ABSTRACTS

EXTRAITS

Conference on Second Language Performance Testing

Symposium sur le testing de performance en langue seconde
Recent years have seen a shift from large-scale established tests (e.g. TOEFL, Michigan) to instruments which are consciously related to specified demands of various university disciplines. The major characteristics of such instruments and procedures used in their development will be discussed and exemplified, including techniques for needs analysis, description of performance levels, general and specific elements of tests, the principle of replication, the incorporation of target use features into tests, balance between reliability and validity, complementation of subjectivity and objectivity, defects in correlational treatments, and the relationship between tests and curriculum.
In the literature of human measurement a distinction is frequently made between a knowledge test (sometimes referred to as a "paper-and-pencil" test) and a performance test. The difference between the two should be obvious from the respective names. A knowledge test measures the amount of information an examinee possesses about a given topic. The format is often multiple-choice, but can also be true-false, matching, fill-in, short answer or essay. A performance test is generally used only if there is a requirement to demonstrate proficiency in a specific task which cannot be measured by a knowledge test, e.g., musical ability, artistic skills, public speaking, foreign language proficiency, etc.

Knowledge tests have been the standard in most second language programs for the past several decades. A typical language knowledge test isolates parts of the language and asks the examinee to make additions, deletions, changes, etc., e.g., write in (or select from a set of alternatives) the correct suffix for a verb or adjective, change a sentence from the present to the past tense, or rewrite a sentence with a new subject. It could be argued that in all of these tasks some degree of language performance is taking place, but not the same kind of performance that is typical of a spontaneous communication situation. These kinds of tests provide useful information, but they do not necessarily show how well the examinee can coordinate all facets of communication in a given sentence.

Research has shown that in spite of the subjective nature of performance tests, if administered properly they can be scored with a high degree of reliability. Certain procedures must be followed, and an adequate training program is necessary for anyone involved in administering or scoring language performance tests.
This workshop will acquaint participants with the work being done by the Tests, Measurement and Evaluation Service of the Public Service Commission of Canada in the development of systems of evaluation to assess student performance in work related second language training programs of both English and French. It will describe the various stages of the needs analysis approach which is fundamental to the development of any evaluation system and which leads to the development of a framework of evaluation of language performance. This framework includes the underlying concept of language proficiency and the various uses which the measurement instruments will be put to.

Emphasis will be placed on the variety of measurement instruments and evaluation purposes of any framework and on the differences that can exist from one system to the next depending on the results of the needs analysis. Four major projects will be discussed by the testing specialists assigned to each and they will focus on a particular stage of the needs analysis process or outcome and on the uniqueness of their particular project.

Some of the topics to be discussed include the use of various instruments to assess evaluation needs; the development of language proficiency models based on the results of needs analysis; the development and validation of an evaluation grid of communicative competence; the analysis of various observation tools and of the results obtained from these tools; and the development of a performance oriented testing system based on specific work operations.

Participants will have a choice of four areas of interest within the framework of the workshop. They may choose to 1) engage in a construct development exercise for testing designed for Francophone public servants learning English; 2) apply the evaluation criteria of a communication grid to a sample of recordings of Anglophone public servants learning French; 3) analyse various observation tools used to assess the language progress of French and English public servants in an advanced language training program; 4) or devise a plan for a mastery/non-mastery criterion-referenced final examination for Coast Guard Cadets learning French.

Because of the number of presentations and the time limitation we propose an initial introduction of approximately 20 minutes giving an overview of the process used to arrive at systems of evaluation in the Service. The participants will then be invited to choose one of the four areas described above. All presenters are bilingual and discussion will be in either French or English.
Cet atelier a pour but d'initier les participants au processus utilisé par le Service de Tests, mesure et évaluation (STME) de la Commission de la Fonction publique canadienne, lors de l'élaboration de systèmes d'évaluation de la performance linguistique des étudiants/fonctionnaires inscrits à des programmes de français et d'anglais langues secondes. Cet atelier de travail permettra aux participants de connaître les diverses étapes d'une enquête de besoins dont les résultats sous-tendent l'élaboration de tout système d'évaluation conçu au STME ainsi que le construct qui se trouve à la base de chacun des systèmes. Ce cadre comprend donc le concept même de la performance linguistique ainsi que les divers buts et utilisations des instruments dont les systèmes sont constitués.

L'atelier a comme objectif principal de démontrer comment, dans tout système, il peut y avoir une diversité d'outils de mesure et comment chaque système peut être différent d'un autre, compte tenu des résultats obtenus lors de l'enquête des besoins. Les spécialistes en évaluation présenteront donc, dans leur atelier respectif, leur projet particulier en portant une attention spéciale à ce qui différencie leur système des autres systèmes élaborés au STME.

Voici quelques-unes des préoccupations qui seront touchées dans les sous-groupes: l'utilisation de diverses procédures dans une enquête de besoins en vue de l'établissement d'un système d'évaluation linguistique; l'établissement de modèles d'évaluation de la compétence linguistique reflétant les résultats obtenus lors de l'enquête de besoins; l'élaboration d'une grille de compétence à communiquer et sa validation; l'utilisation d'outils d'observation; enfin, l'élaboration d'un système d'évaluation de la compétence linguistique reliée à des tâches spécifiques de travail.

Les participants peuvent assister à l'atelier de leur choix. Le premier atelier a pour but d'impliquer les participants dans l'élaboration du concept théorique (construit) en vue d'un système de testing de l'anglais langue seconde; le deuxième atelier permettra aux participants d'évaluer la compétence linguistique des étudiants du français langue seconde à partir de critères de la compétence à communiquer tels que conçus et élaborés par le Service; dans le troisième atelier, les participants pourront faire l'étude de divers outils d'observation utilisés dans un programme où la formation linguistique tant du français que de l'anglais est plus poussée; le dernier atelier permettra aux participants de travailler à l'établissement de critères en vue de la certification des Cadets de la Garde Côtière en formation linguistique, français langue seconde.

