E, I + I and E + I. Each pair of students completed three test tasks of differing degrees of interactivity and were initially rated on a six-point scale on each task by a trained assessor. All interactions were recorded and further ratings were awarded by independent assessors. The recordings were subsequently transcribed, thus enabling both quantitative and qualitative analyses to be performed.

Analysis of variance revealed significant 2-way interactions between personality type and test task, thus providing considerable support for the original hypothesis. Examples of authentic discourse produced by each different personality type on each task will also be presented and salient features examined. The results of this study have disturbing ethical implications concerning the interpretability and validity of the learners’ scores as a measure of oral proficiency.

Gary Buck and Kumi Tatsuoka

Towards the Building of Theory: Developing a List of Attributes Underlying Performance on a Free Response Listening Comprehension Test

At LTRC 1995 a pilot analysis of a test of second language listening comprehension was reported. Using 35 free-response, short-answer comprehension questions, this test was chosen because in an MTMM study of listening and reading comprehension it had high reliability, showed a strong trait effect and a non-significant method effect. A theoretical model of performance on listening tests was developed from the literature, and from this a list of underlying attributes was postulated. Rule-space analysis indicated: a) 15 of the attributes predict 89% of variance in item difficulty, b) 91% of students were successfully classified into their respective knowledge states, and c) for these classified students, the attribute scores predict 94% of the total score variance.
These list of attributes make theoretical sense and behave in accordance with current understanding of how second-language listening comprehension works. However there were still 9% test-takers who were not classified in the analysis. This paper will report the second stage of the process in which the results of the analysis were used as the basis for revision of the attribute list. Through an iterative process of analysis and revision a far more powerful list of attributes was developed which provide the basis for a theory of second language listening comprehension.

Gary Buck, Kumi Tatsuoka, and Irene Kostin

The TOEIC Linguistic and Cognitive Diagnostic Assessment Project: An Application of the Rule Space Methodology to Large Scale Diagnostic Assessment

Part 1: An overview

This will review the whole TOEIC project. The Test of English for International Communication is administered internationally, with almost one million administrations per year. There are seven parts to the test divided into two sections, listening comprehension and reading comprehension. The aim of this project is use Rule Space to a) identify the attributes underlying performance on all the parts of the TOEIC, and b) set up a system to provide test-takers diagnostic information regarding their knowledge and abilities with these attributes. The project will be introduce and describes in detail.

Part 2: Listening Comprehension

The TOEIC listening section has four different item types: picture recognition, statement-response evaluation, questions on short dialogue, questions on a short mini-talk. An initial set of attributes were selected from the literature and previous research; the items were coded on each of these, and analyses run on six forms of the test cases from secure administrations in Japan. Each data set had an N-size of 5000. Interactions between attributes were identified, and further analyses were carried out. Through a number of analyses, the most parsimonious set of attributes was identified.

The analyses will be presented in detail. The attributes which account for performance will be discussed, and related to current theories of listening comprehension and language processing. The generalizability of the attributes across the six forms, and across populations, will be examined, and discussed. The attributes were then combined into higher-order attributes for the purpose of score reporting.

Part 3: Reading Comprehension

The TOEIC reading section three different item types: sentence completion, spot the mistake, and comprehension questions on reading. An initial set of attributes were selected from the literature and previous research; the items were coded on each of these, and analyses run on six forms of the test cases from secure administrations in Japan. Each data set had an N-size of 5000. Interactions between attributes were identified, and further analyses were carried out. Through a number of analyses, the most parsimonious set of attributes was identified.

The analysis will be presented in detail. the generalizability of the attributes across forms and across populations will be discussed. The attributes will be related to current theories of reading comprehension, and other research into reading sub-skills. We will discuss the combination of these attributes into more general higher-order attributes for the purpose of score reporting.
Towards Enhancing Verbal Reports as a Source of Insights on Test-Taking Strategies

This presentation will start by contrasting the three forms of verbal report—self-report, self-observation, and self-revelation, and by noting the contribution that verbal report methods have made to the understanding of test-taking strategies. The presentation will then focus on concerns about the appropriate use of such measures and about the nature of reports which include the findings from the use of such measures. The issues will include: the immediacy of the verbal reporting, the respondents’ role in interpreting the data, prompting for specifics in verbal report, guidance in verbal reporting, and the reactive effects of verbal reporting.

Considerable attention will be paid to what needs to be included in a write up of verbal report procedures. Emphasis will be given to the need for rigor in the write up of verbal report procedures so that others will understand fully what was done, be able to make comparisons to other studies, and also be able to replicate the studies. The point will be made that care in the write up can help to dispel arguments that such methodological approaches are not adequately rigorous. Since by now there have appeared a number of studies using verbal report techniques to better understand the processes involved in responding to tests, the time has come to provide greater systematicity both in the collection of such data and in the reporting of such studies through the research literature.

**********************************************************************

Alan Davies

Migrant gatekeeping through English proficiency: is efficiency replacing equity?

Australia has an unhappy tradition of ethnic discrimination with regard to migration, so much so that its record pre 1973 was often labelled ‘racist’. Since that date official rhetoric has suggested a policy of open access and equity. What is overlooked is that, almost by stealth, a new gatekeeping filter has come into operation, the filter of English language proficiency. It is now universally applied but little known to the public. Immigration policy since the 1970s has emphasised the importance of transferable skills; these include English language skills. The idea of using English in this way is simple: it is based on the belief that adequate English guarantees ease of settlement and certainty of employment. Yet there is no evidence that English makes any difference to settlement or to employment. (Hawthorne 1994).

The paper considers the threat to social justice and equity from this new filtering through English language proficiency tests (Spolsky 1995), using as examples two such tests currently in use in Australia (N=5 15,000). These are:

**Occupational English Test (OET):** designed for health professionals and now mandatory for admission to one of eleven Australian health professions.

**access test:** this test is administered overseas and is now required by most categories of NESB intending migrants.

The arguments for and against such tests are mirror-images of one another:

1. **linguistic/cultural:***
   - **for:** equal access, sharing a common community and avoidance of ethnic conflict by ensuring a threshold of communicative competence between individuals and across communities;
   - **against:** they discriminate against those lacking proficiency in English but possessing equivalent proficiencies in other languages.

2. **economic:**
   - **for:** maximising the contribution of immigrants joining the Australian work-force and at the same time ensuring that they are not disadvantaged on the grounds of lack of communication skills;
   - **against:** maintaining trading contacts and other advantages brought to Australia by speakers of languages other than English.
3. political:

for: releasing the nation-building potential of all members of society, thereby successfully combining their cultural and linguistic identities with a common basis of participation in the Australian political arena;
against: their discriminatory and stigmatising categorising of non English speaking background migrants, especially in relation to English speaking ones.

4. professional:

for: the health professions, which make use of the OET as an entry requirement, make normative English language demands parallel to their normative health profession demands;
against: the exaggerated emphasis they give to language as opposed to professional skills.

The paper presents results of a survey eliciting views of key stakeholders: government officials, politicians, employers, ethnic organisations, professional associations and candidates. On the basis of these results the 'transferable skills' rationale for English language tests is rejected and a general socio-political model of language as ethnicity (Barth 1969, Kiernan 1991) offered, which uses the myth of homogeneous 'imagined community' (Anderson 1991) to explain the language testing requirement. The likelihood of reaching agreement on a code of professional ethics for language testers is considered.

References

Anderson B. 1991 *Imagined Communities* London Verso
Barth F. (ed) 1969 *Ethnic Groups and Boundaries: the social organisation of culture difference* Bergen Universitetsforlaget and London Allen and Unwin
Hawthorne Lesleyanne 1994 *Labour Market Barriers for Immigrant Engineers in Australia* Canberra AGPS

---------------------------------------------

Alan Davies

Theory, research and practice in language testing: another case of fear and loathing?

Theory and practice in language testing, as in other areas of applied linguistics, are not easily reconciled. Just as what linguists and applied linguists do (in description, in analysis, in production and so on, Lyons 1971) is as important as what they say (in e.g. theorising), so we should attend to the practice of language testing (e.g. in constructing language tests) as well as to what language testers write and say in their theoretical modelling.

Claims of a pragmatic or authentic (or even communicative) revolution in language testing practice need to be looked at with a cold eye. Unfortunately, it is rare for theorists to make clear what the practical implications of their theories are, but when they do, by providing exemplar items and tasks, the contrast can be particularly salient (Hill and Parry 1994).

Two theoretical models of language testing are examined, the communicative model of language testing (Canale and Swain 1980) and the performance model (Bachman 1990). Both models attempt to account for the enduring problem of the knowledge-ability/skill balance in language proficiency, arguing (Bachman explicitly and Canale and Swain implicitly) for the central role of the testing event. Neither argument has had much influence on the construction of published tests, as is shown by (a) an examination of TOEFL and IELTS and (b) language test constructors' self-reporting of how it is they construct language tests. Evidence for (a) is provided by a content analysis of the tests and for (b) by a questionnaire administered to 50 language test constructors. The paper discusses the content analysis of the tests and of the questionnaire results and their relationship to the theoretical modelling of the authors quoted.

TOEFL is shown to have been highly resistant to theorising from without, while the IELTS (and ELTS before it) input from theory led only an awkward dalliance with the specificity aspect of communicative approaches, a dalliance from which the new IELTS has gratefully withdrawn. In their responses to the questionnaire, what
language testers (say they) do is to take primary account of practical considerations. That does not mean that they eschew theoretical models but it does mean that they take seriously the traditional views of stakeholders. Test construction is considered to be both a practical and a theoretical pursuit, a question of appropriate task sampling. It is concluded that a three-way distinction is needed: theoretical models - research - tests in use. The purpose of theoretical models, it is suggested, is not to relate research to practice but to explain them independently.

If theory and practice have legitimately different agendas, with research intervening, the question remains of just what is the connection between language testing research and language tests in use. The paper concludes with an attempt to answer this question and to consider the implications for applied linguistics research in general.

References:
Lyons J. 1971 Introduction to Theoretical Linguistics Cambridge: Cambridge University Press

Janna Fox, Bruno Zumbo and Timothy A. Pychyl

An Investigation of Background Knowledge in the Assessment of Language Proficiency

One of the greatest challenges to the validity of inferences drawn from a topic-based, integrated test of language proficiency is the role that background knowledge may play in test-taker performance (e.g., Alderson and Urquhardt 1983; Clapham, 1995; Hansen and Jensen, 1993). In this paper, we report on a study of the effects of background knowledge on test-takers' overall test performance as well as their cognitive processes during a topic-based test of listening.

SAMPLE
A sample of 265 students was recruited from EAP classes in a large Canadian university (n = 249) as well as from a first-year introductory Psychology class (n = 16). The participants had differing levels of English language proficiency, from non-native speakers (NNS) with modest through advanced proficiency to native speakers (NS) of English.

METHOD
Immediately before taking the listening test, roughly half of the participants (n = 129) were exposed to the topic of the test in the form of a 30-minute pre-teaching lecture and notetaking activity. The other participants (n = 136) received no prior specific exposure to the topic. As part of a think-aloud technique, twice during the lecture in the Listening test, the lecture was stopped and participants were asked: "What are you thinking about now? What are you doing now?" These prompts were inserted following two tasks: Task 1 was a guided notetaking task, which Kintsch and van Dijk (1978) would describe as a "macro-processing" task; Task 2 was an item matching, "micro-processing" task. The responses of the participants were recorded, transcribed, coded and analysed. Of particular interest was the effect that the pre-teaching lecture and notetaking activity had on the test-takers' processing during the two tasks.

ANALYSIS AND RESULTS
Chi-square and multiple regression analyses were applied to the data. The results suggest that listeners draw on background knowledge differentially in responding to a topic-based language test. The data support the notion of threshold levels. NNS participants with lower levels of language proficiency did not benefit from topic-specific pre-teaching. NNS participants with higher levels of language proficiency, who had been exposed to pre-teaching of the test topic, performed in a significantly different way from participants who had not. For example, NNS participants at higher levels of language proficiency were less likely to say that the notetaking task (i.e.,
macro-processing) was "too fast" if they had pre-teaching (chi square = 8.05, df = 1, p < 0.01). The results discuss the issue of background knowledge as a source of construct irrelevant variance and the use of test-taker processing information to inform decisions on test development and use.

******************************************************************************

Margo Gottlieb

A Peek into Portfolio Practices

Recently, in the United States, teachers throughout the entire educational spectrum (from prekindergarten through the university level) have been barraged with the notion of portfolios as a multi-purpose tool. As with any innovation, teachers and students must perceive its intrinsic value and usefulness before becoming vested in adopting the practice. With scant research on portfolios as a form of alternative assessment, practitioners, for the most part, have come to center their efforts on portfolio design, organization, and management.

The purpose of this paper is to ascertain the extent of change in teacher practices over a five year period regarding portfolio use. The sample consists of a cross section of Illinois teachers (N = 200) enrolled in a graduate level course on the Assessment of Language Minority Students. A comparison of the results of a survey administered during the 1991-92 and 1995-96 academic years will be shared.

This microcosm of the greater educational community will shed light as to the usefulness of portfolios, the kinds of information teachers feel that portfolios can yield, and the degree of implementation of portfolios by teachers working with second language learners. The role of portfolios as an expression of instruction and learning will be discussed within an assessment framework. Based on the findings, goals and contexts for language assessment for students at the elementary and secondary levels will be revisited.

******************************************************************************

Liz Hamp-Lyons

Uncovering problems with portfolio assessment in the ESL context

Authentic assessment is becoming both respectable and common in educational assessment generally, and performance assessments are increasingly conducted in language testing and reported at LTRC. Portfolios are regularly cited as the epitome of authentic/performance assessment methods.

