Shiken: JALT Testing & Evaluation SIG Newsletter. 5 (1) April 2001 (p. 2 - 3)

The challenge of speaking:

Research on the testing of speaking for the new TOEFL®

Tim McNamara (The University of Melbourne)

Since 1997, teams of researchers have been working intensively on the design of a new fully communicative TOEFL® to replace the existing one, whose specifications still reflect the assumptions of an earlier era of language testing. The current TOEFL is ultimately based on models of language and measurement dating back to the 1960s, to what Spolsky (1975, cited in ALTE 1995, p. 13) calls the psychometric-structuralist period of language testing. Although the current test has changed somewhat in the direction of whole task performance, in that reading and listening passages are longer, it provides a contrast to tests such as the IELTS™ test in that it does not explicitly focus on the contexts or roles in which international students will be involved in communication in academic settings. The impact of the current TOEFL on language teaching has long been considered problematic, although actual studies of the washback of TOEFL provide a somewhat mixed picture.

In response to ongoing critical discussion of the validity of the existing TOEFL, the Educational Testing Service some years ago introduced a project entitled TOEFL 2000 whose goal was a blueprint for a new test. In 1996 this project was restructured and placed under the direction of Dr. Irwin Kirsch of ETS. An initial framework document was produced which set out the conceptualization of the new test in broad terms. The new TOEFL would be a contextualized test of English for academic purposes, reflecting the roles and tasks of international students on a campus in an English speaking country. Four specialist teams were created, one for each of the macroskills, to develop the specifications for the test in each area. Each team consisted of five people, two from ETS (an experienced test developer and a psychometrician) and three external consultants, applied linguists with specialist research competence in the area. The teams have met three or four times a year since the middle of 1997 and have developed detailed specifications for each macroskill, and have commissioned a number of research studies to inform their decisions. The process is now more or less complete and the project has moved into the implementation stage, with the new test to be introduced by 2003. A feature of the test is the use of integrated tasks. These tasks integrate performances across the language macroskills, so that material from a lecture may be used as the basis for a speaking task or a reading passage may lead to a writing task which uses content from what has been read.

In the area of speaking, a number of research projects have been carried out. The first of these attempted to implement the framework used in research on language tasks by Peter Skehan (Skehan, 1998). Skehan has argued that the framework, which controls conditions of the implementation of tasks to vary their cognitive demands, could be used to develop a scale of difficulty in speaking tasks in terms of which ability in speaking could be mapped and thereby defined. This approach had great appeal for the project, as Irwin Kirsch's work in literacy had developed techniques for describing task demands and using them to define ability in reading. A large-scale project was carried out at the Language Testing Research Centre at The University of Melbourne to explore the use of Skehan's framework in this language test. Skehan himself came to Melbourne to assist with the design of the study. A series of narrative tasks was used, based on picture stimuli. Candidates told the narratives under performance conditions derived from Skehan's framework and predictions were made about the impact on performance of those conditions (they were predicted to affect the difficulty of the task). Interestingly, the data did not support the hypothesized effects of the performance conditions: there were no significant differences for

all but one of the conditions, and for this one the direction of the impact was the reverse of that predicted. While the results of the study were in a sense disappointing, in another sense they were exciting as they forced a re-evaluation of the underlying framework. This was a case where language testing research which drew on work in second language acquisition was in position to feed its results back into SLA, where they are likely to have substantial impact (a study based on this research is to appear this year in Language Learning).

The integration of language proficiency assessment within contextualized tasks raises issues of the criteria by which performance will be assessed. A further study, again in Melbourne, has looked at what raters pay attention to in assessing performance on contextualized tasks, especially integrated tasks. A number of performances from piloting of test materials were presented to raters who were asked to carry out think aloud protocols as they listened to the performances. The comments of raters were then categorized to see what was salient in their judgements (the raters were not given a rating scale to use but were asked to comment on what they noticed in the performances). The data were used to make recommendations about the development of rating scales, and to explore the issue of narrower and broader definitions of proficiency on cognitively demanding tasks in a second language.

Despite the enormous amount of work that has gone into the development of specifications for the speaking component of the new TOEFL, and the exciting program of research that has accompanied it, the constraints on delivery of a speaking component in a computer-mediated testing environment have proved very difficult, and at the time of writing it is not clear to what extent these constraints will permit the introduction of a test of speaking in the short term. It is clear however that speaking is firmly on the TOEFL agenda, and its implications for English language education in Japan and other countries in the region will be profound.

References

Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P., & Taylor, C. (2000). *TOEFL 2000 framework: A working paper*. TOEFL Monograph 16. Princeton, NJ: Educational Testing Service.

Kirsch, I. S., Jungeblut, A., & Mosenthal, P. B. (1998). The measurement of adult literacy. In T. S. Murray, I. S. Kirsch, & L. Jenkins (Eds.), *Adult literacy in OECD countries: Technical report on the first international adult literacy survey*. Washington, DC: U.S. Department of Education, National Center for Education Statistics. Retrieved March 10, 2009 at http://www.uis.unesco.org/TEMPLATE/pdf/LAMP/09_FRAMEWORK_ALL_Prose%20and%20Document%20Framework_Dec2005.pdf

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Spolsky, B. (1975) Language testing: Art or science? Paper presentation. Stuttgart: fourth AILA International Congress. (Cited in ALTE *Materials for the Guidance of Test Item Writers* 1995, p. 13. Retrieved March 10, 2009 at http://www.alte.org/projects/item_writer_guidelines.pdf.)

HTML: http://www.jalt.org/test/mcn_1.htm / **PDF**: http://www.jalt.org/test/PDF/ McNamara1.pdf