Vu le nombre de présentations et les contraintes de temps, une introduction générale de 20 minutes précédera les présentations en sous groupes. Lors de la présentation générale, l'animateur expliquera les démarches utilisées au Service de Tests, Mesure et Évaluation en vue de l'élaboration d'un système d'évaluation. Suite à cette présentation, les participants pourront faire le choix de l'atelier qui les intéresse plus particulièrement.

Les présentateurs sont tous bilingues. Les participants pourront donc poser leurs questions en anglais ou en français.
Cito, the Dutch National Institute for Educational Measurement, in charge of the production of most of the central examinations in the Netherlands, has been developing a number of oral proficiency tests in German, French, and English to be used in secondary education. The purpose of the tests is to measure a student's ability to handle a foreign language in everyday life-like settings. Research was carried out in the following areas:

1. The validity of such tests and their format
   What should a test be like in order to measure the sort of performance we expect from students in everyday situations?

2. The development of reliable rating scales
   Teachers should have simple, but reliable, rating scales in order to assess their students' performance.

3. The training of teacher raters
   As these tests covered an area in which most teachers had little experience, it was found that a taped training programme increased agreement among raters considerably.

It is to this research programme, its results and the actual tests as they now stand, that the first part of this workshop will be devoted. Some attention will also be paid to a further development. The tests developed so far did not quite satisfy the needs of more advanced and more experienced learners. It was felt that these learners should also be able to engage in conversations on such general topics as unemployment, energy consumption and environmental problems. Cito has now developed some experimental tests which are based on simple texts on this sort of topic.

In the second part of this workshop the participants will be invited to assess a number of taped performances by Dutch pupils with the help of the training programme mentioned above. Actually working with the newly developed Cito oral proficiency tests may give participants a good idea of how these tests work.
A recent paper by Swain (December, 1982)* discusses four important points in the development of communicative measures for testing productive skills in classrooms. The points are:

1. Start from somewhere
2. Concentrate on content
3. Bias for best
4. Work for washback

The principles refer to (1) building on existing experience in communicative test development, (2) incorporating content in the test materials that is motivating, substantive, integrated thematically and interactive communicatively, (3) doing everything possible to elicit the learners' best performance, and (4) constructing the type of assessment instruments that will positively influence curriculum planning for teachers who naturally "teach to the test".

These principles will be illustrated with reference to a package of assessment materials currently being developed in the Modern Language Centre at UIDE for the testing of grade 9 and 10 students in French immersion programs, post-immersion programs, and minority French language programs. The package includes four writing tasks and two oral tasks in a unified framework, using source material whose theme is summer employment experience projects for youth in French-speaking communities in Canada. The package is called "À vous la parole" (Have your say) and refers to students' expressed desire to have input in the type of programs and opportunities which directly affect them.

*"Large Scale Communicative Language Testing: A Case Study"
(Plenary paper presented at the International Symposium on Language Testing, University of Hong Kong, Dec. 18-21, 1982.)
Conférenciers: COURCHENE, Robert  
Institut de langues vivantes  
Université d'Ottawa  
Ivan de Banheera  
Université de Montréal

Communication: EBAUCHE D'UN CADRE THEORIQUE POUR L'ELABORATION DE TESTS VISANT LA PERFORMANCE

Date: le vendredi 11 mars, 5h00 - 5h45

Dans le domaine des tests de langue la nouvelle importance accordée à ce que l'on peut faire avec la langue en opposition avec ce que l'on sait ce la langue a eu des répercussions importantes sur l'élaboration de tests visant la langue seconde.

Les chercheurs dans leurs tests mettent maintenant l'accent sur la performance du sujet soumis à l'épreuve: ils veulent ainsi préparer des tests qui permettront aux spécialistes de langues secondes de prédire le niveau de performance des sujets dans des situations de communication réelles, bref... les chercheurs veulent construire/élaborer des tests qui forceront l'apprenant d'une langue seconde à démontrer oralement la compétence plutôt que d'en rendre compte par un test écrit.

Cependant avant que les chercheurs ne puissent s'attaquer à l'élaboration de ces tests de performance d'importantes décisions d'ordre théorique et pratique doivent être prises, qui concernent le contenu, les paramètres des habiletés, des juges, la méthode, etc.

Nous examinerons donc ces décisions au cours du présent exposé/document et, notre but étant de proposer un genre de grille pouvant guider ceux qui travaillent dans le milieu à l'élaboration de tests, nous ferons fréquemment référence aux théories et aux tests couramment décrits dans les publications spécialisées.
This presentation examines the relationship of the notional/functional approach to pragmatic teaching and testing. Claims made on behalf of functional testing and against pragmatic tests (e.g., cloze procedure) by Farahdy (in press) are examined in light of recent research (J.D. Brown in press). In addition, the discussion includes reactions to papers presented at the conference and addresses the question: what is a performance test? It is suggested that pragmatically viable testing procedures must always relate to specific performances in specific communication contexts with well-defined purposes. The relationship of communicative competence and performance contexts is considered.
Nous nous proposons de présenter certains types d'épreuves d'évaluation formative ou certificative en français langue seconde ou étrangère, en faisant spécialement référence aux expériences qui ont concerné l'évaluation des compétences de travail des fonctionnaires internationaux de l'ONU à Genève et de l'UNESCO à Paris, mais aussi celle des compétences langagières générales d'adultes débutants.

Les caractéristiques de ces épreuves sont de tenter, pour faire porter l'évaluation sur les savoir-faire langagiers plutôt que sur les connaissances linguistiques, de faire appel chez les apprenants à des opérations mentales qui se rapprochent le plus possible de celles qui sont mises en œuvre dans des situations réelles de communication.