However, the educational research literature is beginning to show us that while portfolios may solve some problems, they create others. Among these problems are reduction in generalizability (Brennan & Johnson, 1995), performance equating (Jaeger, 1995), and equity and fairness (Bond, 1995). It is becoming increasingly urgent to resolve these problems, because, as Guion (1995) says, "performance testing is now coming out of the wings and moving to center stage."

This paper reports a classroom-based case study of four nonnative English users' problem-solving strategies in a portfolio-based assessment context. The study followed the case study writers' portfolios through the scoring process, and looked at raters' problem-solving strategies as they individually and as groups (two groups of three) negotiated scores, in the context of a program with many (supposedly) equivalent sections of the same course, and with program-wide, fixed, criteria. Raising some questions about the strategies used by both writers and raters, the paper goes on to explore theoretically the potentials and difficulties for improving strategic performance in portfolio creation and evaluation.

The focus in this paper is on the theoretical issues; the case study data are used to initiate and to reflect on the issues.
References


Erna van Hest

Self-repair as a measure of L2 proficiency

Theoretical background and rationale

Self-repairs are corrections speakers make on their own initiative, without the intervention from interlocutors. Self-repair behaviour is a recurrent phenomenon in everyday speech. People can correct themselves because they think their utterance not appropriate enough (so-called appropriateness repairs), or because they have made an error (so-called error-repairs). Depending on the type of error made the error-repairs can be divided into phonological, morphological, lexical and syntactic error-repairs.

Self-repair research with children learning an L1 has pointed to a relationship between self-repair behaviour and level of language development. In these studies it was found that what gets repaired at different ages, at different stages of language development, reflects those aspects of the language the developing speaker is working on (Clark & Andersen, 1979; Evans, 1985). Since L2 speakers, just like developing L1 speakers, go through different stages of language proficiency it is interesting to find out if L2 self-repair and L2 proficiency are similarly related.

Purposes of the research

The purpose of this paper is to show how results from L2 acquisition data can provide an empirical basis for L2 assessment and, in particular, how L2 self-repair data can be used as a measure of L2 proficiency. The discussion will be based on the analysis of a large set of self-repair data collected as part of a four-year research project investigating the L1 and L2 self-repair behaviour of 30 Dutch learners of English at three different levels of L2 proficiency.

Research design and method

The self-repair data were collected by means of a story-retelling task and an interview task. Both tasks were performed in the L1 (Dutch) as well as in the L2 (English). All instances of self-repair were transcribed and classified.

Results

Statistical analysis of the L2 self-repair distributions showed significant interactions between type of self-repair and level of L2 proficiency. The data showed a significant shift from simple self-repairs (e.g. phonological repairs) to more complex self-repairs (e.g. syntactic repairs) as speakers became more proficient in the L2.

Implications

The results show that the linguistic analysis of L2 acquisition data, in this case L2 self-repair data, may be used as an extra tool, besides proficiency tests, for assessing learners' L2 competence at various stages of L2 acquisition.
These observations fit in with earlier findings by Clashen (1985) and Pienemann et al. (1987) that L2 acquisition data from spontaneous speech can provide reliable information concerning learners' developmental stages.

Kathryn Hill

**Who should be the judge? The use of non-native speaker raters in the testing of English as an international language**

The spread of English as an international language in South East Asia has implications for the testing of English in the region. This paper reports on an issue raised in the development of a four skill proficiency test for teachers of English in Indonesian (EPTI).

It is one of the tenets of major international tests of English, such as the International English Language Testing System, that assessors must either be native speakers or have 'native-like' proficiency. This practice seems to be based on an assumption that non-native speakers would not be able to assess appropriately at the highest levels on the test, which are typically defined, implicitly or explicitly, in terms of an idealized native speaker. EPTI, however, aims to assess performance against the 'local' model of English as used in Indonesia rather than a native speaker ideal. It was therefore considered appropriate within the design of the test to rely entirely upon local Indonesian raters who would have access to this local model of English. The assumption underlying both practices is that native and nonnative speakers in some sense rate differently.

The research project reported on set out to examine where native and non-native speakers rate the writing section of EPTI differently to each other. 100 writing scripts from a trial administration of EPTI were double rated using a six point analytic scale by 13 lecturers from 2 teacher training colleges in Indonesia. A sample of 30 of these scripts were later multiply rated by a group of 10 native speaker raters in Australia. Data were analysed using FACETS (Linacre, 1987-94). Following Brown (1993), the two groups were compared on a number of measures including:

- the overall harshness for the two groups
- the use of the scale steps and the proportion candidates placed above and below the cut-score by the respective groups
- differences between the two groups in terms of item difficulty and rater consistency
- evidence of rater misfit

Earlier studies relating to standard varieties of the language, have found non-native speakers of the language to rate more harshly than native speakers when using a native-speaker ideal. By contrast, this study found no statistically significant difference in harshness between the two groups. In fact, in real-life terms, a greater number of candidates were judged to be 'adequate' by the non-native speakers than by the native speakers. The two groups were generally found to be interpreting the scoring criteria in the same way and only one misfitting rater was found in each group.

The results confirmed the appropriacy of using local, nonnative speakers as assessors in English proficiency tests in contexts where the highest level of proficiency is defined in terms other than the native speaker ideal. Outcomes also provided insight into the ability of raters to conceptualise appropriate levels of performance in these contexts.
When a Portfolio Programme Adopts the Portfolio

It is a commonplace of portfolio experiences to date that it fosters autonomy and reflexiveness in students, with both teachers and students' awareness of a change in roles. But how often do programs see their nature reflected in an assessment tool and (their identity) consolidated as a result?

Our goal of enabling students to read argumentative academic text in EFL initially called for an approach to reading that would simultaneously make students learn the language, use the language to learn, and learn about the language (Halliday 1980). Immersion into authentic text met the challenge of instructing highly literate students in the Humanities with a developed metalanguage in their fields of study. Needs analysis provided necessary clues to the approach and method that would best meet general objectives, yet it is continuous self-reflection that has been crucial to the appropriation of the method by its initiators, mediators, and beneficiaries. A tendency to reflect has thus been inherent to our program for the past seven years, except we could not capitalize on it properly before the portfolio was adopted for evaluation purposes two years ago.

The recursiveness and pervasiveness of the portfolio concept has proven to be the most natural means of assessment for us all, and has therefore lent cohesiveness to the already empowering nature of the pedagogy implemented right from the start. The portfolio has given the program self-awareness, and with it a conscience/consciousness of the situatedness of our practice. It has been quite a coincidence that, as our program is coming of age, a name for its essential capacity to self-inform has presented itself and helped define its identity one step further.

The portfolio has also highlighted the sharp contrast between our methodology and the traditionally reductionist approach to instruction and to vertical program design and implementation alive and well in our academic environment. In a context of time and material constraints, the portfolio contributes a concept as well as a means of looking back and ahead. It has become a meaningful way of recording experience not only for students developing a critical eye, but also for instructors as they continue research and become facilitators. In response to the mutual caring that the portfolio has strengthened, it is the enhanced sharing that we look forward to. With autonomy achieved, our goal is now interdependence between learners, instructors and administrators.

**************************************************************************

Christine Jensen, Christa Hansen, and Samuel B. Green

An Investigation of Item Difficulty Incorporating the Structure of Listening Tests: A Hierarchical Linear Modeling Analysis

Objective

We are interested in investigating what factors affect the difficult level of items on a listening for academic purposes test that includes lectures with a series of short answer questions. For these items, it is hypothesized that item, text, and speaker characteristics affect item difficulty. In addition, it is postulated that the interaction of item and text/speaker characteristics explain why some items are more difficult than others.

Background and Rationale

In trying to look at item difficulty levels, some have looked at test taker performance (Buck and Tatsuoka, 1995; Perkins, et al. 1994). Others have looked at text/speaker characteristics that contribute to test difficulty levels (Chiang and Dunkel, 1992; Shohamy and Intar, 1991). The focus of this study will not be on test taker performance, but on the contributions of item characteristics and text/speaker characteristics to item difficulty levels. This study goes beyond these previous studies in that it considers multiple text/speaker characteristics simultaneously as well as the interaction of text/speaker characteristics with item characteristics.
We will analyze data from twelve lectures with twelve to fifteen items each. For any one lecture, the number of subjects range from 110-340. We will use hierarchical linear modeling (HEM), a statistical analytic technique that takes into account the hierarchical structure of the data, to separate out the contributions of text/speaker characteristics from those of item level characteristics. The data are hierarchical from a structural point of view in that groups of items are associated with texts; that is, items (level 1) are nested within texts (level 2).

Item difficulty levels will be the dependent variable and item characteristics -- question type, time between tested information, length of response, syntactic complexity of question, and response type (verbatim, direct, constructed) -- will be the level 1 predictors. At this level we can identify item level characteristics that make unique contributions to item performance. Text characteristics -- type of discourse marker, placement of marker, technical level/difficulty of vocabulary, syntactic complexity, and density of propositions and speaker characteristics -- rate of delivery, pauses, number of repetitions, and types of repetitions -- will be the level 2 predictors. We can examine the contribution of these level 2 predictors on item difficulty. Finally, we will assess whether text/speaker characteristics moderate the relationship between item characteristics and item difficulty levels.

Findings and Implications

Hierarchical linear modeling allows us to tease out the unique contributions of text, speaker, and item characteristics and their interactions. These empirical findings can inform test constructors of how text/speaker characteristics affect performance on items.


Discrimination Parameters and Test Dimensionality

Background and Rationale

Language testing is involved in a debate on the number of factors implicated in language performance. Most test results are reported in the form of a single expression of the raw score or some transformation of the raw score and are therefore implicitly based on the assumption of unidimensionality. Researchers using Item Response Theory (IRT) disagree on the number of item parameters needed to estimate the probability of a correct response given the ability of a subject. Models have been proposed using one, two or three parameters. Whereas some claim that the one parameter (Rasch) model should be preferred because of its property of sufficient statistics, others argue that a two or three parameter model may be used if the data do not fit the one parameter model.
Purpose of the research

This research was undertaken to investigate the relationship between the dimensionality in a set of items and the number of parameters needed to obtain statistical fit of the data to a measurement model and the implications for particular groups of subjects taking the tests.

Research design and method

Data were gathered from over one thousand candidates taking the listening section of a language examination testing proficiency in a second language. Candidates were learners of the second language and represented a number of ethnic minorities. The data were fitted to two unidimensional IRT models, the one parameter Rasch model and a model allowing the discrimination parameter to vary, a two parameter model. To test the unidimensionality assumption underlying both models factor analysis, DIMTEST (Stout, 1987), variance in item correlations and cluster analysis were used. Finally, differential items functioning (DIF) was investigated for various subgroups of the candidate population.

Results

The study shows that obtaining fit for a two parameter unidimensional IRT model is insufficient to retain the assumption of unidimensionality. Introducing a discrimination parameter seems to absorb dimensionality in the data thus leading to possibly biased estimates of student ability. DIF analyses of suggest a relation between a nuisance variable and ethnicity. On the other hand, a dataset that fits the one parameter model leads to test results that can be interpreted as the expression of a unidimensional underlying trait and that are invariant across subsets of items.

Implications

The dimensionality discussion in language testing research may profit from new insights in the relationship between IRT models and the robustness of the assumption of unidimensionality. From a test fairness point of view deviations of unidimensional measurement should be considered with extreme caution.

Antony Kunnan

Connecting Fairness and Validation in Assessment

Interest in fairness in language tests has shifted from cross-cultural ESL testing two decades ago (see Briere, 1968, 1973; Briere & Brown, 1971) to gender, native language, and ethnicity based DIF studies more recently (see Chen & Henning, 1978, Zeidner, 1986, 1987; Kunnan, 1990; Ryan & Bachman, 1992). Another research interest in fairness that emerged at about the same time as the DIF studies was that of field specific/major field and ESP testing (see Alderson & Urquhart, 1983, 1985; Hale, 1988; Olman, Stricker & Barrows, 1988; Douglas & Selinker, 1993, Clapham, 1993). These different research interests can be collectively included under the topic area of fairness studies as the overall goal of all these studies is to bring about fair tests. However, these studies (with two exceptions: Hale, 1988; and Olman, et al., 1988) primarily focussed on test items and therefore did not explicitly bring into their analyses and discussions the critical connection between fairness and validation in language assessment; a connection that is crucial in bringing about fair assessment practices.

This study will use exploratory factor analyses and structural modeling to present multiple-group analyses of test performance data of a large sample (n=1000) from eight cities around the world collected for the comparability study of the Cambridge-TOEFL test batteries (Bachman et al., 1995). Through the multiple-group analyses of test performance on the separate tests, the study will focus on the internal structure of the tests for these groups. Additionally, these analyses will show that multiple-group analyses are necessary when tests are used world-wide in multilingual and multicultural contexts. The results from these analyses will be the basis for a discussion of the construct validation of the tests for each of these groups and also be a forum for the discussion of generalizability boundaries among different populations. In this manner, the critical connection between fairness and validation will be demonstrated.
Constant Leung and Alex Teasdale

**What do teachers mean by Speaking and Listening? A contextualised study of assessment in multicultural classrooms in the English National Curriculum**

Theoretical background and rationale

The National Curriculum for England (Education Reform Act 1988) requires all school pupils to be assessed at 7 years of age ("Key Stage 1" in the National Curriculum). No special testing arrangements are prescribed and the Level Descriptions for Listening and Speaking are applied during ordinary classroom activities. The Level Descriptions are used to assess both mother tongue English speakers and ESL learners. Given the multicultural composition of many classes in English schools, it is not entirely clear how these Level Descriptions enable teachers to make fair assessments of both types of learner.