Il s'agit donc d'épreuves intégrant les composantes de l'analyse linguistique, faisant fréquemment appel au procédé de la simulation et tendant, par leur diversification, à prendre très partiellement en cause l'infinie variété des situations réelles de communication.
The workshop grew out of the experience of developing an oral interview test to be used in a study of the cultural assimilation of bilingual francophones. The test was administered to 273 francophone students at the University of Ottawa by two English-speaking psychology students. These "peer-interviewers" had been trained in the use of two evaluation instruments: Carroll's nine-band system of global evaluations and a system in which five criteria (pronunciation, accuracy, range independence, size) were scored separately. The purpose of the interviews was to assess the impression of language proficiency the subjects would make in social conversation with a peer. The limited time and funds available for data collection necessitated a quick, structured interview format and evaluation criteria that could be used by essentially naive raters.

The workshop is designed to share the lessons of this experience with classroom teachers and others who are often tempted to avoid oral testing, believing it to be too time-consuming and subjective to be of value in a testing program. Preliminary results indicate that fairly simple procedures can be learned and administered with relative ease and provide comparable results to those achieved by experienced raters using more complex procedures and scoring grids.

Participants in the workshop will examine and discuss guidelines for a) developing elicitation procedures, b) specifying an evaluation system and c) training interviewers. Participants will score taped interviews using different criteria and grids, and discuss the relative merits of the different procedures.
A la demande de la Faculté des Sciences sociales, l’Institut de langues vivantes a élaboré des tests de compétence pour mesurer le niveau linguistique de quelques-uns de ses membres. Les tests "cibles" doivent répondre aux exigences du règlement sur le bilinguisme de l'Université d'Ottawa en mesurant la performance de l'examiné en français ou en anglais langue seconde dans son domaine spécifique.

Puisque la Faculté des Sciences sociales comprend les écoles suivantes: criminologie, économie, psychologie, récréologie, sciences politiques et sociologie nos tests ont dû être pour ainsi dire taillés dans la matière même de ces diverses disciplines; en un mot "faits sur mesure".

L’exposé présentera la procédure suivie pour établir de tels tests.
Many U.S. universities have developed ESL-based training programs to improve the communication skills of non-English-speaking foreign teaching assistants (TA's). Some programs use performance tests to screen their graduates in terms of their potential success as classroom laboratory TA's. This presentation takes a critical retrospective view of a performance test developed for screening foreign applicants for teaching assistantships at U.C.L.A (Hinofotis, Bailey and Stern 1981). The testing instrument which had proved successful in an experimental setting (Hinofotis and Bailey 1981), requires some modification on the basis of an on-the-job needs assessment of foreign TA's. The problems of performance testing in this context are discussed in relation to Canale and Swain's (1980) model of communicative competence, with sociolinguistic competence emerging as a primary concern in the success of foreign TA's.


Many U.S. universities have developed ESL-based training programs to improve the communication skills of non-English-speaking foreign teaching assistants (TA's). Some programs use performance tests to screen their graduates in terms of their potential success as classroom laboratory TA's. This presentation takes a critical retrospective view of a performance test developed for screening foreign applicants for teaching assistantships at U.C.L.A (Hinofotis, Bailey and Stern 1981). The testing instrument which had proved successful in an experimental setting (Hinofotis and Bailey 1981), requires some modification on the basis of an on-the-job needs assessment of foreign TA's. The problems of performance testing in this context are discussed in relation to Canale and Swain's (1980) model of communicative competence, with sociolinguistic competence emerging as a primary concern in the success of foreign TA's.


Dans le cadre de cette table ronde, nous aurons l'occasion d'entendre des représentant-e-s de cinq organismes, qui ont des responsabilités d'évaluations langagières, nous présenter les divers problèmes auxquels ils doivent faire face. Les résumés qui suivent sont rédigés dans la langue de présentation.

During this round table, we will have the opportunity to hear representatives from five different institutions with language evaluation responsibilities tell us about the various problems facing them. The abstracts that follow are written in the language of the presentation.

PRIMEAU, Diane
Office de la langue française

L'Office de la langue française a produit de nouveaux examens de connaissance du français destinés aux candidats des trente-neuf ordres professionnels du Québec. Après quelques années d'utilisation d'un instrument s'appliquant à tous les candidats, l'Office avait décidé quelques problèmes et on était venu à se poser cette question: "Comment mesurer la connaissance du français appropriée à chaque profession en tenant compte du niveau de formation des candidats (secondaire, collégial et universitaire) sans pour autant multiplier indûment les instruments?" On discutera, au cours de la présentation, les solutions mises à l'essai et les résultats obtenus.

VOYER, Mireille
Canadien National et Air Canada

Pour résoudre certaines contradictions auxquelles nous faisions face avec des programmes axés sur la communication et des tests qui ressauraient surtout la connaissance du code, nous avons examiné un certain nombre d'instruments susceptibles de nous aider. Nos recherches nous ont permis de découvrir une formule de test entretien qui faisait partie d'un test plus général utilisé dans la Fonction publique canadienne. Partant de ce test entretien, nous en avons fait une adaptation en fonction de nos clientèles en gardant comme principales composantes la conversation libre avec un intervieweur et la discussion à partir de sujets proposés. Le contenu entier du test entretien est enregistré. C'est à l'interviewer, entraîné à cette fin, qu'incombe la responsabilité d'assurer autant de naturel que possible à l'entrevue. L'interviewer n'est pas impliqué dans la correction qui est faite par un jury de deux personnes, toujours les mêmes, qui se réfèrent, pour rendre leur décision, soit à leur expérience, soit à des enregistrements modèles des divers niveaux. Les cas, rares, de conflit sont confiés à une tierce personne pour vérification plus approfondie.
PROULX-DUBE, Renée
Service de Tests, mesure et évaluation, Fonction publique du Canada

La Direction générale du programme de formation linguistique de la Commission de la Fonction publique, dès 1978, modifiait ses programmes d'enseignement afin de répondre à des besoins langagiers reliés aux exigences langagières des postes en adoptant une démarche andragogique et en favorisant avant tout l'apprentissage de la langue comme moyen de communication. Le Service de Tests, mesure et évaluation a donc dû concevoir, élaborer, expérimenter et mettre en place des systèmes d'évaluation du rendement des étudiants en tenant compte de ces éléments. Le Service a élaboré des tests de performance dans les quatre habiletés, lire, écrire, parler et comprendre. Il a aussi élaboré des barèmes de correction qui tiennent compte des théories de base sur lesquelles reposent les tests. On présentera les problèmes que peuvent occasionner les tests de performance ainsi que ce genre de barème.

CHARLEBOIS, Pascal-André
Canadien Pacifique

Au Canadien Pacifique, la formation linguistique en milieu de travail vise à développer chez les employés l'habileté à accomplir dans la langue cible les tâches qui leur sont assignées. Cette emphase sur la performance nous a amené à repenser nos contenus, nos stratégies pédagogiques et nos outils d'évaluation. Nous avons mis au premier plan le vécu professionnel des employés, la fonction sociale de la langue et le savoir-faire. Afin d'articuler un testing de performance efficace, nous avons opté pour une approche systémique. Par divers moyens, nous mettons en scène les acteurs impliqués dans le processus de la formation linguistique des employés : les chefs de service, les employés apprenants, les collègues de travail et les agents de formation linguistique. Les chefs de service sont responsables du développement linguistique de leur personnel et de la mise en œuvre de la politique linguistique de la compagnie dans leurs services. Gérés par la base, la formation linguistique et le testing de performance sont ainsi intégrés dans la réalité de gestion des questionnaires. Cette stratégie intégrative favorise grandement l'utilisation de la langue cible par les employés.

WESCHE, Marjorie
Centre for Second Language Learning, University of Ottawa

One of the ongoing functions of the University of Ottawa's Centre for Second Language Learning is the preparation of second language tests in English and French. Most of these tests are destined for use within the bilingual university setting; however, tests are sometimes developed for outside organizations as well. Recent projects for well-defined clienteles have been characterized by a performance-oriented approach, in a departure from testing techniques used in large-scale French and English proficiency testing of undergraduates. These clienteles, ranging from graduate students in history to parking lot attendants, have in common the fact that their second language needs are real, and are definable in terms of specific language use situations and types of discourse. Considerations and procedures in the development of second-language performance tests for such target groups will be discussed.
First, for clarification, a definition of the term "bridge" as used in this paper:

(a) A "bridge student" is an ESL student who is almost ready to move out of ESL classes into regular programming (also referred to as "mainstreaming").

(b) A "bridge course" is an advanced ESL course which emphasizes writing skills, but also includes listening comprehension, speaking practice and reading. Students are encouraged to regard the "bridge" classes (which are offered at two levels) as the last English classes in which they should expect to receive special help as ESL students.

(c) "Bridge" testing procedures are used to assess student readiness for entrance to and exit from ESL "bridge" classes.

This paper describes the efforts of the ESL Department of one community college to help ESL students to cross the bridge from ESL classes into programs offered to English-speaking students. This mainstreaming is usually into such programs as academic upgrading, job training courses or college/university level courses.

The entrance and exit assessment procedures are described, as well as the reasons for them.

Procedures: Students requesting admission to "bridge" classes may be ESL students from our own program, non-native speakers of English referred from other college departments, or "walk-in" applicants.

All students are asked to take a 30-minute multiple-choice grammar test and to write a one-page 30-minute assignment on the subject of their future plans. They then have a personal interview in which their educational background and career goals are discussed. The writing sample is the main factor in determining the starting point for each student.

Exit procedures from ESL "bridge" courses are designed to meet the entrance requirements of the program or course which the student wants to take next. With the co-operation of the Adult Basic Education Department and the English Department, a final evaluation is made as a team. Again, a written essay forms the major part of this evaluation. In addition, a "report" on the student's overall performance is prepared by the ESL Department.

Reasons for choosing these procedures include:
- ESL students often fail the English course required in most regular programs (even though their academic ability in other subject areas for the program is adequate).
- ESL instructors find it difficult to estimate student readiness for mainstreaming because the standards of English for English-speaking students are more variable.
- Admissions/Advising sections of colleges and other institutions are frequently unable to place students whose first language is not English in appropriate English courses.
The Test in English for Academic (TEAP) Purposes is the result of several years of research into the ESL language needs of overseas students in Britain and into appropriate testing techniques related to these needs. The first stage of the project consisted of a survey of the levels, disciplines and institutions into which such students enroll. Subsequently, observational and questionnaire techniques were used to determine the general and discipline-related ESL language needs of these students. The data gathered served for establishing descriptive parameters, and for the design and validation of appropriate test formats for a test battery to assess performance levels in academic English for these students. Test development procedures, the resulting battery, and results of field-testing will be discussed.
L'Institut de langues vivantes de l'Université d'Ottawa a été chargé par le Centre National des Arts d'élaborer un test destiné à mesurer l'aptitude des candidats à l'emploi du CNA à exercer leurs fonctions ou partie de leurs fonctions en français.

Après une étude soigneuse du milieu de travail, deux grandes catégories de personnel ont été déterminées: le personnel administratif dont les tâches à accomplir en français n'étaient pas clairement définies et le personnel directement en contact avec le public pour qui le type de langue à utiliser était très spécifique, appartenant à des domaines d'activités précis (restauration/gestion des salles/guichets).

Deux tests différents ont été préparés: pour la première catégorie un test de compétence générale portant sur les quatre habiletés linguistiques; pour la deuxième catégorie un test de performance consistant en jeux de rôles et en une mesure de la compréhension auditive. Des grilles ont été préparées pour que la correction et l'évaluation des résultats soient rapides et aussi objectives que possible.
A study is presently being undertaken to investigate the value of certain aspects of a reading/writing assessment instrument. The instrument consists of a reading/writing task and a set of thirty hierarchically ordered questions used dichotomously to judge the quality of the learner's written product.

The reliability of the criteria to (NS) graders has already been established by the high inter-rater reliability of 80% (reported by Cumming and Mackay, 1982). The use of the evaluation criteria as (i) overt teaching components in an ESL writing course and (ii) the focus of self-assessment by learners, has yet to be demonstrated.