**Purposes of the research**

The key question addressed is, "How do teachers arrive at overall assessment decisions for ESL and mother tongue students?". And specifically:
- which of the performance features described in the Level Descriptions are important for teachers?
- which performance features not described in the Level Descriptions are important for teachers?
- what are the relationships between the features identified by teachers as being important?
- to what extent is there a common set of criteria which primary school teachers in England use to make assessments?
- to what extent do teachers use different criteria to those contained in the Level Descriptions for (i) mother tongue students (ii) ESL students?

Research design and methods

Edited video footage of mother tongue and ESL learners engaged in normal classroom activity is shown to groups of teachers and they are asked to rate the performances using the National Curriculum Level Descriptions in order to establish baseline evaluation data. Different panels of teachers are then asked to rank order the performances according to individual criteria distilled from components of the Level descriptions, from a survey of ESL scales in use among 70% of the Local Education Authorities in England (Leung forthcoming) and from ESL assessment literature. They are also asked to judge the relevance of individual criteria. The video clips are also transcribed and analysed.

**Results**

The data from the teacher judgement sessions are analysed using discriminant analysis to ascertain which individual cluster of criteria best predict the final score awarded, correlational techniques are used to ascertain the degree to which individual criteria co-vary and descriptive statistics are used to estimate the degree of importance accorded by teachers to particular criteria. An attempt is also made to link features from the analysis of student performance to specific criteria recognised as important by teachers.

**Implications and the results**

The findings have implications for assessment practice within the National Curriculum. Issues of standardisation and the relevance of Level Descriptions are discussed with reference to the findings. However, most powerfully, there exists the possibility, by extending the research, of developing from data a theoretical model of language performance which is contextualised to the distinctive uses of language of pupils in school settings. Such a model could then be used to underpin assessment (and perhaps also pedagogical) practices.
References


**Jo Lewkowicz**

**Authentic for whom? Does test authenticity really matter?**

This paper takes a critical look at our present understanding of authenticity in language testing. It briefly reviews some of the ongoing work in the field, including that by Bachman and Palmer, and then goes on to suggest that the language testers' perspective on the question of authenticity has been too constrained. Much of the research in the field has been by language testers, for language testers. As a result, the question of authenticity has been given much prominence and it is now considered one of the key attributes of any test, on a par with reliability and other aspects of validity. That does not mean that present testing philosophy requires all tests to be authentic, but rather that it proposes that in designing any test a 'desired level of authenticity' needs to be ascertained at the outset.

But what is test authenticity? Is it a single attribute of a test or an attribute of the component parts of a test, that is, test input, test tasks as well as test outcomes? Initial proponents of authenticity including Morrow (1979) and B. J. Carroll (1980) argued test input and tasks had to be authentic. However, Bachman (1991), among others, has also stressed the need for authentic test outcomes. Yet the question remains whether stakeholders in the testing process, including testers themselves, can distinguish between authentic and inauthentic texts, tasks and outcomes and if they can, whether these attributes appear to be of the same importance in the testing process for the various parties involved, that is, test developers as well as test takers.

The above questions will be addressed with reference to a study, the results of which showed that test takers do not necessarily perceive authentic texts, tasks or outcomes as important attributes even when these reflect teaching that has preceded the test and test takers' performance on such tests does not necessarily differ markedly from their performance on other test types.

It will be argued that in designing a test the question of authenticity may not be as important as previously considered. If it is not possible to distinguish the authentic from the inauthentic, then it may not be necessary for testers to place a great deal of weight on ascertaining a 'desired level of authenticity.' Other attributes such as test practicality, in reality, may be of more importance in designing appropriate tests which are acceptable to the test-taking community.


**Tom Lumley and Annie Brown**

**Interlocutor Variability in Specific-Purpose Language Performance Tests**

Role play is a commonly used task type in occupation-specific language performance tests. Role plays in these contexts are designed to simulate professional communication and typically generate interaction between a candidate, in his/her professional role, and an interlocutor, in the role of client or consumer of the professional services. The role plays are generally developed using specifications based on expert analysis, and may also in some cases involve occupational experts in the writing of the materials.
The issue of interlocutor behaviourvariability has been investigated in relation to general-purpose interview-based tests (Ross, 1992; Ross and Berwick, 1990; Wigglesworth, Williams and Morton, 1995). However, there has been little investigation of behavioural variability in the role plays commonly used in specific-purpose tests. There appears to be an assumption that the issue of variability in interlocutor behaviour can be largely dealt with by providing explicit task instructions and interlocutor training.

This study reports on a series of workshops held with occupational informants as part of the validation process of two occupation-specific language tests. It draws on data from the oral interaction component of two occupation-specific tests, the Occupational English Test (McNamara 1990, Lumley and McNamara 1995), a 4-skills test for overseas-trained health professionals (with attention in this study to nurses), and the Japanese Test for Tour Guides (Brown, 1994) an oral/aural test for non-native Japanese speakers who wish to work in the Australian tourism industry.

10 representatives of each relevant profession (i.e. tour guides and nurses) provided reactions to audiotaped data from live administrations in each test, regarding the extent to which the tasks reflect the demands of authentic professional communication. This included consideration of the realism of the task content (as presented in the written materials) but concentrated particularly on the realisations of the tasks. Features considered included interlocutors’ interpretations of the role defined in the task, task equivalence, amount of support given to candidates and the influence on the interaction of the subject knowledge of both candidates and interlocutors (who may not have the relevant occupational knowledge).

The study found that although the examples of test interaction examined were generally perceived by the occupational informants as authentic representations of professional communication, substantial variability was observed in interlocutor behaviour, leading to differences in the challenge faced by test candidates. A significant problem for test developers who rely on performance tests of this nature is raised, that of fairness to candidates. Implications for the selection and training of interlocutors for occupation-specific language tests as well as for the issue of test validity and task design are also discussed.


Brian Lynch and Liz Hamp-Lyons

**Positivistic versus Alternative Perspectives on Validity within the LTRC**

Discussions of new perspectives on reliability and validity have been frequent in the educational measurement literature for the past several years (Cronbach, 1989; Linn, Baker, & Dunbar, 1991; Messick, 1994; Moss, 1994; Wolf, Bixby, Glenn, & Gardener, 1991). In an attempt to describe the range of perspectives on validity within the language testing community, Hamp-Lyons and Lynch (1995) investigated the available abstracts for papers
Presented at the Language Testing Research Colloquium (LTRC) over the past sixteen years. They found evidence that research in language testing has been dominated by a perspective on validity best characterized as “positivistic.” There was also occasional evidence, however, of alternative perspectives on validity—from those who remain consistent in this perspective over the years to those who have made a shift, however temporarily, from the positivistic perspective.

This paper seeks to verify the initial impressions of the Hamp-Lyons and Lynch (1995) study of LTRC abstracts, by examining a selection of complete papers. These papers have been selected in order to sample equally from the Hamp-Lyons and Lynch (1995) data set, those clearly identified as positivistic, those clearly identified as alternative, and those ambiguous in terms of a perspective on validity. In addition to a close reading and a Toulmin (1977) analysis of ten papers from each of these three categories, four authors from each category have been identified and will be interviewed.

The detailed analysis of complete papers will focus on looking for more evidence of validity perspectives than was accessible in the abstracts, in order to answer the question:
Are the initial impressions of validity perspective and of “whose voices are represented”, as determined from the abstract, verified upon an analysis of the complete paper?
The interviews will be structured as assisted introspections designed to have the researchers reflect on their research practice and the assumptions underlying it. Through close study of the recordings and field notes of the interviews, understanding of the ways each researcher is viewing and framing talk about their research will emerge. Looking at the interviews singly and as a group, a better understanding of the following issues should be gained:
To what extent do LTRC members see their work as being situated in a particular research paradigm?
How does the nature and degree of awareness of (a) research paradigm(s) for their work affect LTRC members’ understanding of and approach to validity?

As the language testing community examines new contexts, goals, and alternatives in assessment, it is important that a clear understanding of the basis for evaluating these new directions is established. As well as being open to new formats for assessment, language testers will need to be open to new research paradigms and new frameworks for validity. This study will provide a foundation for understanding the influence of our choice of research design (and the implicit or explicit research paradigm) on our determination of the validity of new forms of assessment and, thus, the constraints on new contexts, goals, and alternatives within language testing.

REFERENCES

Tim McNamara and R.J. Adams

Item bundling: new approaches to the analysis of task-related test items in performance assessments

A long-recognized problem in the analysis of data from performance-based language tests is the issue of the appropriate analysis of bundles of items which are linked to particular tasks. For example, items in reading tests may be passage-dependent; in writing and speaking tests, multiple aspects of performance on a single task may be rated.
Typically, the effect of task-dependence is frequently overlooked, although where design considerations permit, approaches using Generalizability Theory may be used to analyse the effect of tasks (e.g. McNamara and Lynch, in press).

The recent development of a further generalized Rasch model, known as the random coefficients multinomial logit (RCML) (Wilson and Adams, 1995), allows a new approach to this problem. It permits the analysis of bundles of items, where items lack the independence requirement of standard item response models.

Data from two performance assessments, the Speaking and Writing Modules of the access: test, a test of ESL for certain restricted categories of intending immigrants to Australia, are used to investigate the application of this new development. The Speaking Module consists of 7 tasks, where for each task ratings are given for a number of dimensions relevant to that task. Data from 83 subjects each rated by 4 raters were available for analysis. The Writing Module consists of 3 tasks, each rated on 4 dimensions of performance; data from 49 subjects rated by each of 19 raters were available.

Two analyses of each data set are reported: one in which the task dependence of items is ignored, and one in which the task dependence of items is recognized. Gains in validity achieved by acknowledging the existence of item bundles are discussed.

References


Carol Lynn Moder and Gene B. Halleck

Resistance vs. Accommodation: Sociolinguistic and sociopolitical factors in performance testing

The trend toward more authentic communicative testing (Shohamy & Reves, 1985; Bachman 1990, 1991; Henning 1992) has led to a broadening of the factors that are considered to be justifiably assessed in testing situations to include both sociolinguistic and strategic competence. However, these competences are generally construed to narrowly include only those aspects which are directly related to the language generated in the test situation. For example, Bachman (1990) characterizes sociolinguistic competence as the sensitivity to and control of language conventions appropriate to the language use context and strategic competence as the test-taker’s assessment of the elements of the situation in light of his or her competence in order to plan and execute the task at hand in an effective manner. This view of sociolinguistic competence adopts a Labovian perspective of viewing language variation as conditioned externally by the situation. However, an alternative view, that of Communication Accommodation Theory (Giles, Coupland & Coupland, 1991) may be more appropriate in certain test situations. In this view, variation in linguistic performance results from patterns of convergence or divergence with the speech or expectations of the addressee. This convergence is in turn dependent on a variety of socio-psychological factors relevant to the discourse participants.

This paper investigates the relevance of accommodation strategies to the performance of international teaching assistants (ITAs) on a screening test for prospective teaching assistants at an American University. These strategies are especially relevant to the testing situation because the task requires the test-taker to make a five-minute teaching presentation to a group of Americans which includes faculty and student evaluators. The diversity of activities required by the teaching situation and the test task itself requires not only language skills and communication strategies but also an understanding of how American classrooms work and how American undergraduates think (Hoekje & Williams, 1992; Kaplan, 1989; Sadow &; Maxwell, 1983; Shaw & Bailey, 1990) ITAs must then be able to use this sociocultural knowledge to develop appropriate accommodation strategies. Subjects for the study
were six ITAs all of whom were rated as having relatively weak language skills. Half of the subjects were given favorable overall evaluations, while the other half were not. An investigation of the accommodation strategies of the ITAs indicated that the more successful ITAs were those who compensated for their weak language skills by adopting strategies which accommodated more to the expectations of the American addressees, such as telling jokes to establish rapport or employing a more interactive discourse style. However, interviews with the ITAs and observations made by the teachers during an ITA training course indicated that use of these strategies was not dependent on sociolinguistic knowledge alone. The ITAs who did not use these strategies did have the necessary linguistic ability and knowledge of the situational context. However, they chose not to accommodate to the sociocultural demands of the situation, often as a result of strong feelings of personal, social, or political identity. Thus, success on the test was due, at least in part, to the sociopolitical orientation of the ITAs.

The findings of this study suggest that the range of factors assessed in performance tests of this type is much broader than current views of sociolinguistic and strategic competence would indicate. It is incumbent on language testers to re-evaluate their approaches to assessment in the light of these sociopolitical ramifications.

Brian North

The Development of a Common Framework Scale of Descriptors of Language Proficiency Based on a Theory of Measurement

Context

This paper reports results from a Swiss National Science Research Council project which was set up in the aftermath of the Council of Europe Symposium "Transparency and Coherence in Language Learning: objectives, assessment and certification" hosted by Switzerland in November 1991, which recommended the development of a Common European Framework for Language Learning and a related European Language Portfolio to report learner achievement in terms of a common scale. Using the descriptive categories in the Council of Europe Framework and the content of existing scales of language proficiency as starting points, the project has developed and calibrated a "descriptor bank" which could be used to profile achievement in different aspects of both "real life" and "interpersonal/ability" proficiency in the Swiss educational system, and to provide an illustrative set of scale descriptors for the consultative draft of the Council of Europe Common Framework.

Switzerland is in many respects an interesting test-bed for experimentation with the development of a common framework scale. Switzerland is a multi-lingual country with four language regions, each with a distinct pedagogic as well as linguistic culture, which gives the Swiss educational system a distinct pluralism and offers an opportunity to investigate the effect of demographic variables (educational sector, language region) on the interpretation of descriptor difficulty in relation to learner ability. Secondly, the concepts of scales of language proficiency and of descriptors of proficiency were novel to almost all the teachers who took part in the project making it possible to investigate the interpretations of teachers from different contexts to the same formulations without those reactions being significantly coloured by familiarity or training.

Problems with Proficiency Scales

Most scales of language proficiency appear to have been produced by appeal to intuition and to those scales which already exist rather than to theories of linguistic description or of measurement. In addition, different aspects of proficiency are commonly collapsed together into holistic paragraphs in which sentences have only relative meaning with an impression of progression often produced by alternating qualifiers like some and many; or adjectives like good and moderate. Whilst this approach may be appropriate in the development of an in-house system for a specific context with a familiar population of learners and assessors it is questionable whether it can be regarded as criterion-referencing and it has been much criticised in relation to the development of national framework scales. One type of response to these validity problems has been to limit proficiency scaling to qualitative aspects of language use demonstrated in the specific test context, scaling content either through an analysis of representative discourse produced in the test or through an algorithm deriving descriptors from the judged adequacy of responses to test tasks. Such approaches are difficult to apply to the development of a metasystem which by definition lacks a
specific test context, and in addition such approaches tend to focus on assessor- or diagnosis-oriented aspects which may be of more interest to insiders as opposed to real life task completion which may be more relevant to what has been called the strong view of proficiency taken by outsiders (user-oriented) as well as to the process of the design of syllabus, activities, continuous assessment and self assessment instruments (constructor-oriented). All these perspectives appear relevant to a meta-system.

Methodology

A descriptor pool totalling approximately 2,000 items describing aspects of spoken interaction produced by analysing and supplementing available proficiency scales was refined and edited through workshops with teachers from different educational sectors into a pool of 400 stand-alone criterion statements which could be sorted to the appropriate category and which were stated to be clear, useful and relevant.

This pool of criterion descriptors was used to create a set of overlapping questionnaires which were employed by 100 teachers to rate 945 learners of English and scaled using the Rasch rating scale model. Rater severity was estimated with the program FACETS and taken into account in estimating ranges of achievement for different educational sectors.

A new descriptor pool for listening and reading was established and refined in a similar way and used to create a series of overlapping questionnaires for a second study (a) replicating the first study (b) anchored to the descriptor bank produced in the first study (c) extending the areas studied to receptive skills as well as interaction, to French and German as well as English, and - to a limited extent - to self assessment as well as to teacher assessment.

Bonny Norton Peirce and Gail Stewart

Culturally fair task-based assessment; The challenge of diversity

This paper addresses the question: "To what extent can a task-based language assessment instrument be culturally fair to diverse adult learners?" In order to address this question, the presenters will draw on their experience as the test development team of the new Canadian Language Benchmarks Assessment (CLBA). This work, which was contracted to the Peel Board of Education by Citizenship and Immigration Canada, is informed by the work undertaken by the National Working Group on Language Benchmarks NWGLB), a group representing the interests of various stakeholders in adult language learning and teaching, including learners, teachers, administrators, and settlement agencies. The paper focuses on the challenges the test development team faced as it attempted to develop task-based reading, writing and listening/speaking assessment instruments that would be relevant to the needs of adult newcomers to Canada and culturally fair to the diverse newcomer population.

The presentation will be divided into three stages. In the first stage, the presenters will provide an overview of the test development project, which took place from April 1995 to March 1996. They will focus on the steps they took to address concerns about the possible cultural bias of the test and will summarize the important contributions of a Cultural Advisory Group and the NWGLB. The presenters will provide samples of tasks that were considered to be culturally unacceptable and indicate how they addressed some of the diverse concerns of reviewers. In particular, they will describe how they attempted to involve as many stakeholders as possible in the test development process in an attempt to share ownership of the test.

In the second stage of the presentation, the presenters will describe and analyse the results of the piloting process, in which over 1000 learners in different parts of Canada (Vancouver, Ottawa, Toronto and Halifax) participated in a nationwide pilot study. The presenters will focus their discussion on findings that had a bearing on questions of cultural fairness, and will discuss how they addressed these issues. In addition, they will draw distinctions between findings on the respective reading, writing, and listening/speaking tests, addressing how the conditions of administration may have affected the cultural fairness of the instruments.

In the final stage of the presentation, the presenters will draw together their observations and analyse their quest for cultural fairness in task-based assessment. They will argue that it is difficult to separate a task from a cultural
context. They will suggest that one way of addressing cultural fairness in testing is to develop the notion of "cultural accessibility". The notion of cultural accessibility presupposes that although not all learners may be familiar with a particular cultural context, they may still be able to access the meaning of that context via carefully constructed prompts, in written, visual, or oral mode.

**************************************************************

Pirjo Pollari

Could Portfolio Assessment Empower EFL Learners? Portfolios in EFL Teaching in Finnish Upper Secondary School

How could assessment assist learning in addition to assessing it?

Learner-centredness, self-direction and self-assessment are some of the keywords of contemporary educational discussion in Finland. Learning is currently seen as a process of knowledge and meaning construction based on the learner's own activity and on his or her prior learning. Emphasizing the student's own active role in learning requires a paradigm shift from teacher- or textbook-centred "transmission of knowledge" to a more student-centred and self-directed approach as well as a shift from a basically quite uniform curriculum towards more individual curricula. The paradigm shift in learning and teaching also calls for more process-oriented and authentic assessment as well as students' self-assessment as part of the assessment procedures. Could portfolios as a learning and assessment tool empower learners and promote both learner-centredness, self-direction and students' self-assessment?

In spring 1994 we - three upper secondary school teachers and I as a researcher - tried out portfolios in the teaching of English as a new vehicle for teaching, learning and assessment as well as a means for students to negotiate their own syllabuses within a given framework. The participating students (c. 100) were all in the second grade of the upper secondary school, aged 17 or 18. The topic area of the portfolio course was culture (mainly in the sense of literature, music, arts, film, theatre, etc.).

In addition to studying English and some aspects of culture, one of the central ideas of the portfolio experiment was to promote learner-centred and self-directed learning. We wanted the students to take greater responsibility for their own work but also to have the freedom and power to make decisions concerning their studying. We also wanted them to learn to set their own goals as well as to assess their work and learning and state their criteria. And, most importantly, we wanted our students to feel ownership of their learning and of their abilities to communicate in English. In other words, we wanted to empower our students as learners of English.

The presentation will discuss the findings of the portfolio experiment based mainly on students' portfolios, their self-assessments, comments and their reflections on their own learning. The following questions will be dealt with: Did portfolios both as a learning and assessment tool foster students' empowerment? Did student-centredness and self-direction enhance the quality and meaningfulness of learning? How did the students react to portfolios as a vehicle for assessment? How did they see self-assessment? Did student-centredness, self-direction as well as self-assessment influence on students' motivation and self-concept as language learners and users? How did the students see the role of portfolios as a tool for learning and assessment compared to their other English studies at school?

Keywords: portfolios; student empowerment; learner-centred and self-directed learning and assessment; self-assessment

**************************************************************
Probing Above the Ceiling in Oral Interviews: What's Up There?

Although a number of researchers have examined interrater reliability with respect to the original Intergency Language Roundtable (ILR) oral interview format as well as the ACTFL OPI format and have demonstrated high reliability (Adams 1978; Bachman & Palmer 1983; Carroll 1967; Clark 1978; Halleck 1995; Ke & Reed 1995; Magman 1987; Shohamy 1983), the issue of interrater reliability deserves further empirical attention, especially with respect to the performance of the tester and its effect on the outcome of the test.

This paper focuses on an important aspect of oral interviewer style and its relationship to examinee performance and interview outcome: the level of difficulty, or "pitch", of the line of questioning and topics discussed. The analysis presented is based primarily on ratings and transcripts from 42 interviews that were conducted in accordance with the ACTFL Guidelines, a corpus consisting of interviews with 21 examinees by two ACTFL-trained interviewers. Each examinee was interviewed twice, once by each interviewer. Reliability was high according to the traditional measures of correlation and percentage of agreements within one scale rating, but there were instances of disagreement that crossed major borders between ACTFL's broad bands of functionally-defined, verbally-described proficiency. The directions of rating differences were analyzed to see whether they corresponded to the pitch of the interviews. Results are discussed in terms of implications for the status of the concept "major border", interview test design, and rater training.

Steven Ross and John Langille

Negotiated Discourse and Interlanguage Accent Effects on a Second Language Listening Test

The testing of second language listening comprehension is commonly conducted under the assumption that the nonnative listener is an auditor of the target language rather than a participant in the speech event. Analyses of interactions between native speakers of a language and adult learners of second language consistently reveal however, a disparity between how language is used natively, and how it is accommodated to the interlocutor. The present research focuses on the impact of non-native accents of participants in interaction and the effect of participant negotiation on dialogues containing novel lexical items, and how accents and accommodation impact on second language listening test design. A quasi-experimental design was employed with two parallel forms of a listening comprehension test each containing two dialogues (with or without negotiated discourse) between a native speaker or Australian English and an ESL speaker (Thai) interlocutor. Fifty native speakers of Australian English, 314 Finns, and 274 Japanese subjects were exposed to either the baseline or negotiated discourse forms. The results of the main study and a replication concur in supporting the hypothesis that negotiation of discourse will aid listeners in comprehending more propositions embedded in the stream of speech. The comprehensibility of the Thai speakers was also assessed by logit scaling propositions both the Australian and Thai speakers uttered, based on their frequency of citation in written summaries provided by listeners. Accent differences were examined with repeated measures ANOVA controlling for negotiation and dialogue content. The results of the accent analysis indicated that Finns and Japanese listeners did not find the non-native (Thai) speakers more difficult to comprehend. This study supports the notion that when listening tests are devised to simulate authentic interaction in international or multicultural contexts, a native speaker-based model is insufficient and that non-native speakers should be included as participants in the interaction from which test content is drafted.
HISAMI SAI TO and MASAHlRO KASAI

An Application of the Rule Space Approach to Provide New Interpretations of Individual Performance Differences on TOEFL. The analyses will focus on the reading comprehension section of the Test of English as a Foreign Language (TOEFL).

The study consists of the following two stages: (1) identifying attributes which underlie the examinee performances on the reading section of the TOEFL, (2) examining the appropriateness of the identified attributes. Twenty-one attributes have been identified through the think-aloud procedure. The Rule Space analysis will be applied to obtain individual attribute scores. The appropriateness of these attributes will be cross validated by calculating the r-square between the attributes and the total score and IRT theta, and examining the predictability of each attribute. These attributes will be refined further by more detailed error analysis. The results will give both examinees and instructors valuable information regarding the examinee's current cognitive skills.

Charles Stansfield

Description and Analysis of Tests for the Certification of Court Interpreters in the USA

English is the unofficial language of the United States and court proceedings must be conducted in English in all USA locations except in Puerto Rico and on American Indian reservations. However, even in Puerto Rico, federal court proceedings must be conducted in English. Defendants, plaintiffs, and witnesses who do not speak English must be provided with an interpreter appointed by the court. Often, incompetent interpreters have been appointed, with the result that testimony given in foreign languages has been distorted by the court interpreter. On appeal, the findings in several cases have been reversed after it was ruled that distorted information rendered by the interpreter affected the outcome of the case. Due to this problem, a dozen states and the federal government have developed tests to identify and certify interpreters who can handle the linguistic demands of proceedings in state and federal courts. These tests have generally been developed by interpreters, without input from language testing specialists or measurement professionals.

The paper will present the results of a descriptive study of state and federal tests for court interpreters. The study was funded through a Mellon Foundation fellowship awarded to the researcher. The study began in September 1995 and ends in August 1996. The study will identify the tests that have been developed and describe and compare the test formats. The standards represented by the various cut-scores that have been established will also be compared, after considering the different rating criteria, scales, and whether a compensatory or noncompensatory approach is used to determine passing performance.

Although the study is still in progress, information gathered thus far shows that such tests have been developed in ten languages. The standard of competency demanded by the tests varies widely, thus providing evidence that systematic differences exist in the quality of court interpreting in different jurisdictions. This situation has implications for equity and justice. Concerns relating to reliability, validity, and practicality are often given inadequate attention. Normally, little statistical information is compiled on the tests other than pass rates.

This descriptive study will bring to light the problems with current practice in court interpreter testing in a linguistically diverse nation.
Introduction to Rule Space Methodology

Statistical pattern recognition and classification theory is a methodology for assignment of physical objects, or events, into one of several predetermined categories. The rule space methodology is an extension of these techniques for diagnosing cognitive attributes underlying test performance. In technical terms, the methodology consists of two stages: (i) extracting the features of a cognitive domain and then determining classification groups (or latent states); (ii) classifying test-taker response patterns into one of the predetermined latent states. The classification with the highest probability value is considered as the test-takers' latent state.

In practice this involves using expert judgments to determine what skills, knowledge or abilities (attributes), are necessary to successfully perform each item (in terms of language testing this is typically such as vocabulary knowledge, syntactic knowledge, inferencing skills, etc). We then construct a matrix of items by attributes, and using Boolean algebra we match that with the matrix of item responses. If the attributes selected by the experts do in fact explain performance, and we can classify students into their respective knowledge states, it is then possible to give each test-taker a score on each of the attributes. Thus, we can provide test-takers diagnostic information about the state of their knowledge on all relevant abilities necessary to perform on the test.