This paper reports upon an experiment intended to illuminate the value of these criteria as instructional components in an ESL writing course and as the focus of self-assessment attempts by students in academic writing courses. The students are undergraduate Science and Engineering students enrolled in an ESL academic writing course in the Centre for Second Language Learning at the University of Ottawa.

Eight intact groups, divided into four high-proficiency and four low-proficiency groups using university placement tests, are being studied. All eight groups were pre-tested at the beginning of the winter 82/83 semester with a reading/writing task graded by trained readers using the experimental criteria. Of the 8 groups, two high and two low proficiency groups are used as controls and follow the normal syllabus of the existing course. The remaining four groups were assigned to two experimental treatments, T1 and T2.

In experimental treatment T1, one high and one low proficiency group were given the results of their reading-writing pretest and the criteria used to assess their writing product were explained to them. They were each provided with a set of criteria and encouraged to use these to assess and improve upon their drafts of the written work required of them throughout the course.

Experimental treatment T2, (administered to the two remaining groups, one high and one low proficiency) duplicated T1, but, in addition, the detailed explanations of the criteria, and overt instructions in the use of the criteria, were provided by the teachers throughout the semester.

In this way it is hoped to provide relevant data to determine the effects of (i) involving the students in the evaluation process by providing them with explained criteria to assess their own work and (ii) providing overt instruction in the mastery of the criteria as learning and performance strategies.

* Participating researchers: F. Donegan, P. Hauptman, T. Paribakht and D. Roberge.
There may be other methods of testing listening than the traditional discrete point listening-reading, pencil-paper approach. However, are these methods cost-effective, practical and reliable? Perhaps there is something in between. Here is a comparison of two standardized ESL listening tests, the SLEP-L and CELT-L, which considers modalities, format, sub-parts, item type, student population, teacher reaction, student reaction, and the actual test results of over 200 adult students.

You be the judge. Which of the two tests does the better job?
ABSTRACTS

The Fifth Annual Language Testing Research Colloquium
THE FIFTH ANNUAL LANGUAGE TESTING COLLOQUIUM: OTTAWA, CANADA
March 13 - 14, 1983

PROGRAMME

Sunday, March 13 — MST 364

9:15 - 10:00  1. Towards Standards of English Language Proficiency for the
Health Professions: A Preliminary Study. (Charles
Stansfield, Don Powers)

10:00 - 10:45  2. An Empirical Investigation of Communicative Competence:
Designing Direct Oral Proficiency Tests in EFL for Hong
Kong Secondary Schools. (Peter Tung)

11:00 - 11:45  3. Development and Validation of Performance-Based Tests of
ESL "Survival" Skills. (John Clark, Allene Guss Grognet)

(LUNCH)

Sunday, March 13 — MST 364

1:30 - 2:15  4. The Relationship Between Background and Learner Variables
and Aspects of Communicative Proficiency (Lyle Bachman,
Adrian Palmer)

2:15 - 3:00  5. The Construct and Concurrent Validity of a New Test of
Listening Performance. (Dan Douglas)

(BREAK)

3:15 - 4:00  6. Comparison of Six Methods for Testing Reading
Comprehension in EFL. (Elana Shohamy)

4:00 - 4:45  7. Correlations and Distinctions among ESL Skills of
University Students. (Virginia Streiff)
Monday, March 14  —  SIMARD 140

9:00 - 9:45  8. On the Interdependence of L1 and L2 in Writing: An Update. (Michael Canale, Normand Frenette, Monique Bélanger)

9:45 - 10:30  9. The Use of Elicited Imitation to Assess Competence in ESL. (John Albertini, Joan Forman)

(BREAK)

10:45 - 11:30  10. The BLUR Exercise/Test: Another Pragmatic Exercise/Test? (Philip Carleton)

11:30 - 12:15  11. Rasch Calibrations of Cloze Tests for Native and Non-native Speakers of English. (Grant Henning)

(LUNCH)

Monday, March 14  —  SIMARD 140

1:30 - 2:15  12. What We Know About Language Acquisition and Its Implications for Practice. (Mary Ann Larsen-Pusey, Cynthis Darche Park)


3:00 - 3:45  14. A Proposed Research Program for Reading Comprehension. (Kyle Perkins) To be presented by Grant Henning.
Presenters: Stansfield, Charles and Don Powers
Educational Testing Service

Paper: TOWARDS STANDARDS OF ENGLISH LANGUAGE PROFICIENCY FOR THE HEALTH PROFESSIONS: A PRELIMINARY STUDY

Date: Sunday, March 13, 9:15 - 10:00

The growing interest in oral proficiency during the past decade prompted Educational Testing Service to undertake the development of the Test of Spoken English (TSE), a standardized test of oral proficiency for non-native speakers of English. The test has been validated for the selection of non-native teaching assistants applying to American universities.

The objectives of this study were to

(a) validate the TSE for the selection and certification of non-native health professionals,
(b) establish a set of procedures for determining standards of proficiency on the TSE, and
(c) pilot test these procedures in selected health professions in which the test could be used.

Groups of judges who were practitioners in four health professions—nursing, pharmacy, medicine, and veterinary medicine—judged the acceptability of the speaking performance of examinees who had taken TSE. Each group was asked to provide for their profession ratings for three distinct work situations, such as communicating with professional colleagues, teaching in the field, or working in a hospital. Judges were asked to indicate whether each examinee was at least minimally acceptable to function in each situation. Judges' ratings were then related to the examinees' TSE scores. A larger group of consumers of medical care gave more global ratings to the same examinees for each of the four professions.

Alternative methods were applied to determine possible ranges of cutoff scores on the TSE. One of these methods involved a consideration of the consequences of different decision outcomes, e.g., whether the certification (or licensing) of an unacceptable speaker is a more serious or less serious error than failing to certify an acceptable speaker. A range of possible cutoff scores was computed for each situation within each profession.