Lynda Taylor and Alastair Pollitt

The Reading Process and Reading Assessment

Validity in reading comprehension assessment

Most reading tests bear little resemblance to any authentic language activity except doing reading tests. In particular it is difficult to think of any natural reading activity that involves the kind of close reading followed by questions requiring written responses that are typical in so-called reading tests.

In this paper an argument is mounted for the method called Summary Completion on the grounds that it both resembles natural reading to a remarkable degree and that it is capable of producing tests that are highly reliable and valid.

Summary completion

The technique has been invented independently several times; for example by Mossenson, Hill and Masters (1987), Pollitt & Hutchinson (1987), Matthews (1988), Courchene (1993), and by the authors of some reading schemes familiar to many British primary schools.

The technique has usually been called 'Summary cloze', but it is clear on a moment's consideration that the process of completing such a test has little to do with the test methods properly called cloze. There has, however, been little effort to validate the method empirically as this paper sets out to do.

Fundamental theoretical and practical problem

In practice, and for good theoretical reasons, the main problem is: What is the best summary of a given text? It is argued that a partial answer can be found in the concept of contextualisation, by which the student's reading of a text is embedded in a purposeful activity: the purpose constrains what can be accepted as an adequate understanding.
Test construction

The paper will report an investigation of adolescents' comprehension of a narrative and a discursive text in their native language. A propositional and macropropositional description of typical summaries and retellings of each text was used to guide the construction of summary completion tasks for younger children.

Analysis

The study shows that such tests achieve unexpectedly high levels of internal consistency; further investigation identifies an effect which compromises the local independence assumption basic to all common measurement models. It is argued that this finding is true for all integrated tests, including cloze, C tests and even - perhaps - every text-based on the comprehension of a text. Any text comprehension test may, if it is good enough, necessarily violate measurement principles.

Of course this may not have serious practical consequences. Comparison of the results with National Curriculum Test results and teachers' assessments of the children's abilities suggests that the scores achieve a satisfactory level of validity in a very short time.

John J.M. Trim

The proposed Common European Framework for the description of language learning, teaching and assessment

The Council of Europe is at present engaged in the development of a Common European Framework for the description of courses, materials and syllabuses for language learning, teaching and assessment. The Framework is intended to promote the more effective exchange of information among practitioners, and also to encourage them to re-examine their existing practice by considering the full range of options open to them in respect of objectives and methods. It is not intended to presuppose or advocate any one approach, but to offer a comprehensive, transparent and coherent apparatus, which will also be non-dogmatic, multi-purpose, flexible, open, dynamic and user friendly.

The Framework is based on an analysis of language use and the language user in terms of the strategies employed by users (including learners) to activate general and more strictly communicative competences in order to carry out the activities and execute the processes involved in the production and reception of texts dealing with particular themes, so as to be able to fulfil the tasks facing them in the situations arising in the various domains of social existence. Categories are developed in some detail for each of the parameters identified above. Options are discussed and illustrated in detail for the description of levels of achievement in respect of each parameter. Methodological options for language learning and teaching are also present.

A draft Framework has been produced and circulated widely to relevant institutions for comment using an accompanying questionnaire. Responses are being analysed and concrete suggestions for improvement being considered. The revised document will then be further considered by an Intergovernmental Conference and submitted to European authorities for endorsement, following which an extensive programme of pilot applications will be organised.

Carolyn Turner and John A. Upshur

SCALE DEVELOPMENT FACTORS AS FACETS OF METHOD

Empirically derived, binary-choice, boundary definition (EBB) scaling is a new alternative procedure for rating samples of language performance. EBB scales are developed by teams working with samples of learner
performance (Turner & Upshur, forthcoming). These procedures have been shown to yield high reliability for brief tests, and to be efficient in time for training raters and for rating language performances. Claims have also been advanced for content relevance of EBB scales (Upshur & Turner, 1995). Rating scales can be conceptualized as a set of method facets (Bachman, 1990). The effects of some of these scale-related facets have been investigated (e.g., characteristics of raters by Brown, 1995, task being rated by Chalhoub-Deville, 1995). Factors associated with scale development have not been considered as method facets, however. In the reported study, development team and learner-performance samples are considered as method facets. Effects of these two facets are explored.

1. Effects of scale development team

Two teams were chosen pseudo randomly from a pool of graduate students and teachers of English as a second language. They used one set of twelve authentic ESL secondary-level writing samples to construct an EBB rating scale. Members of each team then used their own scale to rate a second set of sixteen student writings.

Three types of analysis will be performed. The first compares the indicators for score levels in the two scales. To what extent do the two teams use the same performance indicators at the same positions of the writing ability scale? The second type of analysis applies to ratings. We will compare correlations of raters within scales and across scales. To what extent do raters using the same scale agree with one another more than they agree with raters using a different scale? Differences in scale structure will also be investigated. The third type investigates the reactions of scale makers and raters.

2. Effects of scale development sample

It is not possible in a limited investigation to examine the effect of scale development sample independent of scale development team. A team developing a new scale with a new sample cannot be expected to forget or ignore the indicators they found salient in constructing their previous scale.

A third team was selected from the same graduate student/teacher pool. They constructed an EBB scale from a set of writing samples equivalent to the set used by the first teams. They then used this scale to score the same set of sixteen writings.

Two types of analysis will be employed. In the first, performance indicators identified using the different sample will be compared with the two sets of indicators identified by the first teams. Are there common elements in the stage 1 scales that are not present when the different set of performances was used for scale development? The second type of analysis considers scores. We compare correlations within and across scales. Structure of the three scales will be compared.

This study expands the concept of method facets (Bachman, 1990) to include instrument development variables. It examines the effects of scale development team and scale development sample as facets of method.
Marijke Cascallar

Development and Validation of the Modified Oral Proficiency Interview (MOPI): Impact on adult language training and job assignments

This presentation will provide a detailed description of the development of this new form of oral proficiency assessment. It will also discuss the impact it has in large programs of language proficiency assessment for adults, and its possible applications in job-related language training programs for adults, as well as in the determination of language related job qualification and assignments. Several government and private organizations need to administer a large number of language test batteries to evaluate the language skills of prospective employees and/or contractors. Such individuals must meet specific requirements related to their prospective job assignments and possible organizational needs.

One of the components of the language test batteries is the Oral Proficiency Interview in various languages. Often though, the language requirements involve languages for which there are no available trained oral proficiency testers in the organization. The Modified Oral Proficiency Interview (MOPI) addresses the need to ensure that applicants and personnel in these languages can be assessed in their oral proficiency skills when no trained tester is available on a timely basis.

The MOPI is conducted by phone via a "surrogate" tester who is an educated native speaker of the target language, guided by an experienced tester in any language other than the target language. This poster presentation will describe the MOPI process, and the roles of the surrogate and experienced testers, including the following components: (a) Surrogate Tester Information Form; (b) Preparatory Training Materials; (c) Briefing on the MOPI Process; (d) Surrogate Tester Debriefing Process; (e) Role of the experienced tester; (f) Rating Protocol. The presentation will also describe the MOPI content and the three phases of the test. Examples will be provided and explained.

In addition, the poster presentation will include the data obtained from several pilot implementations of the program in a large scale setting. An analysis of the validity of the MOPI, compared to other measurement procedures, as well as the reliability of the ratings obtained will be presented. Initial assessments indicate adequate levels of interrater reliability. Data presented will include use of the MOPI for the assessment of speakers of various languages. Results will be discussed also in terms of future integration of new technologies, such as speech recognition, for the assessment of specific elements of the productions, also possible in telephone interviews. Implications for large adult education and assessment programs will be drawn.

April Ginther and Leslie Grant

The Influence of Proficiency, Language Background, and Topic on the Production of Grammatical Form and Error on the Test of Written English

Investigations of ESL writing tend to be based on limited samples, on responses to a single topic, and often focus on error rather than on students' error-free production. The purpose of this project is to examine the effects of language background, level of proficiency, and topic on both the production of grammatical form and of error on the Test of Written English (TWE). The part of speech of each word in a sample of essays written by examinees of Arabic, Chinese, Spanish, and English-speaking backgrounds is being identified and categorized. Categorizations are based on 40 selected grammatical descriptors that allow inspection of examinee use of coordination, subordination, adjectival and adverbial modification, gerunds, infinitives, and other verb forms. Errors are being
categorized with respect to word form, word choice, word order, omissions, insertions, reference, punctuation, and spelling. The examination of the interactions among the effects of level, language background, and topic on the production of grammatical form by examinees on the TWE will lead to greater understanding of the variables that influence examinee performance as well as the evaluation of writing assessments that are undertaken under similar conditions.

Gene B. Halleck and Carol Lynn Moder

Performance Testing of International Teaching Assistants:
The Role of the Assessor

Given the current emphasis on authentic language testing, language testers are increasingly making use of performance tests, which typically involve both language knowledge and strategy use. Bachman (1991) emphasizes the need to consider not only situational authenticity, but also interactional authenticity, that is, the degree to which different types of language knowledge and strategy use are involved in the test task. In performance tests the extent to which language knowledge and strategy use interact in overall test scores is determined not only by test task but also by the assumptions that raters make. These assumptions are "an integral component" of the final score (Chalhoub-Deville 1995). Furthermore, the validity of performance tests is directly related to their ability to predict the performance of test takers in non-test situations in which they will typically be assessed by naive judges (Barnwell 1989). Thus, evaluations of the assessment behavior of raters from various backgrounds is necessary in evaluating the validity and effectiveness of performance tests (Brown 1995).

This paper examines the role that different groups of raters play in evaluating International Teaching Assistants on a performance test. This test requires the testee to teach a mini-lesson and respond to questions about that lesson. Raters evaluate Language Skills (pronunciation, grammar, fluency and comprehensibility), and Teaching Skills (use of visuals, teacher presence, method of handling questions, etc.) on a zero to three scale during the initial 5-minute presentation and Interactive Language Skills during the question period following the mini-lesson. In addition, they must give a holistic rating for their Overall Impression. Since the holistic score for Overall Impression can make the difference in whether or not the candidate passes or fails the test, the raters' understanding of their role in the testing process is critical. We analyze the relative importance that three different groups of raters place on language skills and teaching skills as they determine their overall impression of proficiency. The three groups of raters were: 1) trained faculty raters, 2) naive graduate students in an oral proficiency seminar and a language testing class; and 3) naive undergraduate students. Two of these groups, the trained faculty raters and the undergraduates normally participate in the evaluation process.

Questions to be addressed are as follows: 1) Do different groups of raters rank the ITAs differently? 2) Do different groups of raters view the relative contribution of language and teaching skills differently? and 3) Is there a correlation between raters' overall impression and their determinations of language and teaching skills?

Results indicate that the three groups have different views as to the nature of the test task and their role as assessors. Although the overall ranking of ITAs showed a moderate correlation among groups, each group of raters weighted language and teaching skills differently, leading to highly discrepant evaluations of some test takers. Implications for rater training and test validity are discussed.
Liz Hamp-Lyons

**Ethical test preparation practice: the case of the TOEFL**

TOEFL test preparation books are big business. Claims are made in the advertisements for TOEFL preparation materials that they "teach the skills needed for success on the TOEFL." Claims and criticisms are also directed at these textbooks, that they do not teach the whole language, or the "real" language; it has also been pointed out that very little research has been done into the impact of the TOEFL or of materials that prepare students for it (Alderson & Wall 1992; Alderson & Hamp-Lyons, 1993).

To facilitate questions of ethical test preparation practices, Mehrens and Kaminsky (1989) developed a descriptive list of seven test preparation practices that in their view defines a continuum from ethical to unethical practices: they used these criteria to evaluate commercially available test preparation materials. This paper applies the Mehrens and Kaminsky categories to a review of TOEFL preparation materials to illuminate questions about whether such textbooks follow accepted ethicality standards.

Additionally, Mehrens and Kaminsky (1989), having reviewed studies of the effects of test preparation courses, conclude that many or most of these may be fruitless. There exists little or no documented evidence of the efficacy of TOEFL test preparation courses and materials, a fact which raises questions about the representations made by publishers and language schools in selling TOEFL test preparation courses or textbooks.

The paper considers the questions raised by Wall and Alderson (1992) about the impact of a test versus the impact of test materials and the impact of teachers, seeking to increase our understanding of the dynamic surrounding large-scale testing. It closes with a discussion of whether language testers and the language testers' professional organization, ILTA, have any responsibility to set and monitor standards for test preparation materials.

----------------------------------------------------------------------------------

Joan Jamieson and Carol Taylor

**AIMING AT ACCESS: DEVELOPMENT OF A TUTORIAL FOR A COMPUTER-BASED TOEFL TEST**

Developing computerized language tests leads to questions regarding test fairness and validity if one considers that students with little or no prior computer experience will be taking the same test as students who are "computer literate." The members of various TOEFL constituencies have reiterated the concern for TOEFL examinees who have no or only limited access to computers and thus may be disadvantaged by a TOEFL computer-based test (CBT). Moreover, although computer literacy is increasingly expected of incoming university students, there is a potential danger of confounding the measurement of English language proficiency with computer literacy. How can test makers and users be confident that students' scores on the computerized test reflect their language ability rather than some combination of language and computer skill? One way TOEFL 2000 project staff attempted to address these concerns was by developing a computerized tutorial that would precede the administration of a TOEFL CBT. This tutorial has been informed by test taker performance and feedback on prototype tutorials and computer-based test tasks. Its purpose is to familiarize examinees with the range of computer-based TOEFL test tasks. A study is being conducted that will examine the extent to which the use of preparatory materials including the tutorial eliminates any performance differences among groups of TOEFL examinees having varying degrees of computer familiarity. Part of this research effort will involve looking at examinee performance across a range of new TOEFL item types. In addition, we expect that this research effort will provide baseline data regarding the computer familiarity of TOEFL examinees.

This poster/demonstration session will present the results of the prototype studies that lead to the major design features of the current tutorial. We also expect to provide a hands on demonstration of the newly developed tutorial.

----------------------------------------------------------------------------------
Dorry Kenyon

Further Research on the Efficacy of Rater Self-Training

Theoretical Background and Rationale

Can individuals train themselves to score a performance assessment acceptably when compared to the ratings of experts? Stansfield and Kenyon (1993) showed that it was possible for motivated individuals to learn to score the tape-mediated Texas Oral Proficiency Test using the ACTFL Proficiency Guidelines at an acceptable standard without going through the live training procedures.

Purposes of the Research

This poster presents results of further research on the efficacy of rater self-training using data collected from the trailing of self-instructional rater training kits developed for Simulated Oral Proficiency Interviews (SOPIs) in five languages: Spanish, French, German, Chinese and Japanese. For each language, the purpose of the research was to find out whether individuals could train themselves to score the SOPI to the same level as expert raters.

Research Design and Methods

For each language, we developed a final calibration test containing excerpts from performances of 15 examinees. In each language, two to four expert raters independently rated the performances on the final calibration test. The experts had wide experience with the SOPIs and the ACTFL Guidelines used to score them. Each evaluator worked through the trailing version of the kit and then rated the performances on the final calibration test. The number of evaluators ranged from five in Chinese to twelve in German. Most had little prior experience with the ACTFL Guidelines.

We created the key of final ratings for each performance using the ratings of the experts. Informed from data on the operational scoring of ACTFL oral proficiency interview (Thompson, 1995), an acceptable level of performance for the evaluators was set at 60% exact matches with the key and no rating more than one step away on the ACTFL scale. A generalizability approach was then used to analyze the ratings. Phi-coefficients were calculated for the group of experts, the evaluators who passed the criterion, the evaluators who did not pass, and all evaluators. Correlations with the key were also computed for each individual, whether expert or evaluator.

Results

Across Spanish, French, German and Chinese, approximately half the evaluators met the passing criterion. The phi-coefficient for a single rating using a member of the group of experts ranged from .86 in French to .94 in Chinese. In Spanish, French and German, the corresponding phi-coefficient using a member of the group of passing evaluators was similar to that of the experts. The phi-coefficient for a single rating using a member of the non-passing group were unacceptably low. Correlations with the key for experts and passing evaluators ranged from .84 to .99. Data from the Japanese group are yet to be analyzed, though similar results are expected.

Implications

The findings from this research further support the claim that it is possible for individuals to train themselves to score SOPIs using the ACTFL scale.
Linking Performance on a Multiple-Choice Reading Test to Verbally-Defined Proficiency Levels

Theoretical Background and Rationale

How can we convert an examinee's performance on a multiple-choice foreign language reading test into a level on a verbally-defined reading proficiency scale? In particular, how can we theoretically and empirically relate such performance to the levels of the ACTFL Proficiency Guidelines, whose influence has been pervasive in American foreign language education? The Educational Testing Service (ETS) has tried several approaches using tests in Russian and the Test of English as a Foreign Language (TOEFL). However, these strictly empirical approaches failed to use theoretical models or item-level information.

Purposes of the Research

In this poster, a four-step method using an investigation of the construct representation of the test items to make this link is proposed. The poster also reports on an application of the methodology. The application attempted to make a direct link between item-level characteristics, the ACTFL proficiency descriptors, and overall test performance by defining the underlying test construct using the ACTFL proficiency descriptors.

Research Design and Methods

The methodology consisted of four steps. First, the researcher developed a cognitive processing model and scales to operationalize its variables. Seven variables were operationalized using the levels and wording of the ACTFL proficiency descriptors. One additional variable salient to the multiple-choice format of the test was also operationalized. The second step involved collecting scores on each scale for 100 reading comprehension items of a Chinese test. The researcher trained three judges to perform this task. Each judge independently assigned scores for each scale for each item. In the third step, the researcher used an adaption of the Linear Logistic Test Model (Fischer, 1973) to examine the construct representation of the items. Data came from the performance of 2348 examinees. In the final step, the researcher determined the boundary scores for the proficiency levels of the ACTFL Guidelines in terms of test performance by relating raw scores on the operationalized variables to the Rasch item difficulty scores from the LLLTM analysis.

Results

The methodology was successfully applied to this data set. Judges applied the scales with a high degree of consistency. The variables defined according to the ACTFL Guidelines accounted for 40% of the variance in item difficulty. The method could be used to link performance on the test to ACTFL proficiency levels.

Implications

Beyond demonstrating a methodology, this poster sheds some light on the strengths and weaknesses of the verbal descriptors of the ACTFL Proficiency Guidelines for reading.

Pavlos Pavlou

Do Different Speech Interactions in an Oral Proficiency Test Yield Different Kinds of Language?

Recent studies in oral proficiency testing (Bachman and Palmer, 1983; Lantolf and Frawley, 1985; Savignon, 1986; Van Lier, 1989, etc.) have indicated problems with the oral interview, the most common technique in testing oral proficiency. Research has shown that this technique may not be the only or best way to test oral proficiency since
it is insufficient to detect many aspects of the construct.

Certain theoretical models for testing oral proficiency that have been proposed to deal with the above mentioned problem indicate the necessity and usefulness of assessing oral proficiency through various genres (Shohamy et al. 1986, Shohamy 1988). Group discussions, oral reports, oral interviews, and role plays are some of the proposed speech interactions. Shohamy et al., 1986 support that these different speech interactions are necessary in a test of oral proficiency because the test taker produces different kinds of language in each interaction. In addition, through these interactions it can be ascertained whether the test taker is sensitive to contextual variables (such as role, relationship, setting, formality level, topic, function, etc.) that interact in communicative oral language.

A test battery for oral proficiency was developed according to guidelines set by Shohamy et al. (1986) and Shohamy (1988) in order to provide evidence for or against the hypothesis that different speech interactions measure different aspects of oral proficiency. The test was administered with 60 EFL high school students in Cyprus.

In order to assess whether there is a difference in the language produced in each interaction all four interactions were defined independently of possible content or specific topics addressed in a real test situation. Then, after a structural juxtaposition of the four speech interactions, a list of possible features thought to be unique or relatively more frequent in each speech interaction was compiled. Finally, the actual frequency of occurrence of the selected features in the samples produced by the 60 candidates in all four interactions was counted.

The results show that certain features actually occur more frequently in a given speech interaction, a fact that justifies the use of a variety of speech interactions in an oral proficiency test. The study also provides suggestions for the development of similar test batteries which include as many of the selected features as possible.

********************************************************************************

Mirja Tarpanen, Tommi Salmela, Anu Halvari, and Ari Huhta

The Comparison of Different Levels and Languages in a Criterion Referenced Test

The Finnish National Certificate (NC) has been administered four times. The certificates include 7 different languages (English, Finnish as second language, French, German, Russian, Spanish and Swedish) at three levels (basic, intermediate, advanced). The examination is criterion referenced, measuring proficiency levels 1 - 8 (roughly comparable to the ESU levels 2 - 8), and it includes subtests in reading and listening comprehension, vocabulary and structures, writing and speaking. The prospective test-taker can be anybody who has studied or otherwise acquired one of the forementioned languages, thus, the target group is very heterogeneous. The proficiency levels are defined verbally, and the descriptions are the same for all languages, which means that a particular grade, e.g. 3 in German, should mean the same as grade 3 in Swedish or in English in terms of the ability to use the foreign language. The comparability of the grades across languages in a problem typical to multilingual testing systems, and cannot be neglected by the test developers working for the NC, not only because of the criterion referenced nature of the tests. A failure to design at least roughly equivalent tests for the different languages would probably make the test takers question the validity of the whole examination system, as quite a few of them attend the examination in more than one language and have a fairly realistic idea of how well they master the foreign languages that they have studied.

Another major concern for the NC system has been the comparability of grades awarded from the tests at different levels. Would a test taker who is given grade 3 in the basic level test of French be given grade 3 also in the intermediate level test of the same language?

This paper sheds light on how these demands for comparability are met in the National Certificates both in theory and in practice.

The attempt to ensure comparability is an ongoing activity and includes four phases:

Planning phase
- the same test and item specifications for all languages, updated after every testing round
- item writers' meetings: representatives from every language discuss the basic principles and try to learn from past experiences

Item writing phase
- tasks / tests are commented by representatives of different languages and by system coordinators

Assessment phase
- benchmark examples (also multilingual) are used, especially in writing and speaking
- general assessment criteria and guidebooks are the same across languages / levels

Test-taking phase
- comments from individual test-takers who have participated tests in different languages and at different levels (referred to as multi-test-takers, MTT's)

The poster will present quantitative information including test results from all candidates (N = about 4000), and the MTT's (N = a few dozen) responses to questions concerning their views about the comparability of the tests in different languages and at different levels. Also, qualitative analyses will be presented which discuss how well the procedures have succeeded, and what needs to be developed.

Carol Taylor and Irwin Kirsch

TOEFL 2000 PROJECT UPDATE, 1996

Within the past year, TOEFL committees, consultants, and constituencies have recommended a more incremental approach to TOEFL 2000 development efforts that would result in offering further test improvements before the year 2000. In responding to this recommendation, TOEFL 2000 project staff identified and analyzed possible design features for a computer-based TOEFL while keeping in mind the longer-range goals of TOEFL 2000. This interim test will include unique CBT-design features that hold promise for providing a vehicle for further incremental improvements and gaining experience with CBT that would be informative to the longer-range goals of the TOEFL 2000 project.

To achieve these goals, TOEFL 2000 research and development efforts continue a multifaceted approach to project development. Current efforts include, but are not limited to, (1) the design of new CBT TOEFL items and response types, (2) an investigation of examinee performance on a computer-based tutorial and on computer-based test tasks, (3) a study to investigate the effects of multidimensionality on the selection of items and the estimation of abilities for computer adaptive portions of the test, (4) the development of a conceptual test framework that takes in account models of communicative competence, and (5) a narrowly focused research agenda that is directly applicable to inform the transition from CBT to integrated skills CBT (i.e., TOEFL 2000).

This poster/demonstration will outline the direction from TOEFL CBT to TOEFL 2000. Abstracts of currently available TOEFL 2000 monographs will also be provided.

Ingrid Wijgh

On how efficient a test of oral interaction can be

A research project has been carried out to answer the following three questions:
- How must oral interaction ability in a foreign language be defined and how can this ability be operationalized in terms of language behavior?
LTRC 96 POSTER ABSTRACTS

- Is it possible to develop an instrument which elicits the language behavior in question?
- Does this instrument meet with criteria of validity, reliability and efficiency.

To answer the first question a framework has been developed consisting of the different components of oral interaction ability and the distinctive features of its operationalization into language behavior. This framework has been validated through expert judgement.

A prototype of a test of oral interaction has been developed which claims to elicit the language behavior as described in the framework. This has been verified through analysing students' performances on the test. The main part of the study consisted of research into the reliability and efficiency of the test. To establish the reliability of the test an intraclass coefficient for interrater agreement has been estimated. Secondly a G-study has been carried out to examine to what extent the results could be generalized and if the rating procedure or the test length could be reduced. Data were gathered by having 8 raters score 30 students' performances on the test. The test consisted of six parts and of 50 items and took about 20 minutes. The raters followed two rating procedures, a rather analytic and a more global one:
- they scored each item separately on a fourpoint-scale and
- they scored two aspects for each part of the test, of which one was specific for that part of the test and the other was fluency.

The analysis of the data showed that interrater agreement was slightly higher in the analytic rating procedure than in the more global one (ICC .83 versus .86). The results of the G-study showed that the prototype was an excellent measure for oral interaction (g-coefficient .83) when only the global rating procedure per part was followed. If only one half of the test is taken and rated, the g-coefficient is .76 for the first half and .83 for the second half. If only the aspect 'fluency' is rated in each part of the test, the g-coefficient is .80.

In a small-scale replica of the study with three new raters the agreement between the first group of raters and the second group turned out to be almost perfect (ICC .95). In view of these results the conclusion is drawn that in this research project a valid, reliable and efficient instrument has been developed to measure oral interaction ability.

******************************************************************************************

Elaine Wylie

Test and non-test assessment for the professional second language user

The context of this poster is a push to improve the quality of the teaching of languages other than English (LOTEs) in Australia. The Federal Department of Employment, Education and Training has funded projects aimed at measuring the proficiency of teachers who have learned as a second language the language they are teaching, and at providing a curriculum framework for programs to increase the proficiency levels where necessary.

The major focus of the poster will be on an assessment project involving the development of a specified purpose scale and tests for teachers of Indonesian. This project culminated in late 1995 and early 1996 in a 'benchmarking' process involving the following:
- the rating of 50 teachers (in various states and territories of Australia) against the scale by fully direct, adaptive testing procedures
- the administration of 'standard' (i.e. non-adaptive) tests to the same cohort
- the alignment of the ratings on the scale and the scores on the tests.
A sub-set of the teachers also self-assessed their proficiency, and these were compared with formal ratings.