The results of the study support the interpretation of the TSE as a valid measure of oral language and provide some progress towards the setting of appropriate standards of oral language proficiency within the health-related professions.
Canale and Swain, in 1979, had proposed that grammatical competence and sociolinguistic competence would be two of the three areas, or "systems of knowledge", to be included in a theoretical framework of communicative competence. Two direct tests of oral skills, a simulated telephone conversation and a subsequent job interview, were designed for senior secondary EFL students in Hong Kong, in an effort to construct meaningful and valid measures of communicative ability. Various devices were used to elicit testees' performance with respect to specific grammar points, selected vocabulary items, and relevant "sociocultural rules of use", a component of Canale and Swain's sociolinguistic competence. It was felt that two types of rating scales were needed for assessment, graded scales of performance and a checklist of the sociocultural rules observed. The graded scales were used to assess testees' performance in areas such as pronunciation, fluency, grammatical structures, vocabulary, and to record the overall impression of the raters. Two markers were employed on each testing occasion, and a total of 50 students from five different schools and two grade levels were examined. To obtain sociolinguistic data on the testees' background, questionnaires were administered after the tests.

Statistical analysis of the data shows that the graded scales and the checklist recorded different dimensions of the testees' performance and suggests that in a communication test, both integrative and discrete-point marking are necessary. Furthermore, results show that the graded scales all seem to be measuring a single aspect of the testees' performance, and hence a global impressionistic judgment could replace detailed assessments in carefully defined categories of grammatical competence for ordinary classroom purposes at least. Inter-rater reliability has also been demonstrated for the tests.
The recent arrival in the U.S. of large numbers of non- or limited-English speaking refugees from Southeast Asia and other areas has created the need for development of performance-based testing procedures to assess the refugees' ability to accomplish basic "survival" tasks in English, including, for example, giving basic autobiographical information orally, telling time, asking for and giving directions, shopping for food and clothing, describing emergencies, dealing with health and housing matters, reading printed caution signs and other similar "sight-word" materials, and filling out simple data forms.

This paper describes the development, trial administration, and final revision of the such test batteries, one for use in intensive pre-departure language training programs abroad, and the other for use in similar post-arrival training in the United States. The considerations involved in defining the content areas to be covered in the tests are described, as well as the procedures followed in developing testing formats and question types capable of assessing examinee performance in a face- and process-valid way within the stringent and less than optimum administration conditions under which the tests would need to be used. The testing approach finally adopted (booklet-mediated face-to-face interview) is described in detail, as well as the individual tasks represented in the test. Also described is the use of an interactive computer program (SCALAR) to establish sub-scales for listening comprehension, fluency, communication, and reading/writing against which individual item performance was evaluated for final item selection.

The paper concludes with a discussion of the obtained internal test reliabilities, scoring reliability, concurrent validation data based on instructor judgment of overall language proficiency, and description of plans for follow-up validation studies.
Researchers have noted that research in second language acquisition has been hampered by a lack of valid criterion measures of language proficiency (Larson-Freeman and Strom, 1977). Indeed, there is little agreement on what constitutes language proficiency or what aspects of language proficiency various types of tests measure. A case in point is the frequently made distinction between discrete point and integrative tests and models of language acquisition and performance which have been developed on the assumption that this distinction is valid.

Bachman and Palmer (1982) examined several theoretically plausible models of communicative proficiency and found that the model which best explained their data included three factors: a general factor, a linguistic-pragmatic competence factor, and a sociolinguistic competence factor. In this study, we examine the relationship of a number of background variables and learner characteristics to the language competencies previously identified and to performance on different kinds of tests (communicative and non-communicative).

Background variables include length of stay in an English speaking country, amount of exposure to an acquisition intensive environment, amount of exposure to a formal learning intensive environment, and age of first exposure to English. Learner characteristics include age, degree of integrative motivation (desire to acculturate), degree of instrumental motivation, and extent of perceived use of conscious rules.

Specific hypotheses regarding the relationship between type of motivation and different language competencies and between environment of exposure to English and performance on different types of tests are tested.
In the autumn of 1982 a new listening performance test was developed for use in standardized testing programs at the English Language Institute. The test involves the interpretation of stressed elements in English sentences. For example, subjects hear sentences such as the following:

A. I wanted sugar in my tea.
B. Does the movie start at nine A.M.?
C. No, Jane's brother's a novelist...

The underlined words are spoken with emphatic stress, and the subjects are asked to choose a response from among three alternatives. Subjects must decide, in successive sections, what the speaker probably meant (as in example A, the speaker didn't want cream), what the speaker wanted to know (as in example B, did the movie start in the morning or in the evening), and what the speaker would probably say next (as in sample C, ...not her husband).

This paper will be a discussion of the theoretical background to the test with reference to pragmatic theory, and a report on the statistical analysis of the results of some 500 test candidates. Suggestions will be made for the improvement of the instrument, especially in the area of pragmatic validity.
Looking for effective procedures for testing reading comprehension (RC) is important for teachers, administrators, test developers and researchers. Most RC tests used today use the multiple choice (MC) procedure which requires the test-takers to select an answer out of several alternatives. This provides the tester with indication of the extent to which the test-taker can comprehend a given text. Other testing procedures for RC such as open-ended questions, summaries and translations are used as well, although not as commonly as MC because the scoring is not as efficient. MC type tests have been criticized on the grounds of strong method effect over the trait of RC, on the difficulty in writing good questions, and the need to pre-test items before they can be used efficiently. The language that should be used on RC tests is also a controversial issue. In most RC tests the questions are presented in the second language. Thus a student's wrong answer may not necessarily result from not understanding the text but perhaps from not understanding the questions. Some testers suggest giving the questions in the first language of the test taker. These are only some of the questions which are used in an attempt to look for effective ways of testing RC.

The paper reports on a large scale study which examined and compared six common procedures for testing RC in order to find out to what extent the method used for testing affected the score of the students on RC. The six procedures are: 1) MC - questions in L-2 (English); 2) MC - questions in L-1 (Hebrew); 3) open-ended questions in L-2; 4) open-ended questions in L-1; 5) summary in L-2; 6) summary in L-1. The six methods were all based on the same texts - fifteen short passages and two longer ones, and were administered to a sample of 1900 randomly distributed high school students studying EFL in Israel. The results of the analysis provide information on the effect of the methods of testing RC on the trait of reading comprehension.
Recent studies in second language testing have emphasized the unitary nature of adult ESL students' underlying linguistic competence in the second language (Rachman and Palmer, 1981; Oller and Hinofotis, 1980; Oller, 1979). However, ESL test practitioners and instructors typically observe differences in student test performance among the major skills of reading, speaking, listening, and composition. The study to be reported involves 593 international students who attend a major American university, including January, May, and August 1982 test groups subsequently divided into four ESL proficiency groups from low to high. The study investigates the following questions.

1. Does a factor analysis of six ESL subtests (oral interview, composition, listening, grammar, vocabulary, and reading) result in a single underlying factor which may be inferred to reflect global second language competence?

2. Across four ESL proficiency level groups, do analyses of variance reveal significant differences among ESL skills, e.g., listening and composition? Within each proficiency group, do t-tests for non-independent groups reveal significant differences among major skills? Which ESL skills are most closely associated; which diverge most from one another, according to correlation patterns and t-test results?

3. Given a spectrum of test types, from the integrative and subjectively rated oral interview and composition to the discrete point grammar and vocabulary measures, what do linguistic error analyses reveal about the competence versus performance of six randomly-selected students (representatives of the three lowest ESL proficiency groups)? What do such analyses reveal about the relationships between type of test and individual performance?

An oblique factor analysis employing reliabilities on the diagonal and an alpha factor analysis with varimax rotation were conducted with raw scores, with January, May and August test groups treated separately. Both procedures reveal a single factor for all test groups. However, analyses of variance reveal significant differences between proficiency groups, and t-tests reveal significant differences among skills within proficiency groups, as great as 50 percentage points between subtest scores. Although the subtest types differ, speaking and listening are the most closely associated skills consistently across January, May and August groups; speaking and reading consistently diverge the most. Error analyses suggest (i) some specific improvements needed in both integrative and discrete point subtests and (ii) a narrowly-defined grammatical competence superseded by a larger communicative competence in relation to these major test types.
This paper updates the research we presented at the 4th Annual Colloquium on Language Testing Research. There we reported on two pilot studies bearing on Jim Cummins' (OISE) Language Interdependence Hypothesis. This widely influential hypothesis asserts that literacy-related aspects of language proficiency—such as involved in a formal writing task—are interdependent across an individual's first and second (or other) languages, i.e., manifestations of a common underlying proficiency. Our pilot revealed a significant but modest overall correlation (r = .33, p = .01; N = 106) between holistic scores on narratives in French (L1) and English (L2) written by the same Grade 9 and 10 students in advanced language arts classes in four French-language secondary schools in Ontario.

The present paper re-examines Cummins' hypothesis in light of results of our major study of students writing in French and English. Several improvements have been incorporated into this study. For example, the overall sample includes 1722 students enrolled in general and advanced Grade 9 and 10 classes in nine French-language secondary schools throughout Ontario; in both French and English, students wrote in two different genres (first-person narrative and exposition of a procedure); finally, the same four trained scorers have independently evaluated each of the four texts (two genres x two languages) for a randomly selected subsample of 50 students. As in the pilot study, each test has been scored in two different ways on two different occasions: first, holistically in terms of overall quality and second, analytically with reference to a five-level scale for each of five general criteria (language usage, norms for written text, individual expression, unity of forms and ideas, and effectiveness as an act of communication).

Three types of results are to be reported. First, inter- and intrarater reliability estimates will be examined for each scoring procedure. Second, two validity analyses will be discussed: one concerned with the distinctions drawn among the five general analytic criteria, the other with the relationship between the five analytic scores and the holistic score. Finally, we will examine the relationship between students' L1 and L2 writing with respect to genre, holistic scores and analytic scores.
Slobin and Welsh (1973) have demonstrated the usefulness of the task of elicited imitation (EI) for assessing children's linguistic competence. Use of this procedure was based on the premise that sentence imitation is filtered through the child's linguistic system. They argued that since EI is performed without communicative intent or contextual reference, the task strains a learner's abilities and provides a conservative estimate of linguistic competence.

Slobin and Welsh's study dealt with very young children. As children become older, their short term memory span may enable them to repeat a sentence by rote, without full comprehension. When a delay is introduced, however, children apparently are less able to repeat sentences they cannot comprehend (Mcdade, Simpson, & Lamb, 1982). Spitzé and Fischer (1981) demonstrated the usefulness of EI in assessing global linguistic skills of adults learning English as a second language. Since adults appear to have a greater short-term memory capacity than children, a delay should probably be introduced in the EI task. Further, if the goal is to create a diagnostic test of specific syntactic structures, the stimuli need to be refined.

The syntax Specific Test, which is designed to test specific syntactic structures of English, is a visually presented EI task which incorporates a delay in its administration since the response is in written form. The test was administered to 53 non-native speakers (enrolled in three different college-level intensive ESL programs), 47 American college students and 116 hearing-impaired American college students. Results of this test administration indicate that non-native and hearing-impaired subjects reconstructed the stimuli during the EI task. This paper will argue 1) that having subjects write their responses provides delay and thus adequate control for rote memorization and 2) that the errors produced are a result of grammatical processing and thus reflect grammatical competence.

References


A new kind of exercise requiring students to read an entire text in which sentences became progressively blurred was designed to measure their overall English proficiency as well as to help students develop good reading skills. But is such an exercise a valid and reliable measure of a student's English proficiency?