The major theme of the poster - the coherence and practicality inherent in the use of proficiency scales which are independent of any particular assessment procedures - will be developed in an overview of types of proficiency assessment which can be used at different stages of a teacher's career:
- WHEN PLANNING VOCATION, self-assessment against (a simplified version of) the general proficiency model of the scale
- WHEN APPLYING FOR ENTRY TO DIPLOMA OF EDUCATION, adaptive test to rate against the general proficiency version of the scale
- WHEN APPLYING FOR EMPLOYMENT, adaptive or standard test to rate against the specified purpose version of the scale.
- AT/TOWARDS THE END OF PROBATIONARY PERIOD, test or comprehensive observation of performance in situ in professional contexts to rate against the specified purpose version of the scale.

The focal project has been a collaboration between the two testing centres of the National Languages and Literacy Institute of Australia (NLLIA). Reference will be made (partly through handouts) to the research and development contexts of the two centres:
- the evolution of a specific purpose model of the Australian Second Language Proficiency Ratings, a scale which has been used in Australia for 17 years, and the development of the particular specified purpose version involving 'core' registers such as the metalanguage of second language teaching (the NLLIA Language Testing and Curriculum Centre);
- the prior development of tests for teachers of Japanese and Italian, and the extensive needs analyses for those projects which have informed the present project (the NLLIA Language Testing Research Centre).

Reference will also be made to the use of the specified purpose scale as a proficiency framework in national projects to develop curricula for LOTE teacher education.
SHARPENING OUR PRAGMATIC INSTRUMENTS: ADDITIVE ASSESSMENT IN ESL

Doug Absalom and Alejandro Brice

Predictive capabilities or standard tests in University environments have often left much to be desired and it is hypothesized that this is at least partially due to test-focused procedures by candidates, rather than focus on the language in operation. That is, students tend to see the test as an end in itself, a gateway that bears no relationship to the pathway beyond. They thus adopt strategies such as practice tests and extensive private study that help them to pass tests but which do not fit them for the real world tasks involved in their ensuing courses. Survey evidence suggests that skills developed prior to entrance tests are often not carried beyond those tests.

This paper investigates the addition of a further dimension to language testing procedures in an attempt to improve the predictive nature of entrance assessment. Pragmatic considerations such as affecting and responding to an interlocutor, expressing one’s Self, initiating and controlling dialogue, cueing topic shifts and listening actively were assessed using a Pragmatic Screening System consisting of 38 separate items of behaviour observation, and the resultant profiles correlated with some of the regular entrance tests used at Universities. It is hypothesized that students with high pragmatic-language (as measured by standard tests and by the pragmatics screening scale respectively) correlations are those that have acquired language in active environments and that their language skills are more likely to be transferable to University environments than those students with low pragmatic-language correlations. It is assumed that the latter group will have adopted test focussed strategies that tend to preclude the pragmatic aspects observed in the screening system. The investigation is a longitudinal one and is ongoing.

As might be expected, such aspects as rater reliability, presentation techniques in University courses, learner strategies in University courses and strategy emphasis in pre-university ESL courses interfere with results of such an investigation. Variations in cultural pragmatic expectations, not only between Western and Eastern cultures but also between American and Australian cultures, have also provided problems that have caused the Screening Scale to undergo some modification during the process of the investigation. These and tentative results of the program, along with some recommendations for University entrance assessment, are provided in the paper.

Micheline Chalhoub-Deville

Proficiency-Based and Diagnostic Computer Adaptive Testing: Development Issues

The proposed paper focuses on the computer adaptive reading proficiency tests under development at the Center for Advanced Research on Language Acquisition (CARLA). This paper will address the theoretical underpinnings and the research undertaken to construct computer adaptive tests (CATs) for assessing and providing diagnostic information regarding students’ reading proficiency in French, German, and Spanish. The paper will report on an extensive review of various reading and testing models (e.g., ACTFL, 1986; Bachman, 1990; Bernhardt, 1989), and underscore the aspects most useful to construct the CAT content matrix.

The paper will describe in detail the blueprint for the French, German, and Spanish proficiency and diagnostic item banks. Generally, the proficiency bank will include the more integrative items with an emphasis on authentic materials, similar to those that students encounter in real-life situations. Another purpose of the CAT project is to design and develop a bank of items that would help provide students with detailed information regarding the areas/components in which they need further instruction to attain higher proficiency levels. The diagnostic item bank will typically include the more discrete items with the objective of pinpointing a student's area(s) of difficulties. Furthermore, the proposed paper will discuss the descriptive systems, a comprehensive set of content attributes and
constraints, which focus on how the proficiency item banks of the French, German, and Spanish tests will link to their respective diagnostic item banks.

Finally, this paper will report on various theoretical issues and empirical findings in the areas of L2 reading and assessment, technology, item types, scoring algorithms, item selection heuristics, test security and exposure. This discussion will be informed by the arguments that had taken place at the "Issues in Computer Adaptive Testing and Reading Proficiency" conference, (March 20-22) hosted by CARLA to inform the development of the present CATs.

CAT is proclaimed as a "technological solution to many serious and costly problems which plague conventional, standardized testing" (Luecht, 1995). Testing researchers are increasingly exploring the feasibility of CAT for assessing language proficiency. The knowledge, experience, and research presented in this paper will help further the field of computer adaptive language testing research and lay the groundwork for future research in this area.

Carol Chapelle and John Read

Toward a framework for vocabulary assessment

As the profession works toward the design and use of alternative forms of assessments, particularly those which make a positive impact on learning, it must also investigate their validity in an appropriate manner. Validation theory (Messick, 1989) offers a valuable framework for exploring multiple forms of validation, and therefore offers potential for evaluating alternative assessments, however, among the problems in applying validation theory to testing practices is the overwhelming number of potential methods of validity inquiry for any given test. Rational application of validation theory requires the researcher or test user to select and prioritize the types of potential arguments for test interpretation and use on the basis of the test's intended purpose in a given context. As a consequence, progress in the use of validity arguments rests on an understanding of test purpose.

Focussing on L2 vocabulary assessment, this work-in-progress session outlines a research agenda intended to expand our understanding of practical validation issues by exploring the interrelationships among test purpose, construct definition, test design, and validity inquiry. L2 vocabulary assessment is ideal for exploring these issues because the lexicon is recognized to be a significant focus of L2 teaching and research, and alternative forms of assessments are needed for these purposes. The first step in this research agenda -- and the focus of the presentation -- is the definition of test purposes for vocabulary assessment. We wish to discuss and receive feedback on our method of characterizing test purpose, which is comprised of three facets (inferences, uses, and impacts), and on the specific purposes we have identified for vocabulary tests in educational and research contexts. The input received at LTRC '96 will inform our conception of test purpose and therefore subsequent stages of the research.

Patricia Dunkel

Computer-Adaptive Testing of Second/Foreign Language Listening Comprehension: Issues to Consider When Constructing and Calibrating Multi-Media Item Banks

Computer-adaptive testing (CAT) is a procedure in which specific test items presented by the computer vary with the estimated ability of the examinee and her or his response to previous items. The potential benefits of CAT (e.g., its efficiency) are beginning to be recognized in the field of education, in general, and in second/foreign language education, in particular. To illustrate, in the fall of 1993, the Educational Testing Service (ETS) introduced a computer-adaptive version of the Graduate Record Examination. In April, 1994, all candidates for a nursing license in the United States began taking a computer-adaptive test on the content information. With respect to second/foreign language education, several L2 CATs have been developed in the past (e.g., the Computerized Reading Comprehension Proficiency Test sponsored by the American Council on the Teaching of Foreign
Languages, the COMPUTEST: ESL reading structure test, and the S-CAPE: Spanish Computerized Adaptive Placement Exam given at Brigham Young University, to name a few, and more CATs are under development at present. To date, L2 CAT research and development has focused mainly on creating reading- and grammar-skills assessment instruments. Before the advent of large-storage and fast personal computers such as the PowerMacintosh 8100/100 Audio Visual machine, the addition of a speech component to a CAT for listening assessment purposes added a dimension of complexity to the development effort since the storage and retrieval of digitized speech were extremely memory intense. As a result, in the 1980s, few efforts to develop listening comprehension CATs for administration on personal computers were undertaken, or even possible. However, as a result of the increased storage-capacity and sophistication of today's personal computer systems, test developers are not only able to create listening comprehension CATs, but they can also create multi-media listening CATs. Problems facing today's multi-media CAT developers are no longer those involving the storage and retrieval of digitized speech, photographs, and full-motion video, but instead they involve issues related to the following: (1) selection of an appropriate framework and model for development of a multi-media item bank; (2) identification and use of item formats appropriate for multi-media CAT design; (3) identification of suitable and efficient item-trialing procedures to establish calibration of the item in the bank; and (4) utilization of suitable item-selection and test-stopping procedures, to name but a very few. In other words, although the hardware has gotten more sophisticated and easier to use, and the technical difficulties have lessened, multi-media CAT developers face all the hurdles and challenges paper-and-pencil developers face, and some novel ones, in addition. We are very much in the infancy of L2 CAT development, in general, and L2 multi-media CAT development, in particular. Those involved in CAT development (or those contemplating getting involved therein) need to examine the strengths and weaknesses of those operational CATs that have been developed, to date. It is the purpose of this presentation to discuss issues involved in the development of three operational multi-media listening CATs in English as a Second Language (ESL), Hausa, and Russian. The listening CATs will be available for examination on a PowerMac notebook computer, and information on the development effort will be provided.

In summary, the presenter will do the following: (1) outline the structure of the framework used to develop the item banks; (2) provide data related to the trialing and calibration of the tentative item banks; (3) examine issues related to constructing multi-media listening CATs; and (4) encourage participants to discuss issues related to development of multi-media CATs.

Angela Hasselgren

The development and trialling of criteria-referenced tests of spoken English in Norwegian secondary schools

This paper will give an account of the a priori validation and the outcomes of the initial a posteriori validation of a diagnostic test of spoken English, to be used in Norwegian secondary schools. An outline will be presented of a projected large-scale a posteriori validation, the main aim of which will be to shed light on how fluency can be operationalised and evaluated.

It was decided initially to make the following demands of the test:
* it was to adopt a combined interactive-ability/real-life approach to authenticity,
* it was to be conducted with pupils in pairs, and taped for assessment;
* it was to be evaluated according to criteria scales, with the use of 'performance profiles' giving an in-depth description of each pupil's performance on critical aspects of speaking. These profiles were intended as both a guide to assessors and a feedback to pupils.

Trialling was carried out on 70 pupils. Two external raters assessed each pupil.

The paper will emphasise the following aspects of a priori validation:
* test items - how they relate to a model of CLA and its operationalised traits;
* methods and procedures - their contribution to 'authenticity' and the safeguarding against error of measurement;
* scoring procedures - the building of performance profiles and criteria scales on theory and empirical research, to lead to a 'correct' interpretation of scores and reduced rater error.
The following findings from an initial a posteriori validation will be presented:
* inter-rater reliability, across overall scores and individual performance items
* correlations between test scores and teachers as well as pupils' assessment

An outline will finally be sketched of the projected a posteriori validation, involving the following stages:
* reaching working definition of 'fluency' in terms of (sub)traits/operations;
* analysis of data (taskwise) to see how performance on these subtraits correlate with each other and with overall speaking ability;
* analysis of rater's performance profile to see how these subtraits are represented in questions
* correlation between performance on subtraits as exhibited in data and as evaluated by raters;
* evaluation of criteria scale relating to fluency with respect to how well its levels describe what occurs in the data;
* conclusions as to how fluency can be defined in terms of traits and operations, tested in terms of tasks, assessed in terms of questions on a performance profile and evaluated by using a criteria scale.

Ari Huhta

**Variability vs. Generalizability of Oral Proficiency across Different Test Tasks**

The presentation reports on a study which examines whether certain features of speaking ability – and the assessment criteria which reflect them – are more affected by the context of language use than others. The study relates to a fundamental problem in language testing: how generalizable are the grades awarded on the basis of performance on one task to other tasks and contexts?

On the basis of theoretical analyses of the most commonly used assessment criteria, it seems likely that some criteria (features of language) are more generalizable across different tasks than others. That is, a person's performance level on a certain feature of language can remain more or less the same irrespective of the situation where he/she has to speak. An example of such a feature might be the ability to pronounce individual sounds. Thus, for getting a fairly generalizable idea of a person's ability on this feature, it might not matter very much what kind of tasks are used to elicit speech in a test. Almost any task might give the tester a generalizable idea about the person's ability. In contrast, a person's performance on some other areas of speaking ability may vary greatly depending on the situation or task. One task might give a tester a very different picture of the test taker's skill from another task that the tester could have chosen. Examples of these less generalizable features are probably the adequacy of vocabulary or the ability to use appropriate language (sociolinguistically).

Thus, the generalizability of, e.g., an overall grade for speaking probably depends on the particular criteria used: if the criteria consist of predominantly 'generalizable' features of language, the grade will be more generalizable than when all or most of the criteria are very sensitive to the context of language use. Various studies have indirectly touched this problem, but it appears that attempts to systematically study the question are practically nonexistent.

The presenter decided to examine the problem by choosing a number of assessment criteria which probably reflect the hypothesised range in generalizability, and by designing a set of different tasks for a SOPI type of speaking test. The tasks vary along some contextual factors which in sociolinguistic research have been shown to affect the way people use language: monologue vs. dialogue, and different degrees of familiarity with the addressee(s). A number of trained assessors will rate each task separately and use all the criteria. The ratings will be analysed statistically to find out if some of the testees' features of speaking vary across the test tasks more than others. The analyses will be complemented by information obtained from the assessors and testees via questionnaires and interviews.