Over two hundred students in a university intensive English program took a BLUR exercise, a composition test, and several standardized tests measuring listening, structure, vocabulary, reading, and writing skills. The students also answered a questionnaire on two pragmatic tests and the BLUR exercise to determine whether they 1) found each measure interesting, 2) tried to do their best, and 3) considered the measure to be a valid measure of their English proficiency.

Pearson product-moment correlations and factor analyses using principal factor with iterations provided the following information:

1) Correlations of the BLUR exercise with specific skills tests were positive and moderately high, ranging from .50 to .70 (p .01). The BLUR also correlated positively and highly .70 (p .01) with the exit test battery average, indicating that the BLUR was a good measure of overall proficiency.

2) The BLUR loaded positively on the principal factor, loading more highly than the vocabulary and reading subscores of one standardized test and the listening subscore of another.

Cross-tabulations of questionnaire results with class and with native language revealed that the students found the BLUR exercise and the pragmatic tests interesting, tried to do their best on them, and felt that each measured their proficiency in English.

Error analyses showed that the students make many of the same errors when doing the BLUR exercise as when completing other tasks which require writing and speaking in English.
Rasch Model latent trait analysis has recently been applied successfully to foreign language test development. Among the advantages of this procedure have been (1) elimination of boundary effects, (2) identification of invalid respondents, (3) sample-free item measures, and (4) test-free person measures. Cloze testing has long been useful as a measure of global language proficiency and as a means of observing developmental error trends.

The present study purports to Rasch Calibrate cloze item difficulty for native speakers of English and for Egyptian EFL learners at two proficiency levels -- about 100 university students in the total sample. 100 Random Cloze deletions will enable comparative difficulty analysis by language background and proficiency level. It is hoped that this procedure will introduce a response-probabilistic paradigm for the investigation of developmental trends in language acquisition.
This paper reviews twenty-five years of investigation in first and second language acquisition and addresses the implications of this research for the construction of second language proficiency measures. The main issues of language acquisition center on three topics: (1) the psychological processes involved in the development of both competence and performance, (2) the description of developmental stages depending on discrete syntactic/morphological features, and (3) the ways in which psychological processes interact with language input in the acquisition of the target language.

An exploration of these issues leads to an integrative theory of second language acquisition which provides the following implications for the construction of language proficiency measures suitable for both research and adequate placement of students in instructional programs:

(1) Comprehension is probably a more valid measure of competence than either imitation or elicited production.

(2) The sampling of syntactic/morphological features must be large enough to account for the integrative nature of the acquisition process.

(3) That multiple choice items for secondary school students and adults can be written to reflect documented stages of target language development through a careful selection of appropriate distractors.
In trying to identify relevant dimensions of oral proficiency (in English L2 as much as in any other foreign/second language) it is important to go beyond the speech act level and also investigate into the interactional structure of the learner/native speaker discourse. In my paper I will suggest that the ability to give and elicit feedback in its broadest possible sense is one such dimension which contributes to the development of individual differences among learners. This specific ability has hardly been studied in any detail so far (see, for example, Gail 1981, or pragmatically, Perdue 1932). Theoretically it is considered to be a valid concept within the broader framework of social interaction as well as within a modified input hypothesis. Empirically this construct seems to be rather easily accessible insofar as feedback processes (being the verbal or non-verbal indicators of the underlying feedback ability) are to be found in any natural or elicited interactional language data like conversations, interviews, role-plays etc. It is one of our basic assumptions that these feedback processes are changing in number, type, and quality as the learner makes progress towards the L2 (linguistic and social) norms and becomes more proficient interactionally. (The influence of the L1-based prior knowledge of a learner, of his/her linguistic, cognitive, and social background experiences on mastering communicative tasks in L2 cannot be dealt with in this paper; cf., however, Vollmer 1982.)

Having given a brief outline of the study and its theoretical perspective I will then focus on the description of different types of feedback processes and their linguistic means of realization by advanced German learners of English as L2. It will be shown that both self-initiated and other-initiated feedback processes can be ordered in rather systematic ways along criteria like implicit/explicit, perception of role relationship and interactional setting, degree of certainty about one's own knowledge (linguistic, factual, psychological), level of aspiration of comprehension/understanding, degree of awareness/consciousness and evaluation of communicative flow etc. Thus feedback processes will be defined as certain moves in oral interaction which are discoursal and structural in nature, yet can be either metalinguistic or metacommunicative or just as well illocutionary and/or propositional in form. (Repair and face-work are only part of the notion of feedback as introduced here.) Each of the types mentioned will be illustrated by at least one example taken from Osnabrück corpus of elicited interview/role-playing data.

The last section of my paper will be devoted to discussing questions of how feedback processes could be used in evaluating pragmatic competence in L2, of how progress in this area could be defined in operational terms and "measured" quantitatively as well as qualitatively, of how feedback ability in second language learners is acquired and could be further developed (e.g. by instruction).
This paper will present two research proposals for conducting future L2 reading comprehension research. The first proposal involves eye-voice span (EVS); the second, text processing.

This paper will suggest the use of eye-marker and line of sight approaches for data collection. Some interesting research questions for L2 reading comprehension include: does EVS increase with more highly/less highly constrained syntactic structures? across different language groups? across proficiency strata? does EVS extend to natural intrasential boundary points? If so, to what extent is EVS related to language proficiency? to short-term memory? what language groups exhibit significantly different fixation functions? on what kinds of structures? at what proficiency levels?

Studying the processes and products of text structure generation by patients who have impaired language abilities due to damage sustained to the central nervous system could give some insight on the neurological bases/factors that determine how text structure is created. Patients having transcortical motor aphasia or Broca's aphasia would be needed for such research. Patients with transcortical motor aphasia (also called frontal dynamic aphasia) exhibit no phonological impairments and no salient linguistic deficits with respect to lexical or grammatical formatives; their auditory and visual comprehension are relatively intact. Ideally, one should be able to obtain data bases from L1 and L2 readers who have no language impairments for comparison with the aphasia patients data base.

*(Paper to be presented by Grant Henning)*