The study is now at the pilot testing phase where a number of assessors will be asked to provide separate ratings of six different SOPI tasks in a large-scale administration of a proficiency test of English. Also, a computer programme has been designed which will allow the simulation of the data to test how much variation (and subjects) there should be so that a statistical programme such as the FACETS would detect the phenomenon studied. This simulation is felt to be necessary as finding enough subjects and raters is difficult in studies addressing oral skills, and because language testers seem to be uncertain as to how many subjects are needed to use statistical programmes based on Rasch models in a reliable way.
The presenter wishes discuss the pilot and main studies, and the major problems he is faced with such as the amount of data needed for the quantitative part of the study, and the type and reliability of ratings required.

Valmar Kokkota, Hele Pärn and Leeni Simm

Development of the Test of Estonian Language for Naturalization Purposes

This work in progress paper describes the standardization process of the test of Estonian language for naturalization purposes. The test official name can be translated into English as Standardized Test of Estonian Language for Citizenship Applicants (STELCA).

The STELCA development began in mid-April, 1995 and its first official use took place on 21 November, 1995. The background of the STELCA development will be described including criticisms of the previous loose examination system, the political and socio-cultural situation in Estonia during the work on the STELCA etc.

Comparisons will be drawn with the test of Latvian language for naturalization purposes compiled in cooperation with the Council of Europe. Due to the lack of previous agreements, parliament elections followed by the crisis of the new government and fierce political struggle against the modernization of the previous examination system Estonia failed to get any outside expert advice.

The format of the STELCA includes four sections: writing (two short tasks), listening (3 texts, 12 items with fill-in answers), reading (4 texts, 16 items with MC answers) and an interview (up to 15 min.). The specific problems of text selection related to the authenticity requirements, sampling problems for the try-outs, data-processing and interpretation problems will be covered in this paper.

The principles of assessment and rating used for STELCA will be explained with references to the Administration Guide and the Assessment Guide developed during the try-out stage of the work on the STELCA. The Test Preparation Booklet has been compiled for the testees and teachers using the texts and items which has passed the try-outs but were not included into the final forms of the STELCA.

The validity and reliability studies, usual item analysis, the test and its sections' overall suitability studies were carried out during the stages of the STELCA pretesting and official administration. According to the current regulations the STELCA both forms have to be renewed each month. All the data mentioned above will be presented in the paper with the accompanying hand-outs and transparencies.

Jose Lai

Self-Assessment as a Meta-cognitive Strategy in Managing Self-Directed Language Learning

Research in self-assessment has focused on the areas of developing relevant techniques and finding out the extent to which it correlated with other formal objective forms of assessment (Oscarsson 1978). Relatively few studies have investigated the role self-assessment plays in the learner's learning process. With the changed educational paradigm in the 80s (Breen 1987), attention has shifted from teaching to the learning process and how learners can be empowered to take an active role in it. Incorporating self-assessment as an integral part of basic curriculum thus becomes logical and necessary (Nunan 1988).

This project examines the role of self-assessment in self-directed language learning. Self-assessment is defined as the learners' ability to diagnose their own performance on a given learning task whose objective(s) the learners have set themselves. This ability is viewed as a meta-cognitive strategy which allows learners to gain tremendous insights
into their own learning progress and hence modify their learning approach appropriately. It is hypothesized that the encouragement of guided continuous self-assessment will enhance self-direction in language learning. Some sixty subjects from tertiary ESL courses are studied over a three months’ period. The main self-assessment instruments identified include self-rating scales and introspection based on guided questions.

Sari Luoma

**Analysing the language samples elicited in a tape-mediated and a face-to-face test of speaking**

When researchers try to understand what is being tested in various tests of speaking, one of the central concerns is accounting for the representativeness of the language sample elicited. This is particularly relevant where the merits or otherwise of various forms of measurement are being judged. With respect to face-to-face and tape-mediated tests of speaking, this issue has been raised by Shohamy (1994) and O’Loughlin (1995). The merit of the Shohamy study is the number of features covered, from genre and rhetorical structure to lexical density and the number of morphological errors. O’Loughlin’s merit is careful reporting on method of analysis, but at the cost of concentrating on one feature only: lexical density. The present study investigates the usefulness of various methods of analysing spoken discourse for describing the performances elicited in two tests of speaking, one tape-mediated and the other face-to-face. The methods should help comparison of the characteristics of the language elicited in the two tests.

The clearest overt difference between tape-mediated and face-to-face tests of speaking is the absence vs. presence of interactivity in the test situation. The “interaction” in a tape-mediated test is pre-programmed, “canned”, and requires the candidates to adapt to the test, while real-time adaptation of the test to the candidate is possible in face-to-face tests. How this is reflected in the language samples elicited is largely unknown, however. In searching for ways to analyse similarities and differences between tape-mediated and face-to-face tests of speaking the study is thus attempting to find means for describing interactivity and the lack of it in spoken performances.

The search for various methods of analysis was initially informed by working hypotheses and the availability of research instruments. Language functions were thought to be an interesting target of analysis because the definition of language skills in the test system where the target tests came from partially built on a functional description, and because in the initial stages of test development this was considered to be one of the advantages of the tape-mediated test. Transcripts of ten candidates’ performances in each test were tagged for language functions in Childes, as this was one of the instruments available. Grammatical complexity, lexical range, use of phrasal language and overt hesitations were considered as potentially interesting, though possibly difficult to analyse. At the time of writing the abstract, only one other program, ACAMRTT, was available to the researcher for implementing analyses, but the results are not clear yet.

The objective of the presentation is to discuss the methods of analysis used and the results of the analyses. The participants will be invited to judge the usefulness of the methods used for understanding discourse differences between tape-mediated and face-to-face tests of speaking, and for understanding the linguistic realisations of interaction. The presentation will build on actual data, both transcripts and analyses.

Charlene Rivera and Margaret E. Malone

**Assessing Native Speaker Language Development in Bilingual Education Programs: The Spanish Language Assessment**

- **Oral**

**Overview**

The Spanish Language Assessment-Oral (SLA-O) is being developed during the 1995-96 and 1996-97 school years. The SLA-O is an unobtrusive oral language proficiency instrument which will allow teachers to assess the
language proficiency and the language development of native speakers of Spanish enrolled in Spanish/English bilingual programs in Washington, DC and Fairfax County.

Objectives

The objectives for Project OASIS' field-initiated research are as follows:

(1) develop an observation-based instrument to assess native (Spanish) language oral proficiency;

(2) perform a validation study of the observation instrument, including a study of validity, practicality and washback effects of the instrument;

(3) conduct a reliability study;

(4) adapt and implement a teacher training model for the SLA-O for the District of Columbia and other districts and states;

(5) conduct a yearly evaluation of the project.

Presentation

In this presentation of the progress of the SLA-O, we will discuss the process and challenges of developing the SLA-O scale, some of the preliminary results of the validation study, and the results of the reliability study.

Mirja Tamanen, Torumi Salmela, and Sauli Takula

Testing Vocabulary: How Much Does the Method Affect the Scores and Grades?

The Finnish National Certificate examination has been administered four times. The certificates include 7 different languages (English, Finnish as second language, French, German, Russian, Spanish and Swedish) at three levels (basic, intermediate, advanced). The test is criterion referenced, measuring levels 1 - 8 (roughly comparable to ESU levels 2 - 8), including subtests in reading and listening comprehension, vocabulary and structures, writing and speaking.

Because one of the subtests include knowledge of vocabulary, it is vital to examine, what knowledge of vocabulary at a given level actually means. How many words does a candidate have to know in order to reach a certain level? What is the effect of the testing method?

The research questions will be addressed through a carefully designed experiment. The purpose is to ask a fairly large number of foreign language learners at approximately levels 1 - 5 to do vocabulary tests in some of the aforementioned languages. The plan is to create a vocabulary bank of a few hundred words, sampled from approx. 10,000-word basic / intermediate vocabulary. The samples then translated into different languages, and a number of different vocabulary tests are designed on the basis of the translated sample. Several testing techniques will be used, eg. words in a sentence context, decontextualised words, productive and receptive tasks (multiple choice, gap-filling, translation). Test booklet will be rotated in order to achieve a better coverage of the vocabulary sample.

The study aims at producing information on how much difference a specific test method makes for a test taker's vocabulary score. Some of the methods currently used in the examination probably result in higher scores than others (even if the words tested are exactly the same). It is the estimation of the magnitude of these differences which will hopefully be one of the more practical outcomes of the study. This would enable the test designers to estimate more accurately where to place appropriate cut off points for the different grades, when they know which particular combination of test methods was used in the vocabulary test.
language proficiency and the language development of native speakers of Spanish enrolled in Spanish/English bilingual programs in Washington, DC and Fairfax County.

Objectives

The objectives for Project OASIS' field-initiated research are as follows:

(1) develop an observation-based instrument to assess native (Spanish) language oral proficiency;

(2) perform a validation study of the observation instrument, including a study of validity, practicality and washback effects of the instrument;

(3) conduct a reliability study;

(4) adapt and implement a teacher training model for the SLA-O for the District of Columbia and other districts and states;

(5) conduct a yearly evaluation of the project.

Presentation

In this presentation of the progress of the SLA-O, we will discuss the process and challenges of developing the SLA-O scale; some of the preliminary results of the validation study, and the results of the reliability study.
ROUND TABLES

ROUND TABLE A:

Chair: Lawrence T. Frase
Participants: Jared Bernstein, Jill Burstein, Carol Chapelle, April Ginther, and Judith Klavans

TECHNOLOGY FOR LANGUAGE ASSESSMENT AND LEARNING

The goal of this roundtable is to help define a fuller and more systematic approach to the use of technology for language assessment and learning. Six participants, with expertise in computer technology (related to reading, writing, speaking, listening, authoring systems, and database retrieval), applied and computational linguistics (the conceptual basis of tests), and testing (for instance, algorithms for scoring), will each provide a perspective on a critical issue. After all individual presentations there will be a 20 minute open discussion centered around the question of whether we have a basis, in disciplines of language teaching and testing, that can guide the best use of technology and whether language teachers and testers can reasonably be expected to make effective use of relevant technologies.

Based on our recent review of technologies for language assessment for the Annual Review of Applied Linguistics, we think that much could be done to create a more comprehensive and integrated approach to the use of technology for language assessment and learning. The profession now comprises many views about what should be taught and tested, but linguistic performance is often assessed in narrow ways. Poor use of technology might narrow this focus even more. In addition, technology could be used to address a variety of activities in test development, delivery, and use; however, the broad range of functions needed across those contexts are poorly represented in today's technology use. To develop a fully functional test of communicative language requires a clear definition of performance as well as an integrated view of the curricular, software, hardware, and network elements that might make up such a system. Actually, the obstacles to the successful use of technology for language testing now seem less technical than conceptual. Promising practical developments in language teaching and testing tools, such as authoring systems, make it possible rapidly to construct multimedia language instruction. Important human functions in language assessment, such as collaborative decision making, exchange of test results (writing and speech samples), and scoring and summarization of complex data, can be supported by technology. One need only visit the language learning resources on the World Wide Web to see the potential. Nevertheless, dialogue about the use of technology in language teaching and testing is just beginning. Hence, the goal of this roundtable is to broaden that dialogue and to give it some focus, to become more systematic, to lay out areas where opportunities for the use of technology exist, and to identify technology partners for making the best of these new resources. This roundtable will develop an agenda of areas within which work would contribute directly to the furtherance of language testing and teaching, along with a set of action statements concerning critical issues, such as what groups should be brought together, what infrastructure must be developed, and how such work might be funded.
ROUND TABLE B:

Chair: Eduardo Casallas
Participants: Marijke Casallas (and others)

FACTORS INVOLVED IN SPEAKING PROFICIENCY TESTING: COGNITIVE, LINGUISTIC AND STATISTICAL PERSPECTIVES

This presentation will provide a full description of a large scale assessment study of speaking proficiency, which analysed results from over 1300 tests administered in three languages (Spanish, English, Russian) to a total of approximately 350 adult language learners, by various organizations. A brief description of the procedures following the characteristics of the test will be given. Background information of the examinees, including full description of language learning histories and aptitude measures will also be presented.

Detailed results of the various administrations, including tester and examinee differences and characteristics, testing teams differences in performance, reliability data for teams, second ratings, and organizational differences, will be included in the discussion. Reliability and validity data of the new oral proficiency instrument obtained through the three experimental studies in Spanish, English, and Russian, will be presented.

In particular, results from the six rating factors introduced in the studies will be analyzed in detail towards a better understanding of the dimensions involved in oral proficiency, following previous work in this area (Higgs & Clifford, 1982; Sang, et al., 1986; de Jong & van Ginkel, 1992). Results will be interpreted in view of cognitive perspectives on the process of acquisition (Anderson, 1982, 1983, 1990; Carroll, 1986; Rummelhart, et al., 1986). These factors: interactive comprehension, structural control, lexical control, delivery, social and cultural appropriateness, and communicative strategies, showed different proportional contributions to the overall rating at different levels of global proficiency. The analyses to be reported, concentrate on the role of these factors, in terms of their participation in the explanation of variance of final ratings, as well as in their inter-correlations found for various examinee, tester, learning conditions, and proficiency levels. The multidimensional interpretation and findings of the current data in oral proficiency are in agreement with those models of cognitive psychology which establish declarative and procedural stages of cognitive processing, and for hierarchically related dimensions. Implications will also be drawn for the characterization of different types of learners (Skehan, 1989; Ellis, 1990, 1994). Linguistic characteristics and productions for the various languages will be examined and interpreted in view of the models discussed.

Input from those attending the discussion will be very important for the research strategies to follow in this area of research, and input from various perspectives will inform the workplan. Discussion of implications of these results will be encouraged.