SHIKEN: The JALT Testing & Evaluation SIG Newsletter, 7 (3) Autumn 2003. (pp. 2 – 7).

OPINION PIECE:

TOEIC® : Tried but undertested

by Mark Chapman (Hokkaido University)

In 2001 Tim McNamara reported on the development of the new TOEFL® test in the pages of *Shiken* (2001, pp. 2,3). Educational Testing Services (ETS) has recently announced that the new TOEFL will be launched in 2005 as an integrative test of all four linguistic skill areas. Scores will report the abilities of candidates in relation to the skills required for study at an English speaking university. This re-launch of the TOEFL marks a drastic change from the norm-referenced test of listening and reading that it once was to an apparently criterion referenced test of both receptive and productive skills. In his article McNamara (p. 2) stated that ETS started to consider redesigning the TOEFL in response to "ongoing critical discussion into the validity of the existing TOEFL."

This discussion is reflected in a brief Internet search. In response to a search for TOEFL research, Google returns 282,000 possibilities. *Language Testing* published eight separate articles about the TOEFL between 1990 and 2003. This critical discussion is extensive when compared with research into another ETS test - the TOEIC®. Google returns only 13,900 possible sites for TOEIC research, less than 5% of that for TOEFL. Language Testing has no dedicated articles about the TOEIC. Perhaps the most telling figure however, is for research into TOEIC published by ETS. ETS has released 69 research reports into TOEFL, with an additional 17 technical reports on this exam between 1977 and 2002. However, for the TOEIC there are only three full research reports. In addition there was an initial validity study in 1982 and one technical manual. This data begs the question: "Why has ETS produced 23 times more research reports on the TOEFL than on the TOEIC?"

"Why has ETS produced 23 times more research reports on the TOEFL than on the TOEIC?"

There are several possible answers. TOEIC and TOEFL were similar in many ways before the Test of Written English was introduced to the TOEFL in 1996. The only significant difference was that TOEFL reputedly focused on academic English and TOEIC on the language of business and commerce (see Gilfert, 1995 for a fuller comparison of these two tests).

ETS may have seen little benefit in pursuing the same degree of research into TOEIC that had already been conducted into TOEFL. Moreover, ETS' experience of extensively publishing research into their own test with the TOEFL may have caused them to question whether this process invites skepticism and further critical investigation. Shohamy (2001, p. 148) reports that "there is low trust on the part of the public with regard to research conducted by companies that also develop and market tests, in a similar way that there is research conducted by profit-making drug companies on the drugs they produce." Whether or not this was a factor that prompted ETS to avoid publishing extensive in-house research on the TOEIC is, of course, a matter of speculation.

TOEIC came about as a result of a request by the Japanese Ministry of Trade and Industry to ETS. ETS may have felt that in fulfilling the request there was no necessity to further substantiate the test created. The most likely answer lies with the nature of the end user of both tests. TOEFL scores are intended to provide a reliable measure of the linguistic competence of candidates for English speaking universities. TOEIC scores indicate the proficiency of non-English speaking employees of corporations. Many universities have the resources and expertise to investigate the claims made for the TOEFL by ETS, whereas companies are far less able to challenge the validity of TOEIC. Overseas students who enter a university in an English speaking country on the strength of a TOEFL score are likely to be initially enrolled in a language program. The purpose of such programs is specifically to prepare the learners for the linguistic skills required for their studies. Instructors in these programs have a clear view of the skills students come equipped with and the level they need to attain. Hence, the shortcomings of TOEFL scores as a predictor of the competence required to study at an English-language medium university are readily apparent. This has been acknowledged by ETS (Jamieson et al., 2000, p. 3) with the admission that "those who use TOEFL test scores in selecting students for undergraduate and postgraduate programs increasingly express concern that many international students who are admitted with high TOEFL test scores (i.e., above 550) arrive on campus with insufficient writing and oral communications skills to participate fully in academic programs." The feedback mechanism between test maker, test taker and end-user is reasonably effective in the case of the TOEFL. This has eventually resulted in the test being redesigned to better meet the requirements of the end user; in this case, English-language medium universities.

Despite the TOEIC now being in use for almost 25 years it has not changed at all. It is still based on the structuralist, behaviorist model of language learning and testing that informed discrete-point testing. If ETS has accepted this model is no longer suitable as a basis for the TOEFL, why has TOEIC not been treated similarly? Surely the lack of critical research is a major factor along with the lack of an effective feedback mechanism from end user (corporations) to test maker. TOEIC cannot have been ignored by ETS due to its minority status: more people take the TOEIC every year now than the TOEFL. In 2002 more than 2.8 million individuals registered to take the TOEIC in more than 60 countries worldwide (ETS, 2003). This is more than twice the number that took TOEFL in the same time period. Given this importance in business terms of the TOEIC to ETS, it is perhaps even more surprising that there is no indication of TOEIC receiving the same degree of research attention devoted to the TOEFL.

"TOEIC cannot have been ignored by ETS due to its minority status: more people take the TOEIC every year now than the TOEFL."

The small quantity of existing research into TOEIC provides conflicting evidence and can be grouped into three general categories. Firstly, there are the previously mentioned research reports and technical manual published by the test producer. Secondly, two independent reviews of the TOEIC by Kyle Perkins (1987) and Dan Douglas (1992) that are both mainly based on data supplied by ETS. Finally, there are a small number of studies into the TOEIC conducted with independent data (not generated by ETS). As may be expected, the reports financed and published by ETS (Woodford, 1982; Wilson, 1989; Dudley-Evans & St. John, 1996; Boldt & Ross, 1998) provide broad support for the reliability of the TOEIC and its valid use as a direct measure of listening and reading and an indirect measure of speaking and writing. The two independent reviews by Perkins and Douglas both support the claims made for the reliability of the TOEIC by ETS. Perkins is largely supportive of all claims made for the TOEIC by ETS; however, the references he quotes in his review indicate that he only used literature published by ETS in forming his opinion. Douglas is somewhat more critical, but only in the sense of questioning the relevance of TOEIC items to the skills actually required in the world of international business and commerce. Again, Douglas does not appear to have investigated beyond the test items and the ETS reports. These two reviews need to be considered in the light of evidence provided by research conducted with independent data.

Three reports have provided data that conflict with ETS research. Childs (1995) is very critical of the TOEIC. His independent data suggests that the reliability estimates provided by ETS are overstated. He also concluded that the standard error of TOEIC scores is greater than the published ETS figure, making TOEIC scores less reliable as a measure of individual progress as score gains tend to be within the test's SEM. Hirai (2002) also expressed doubts about the ability of TOEIC to predict individual oral and written English proficiency. In a study conducted with employees of a major Japanese company, he suggested that the TOEIC was especially unreliable as a predictor of spoken English for individuals with intermediate range TOEIC scores (approximately 450 - 650). Hirai found that TOEIC scores had a low correlation (around 0.5) with BULATS scores, a test of writing in a business context. Finally, an unpublished MA dissertation (Cunningham, 2002) reported that the TOEIC was a very poor predictor of communicative competence and was not at all suitable for measuring gains in communicative performance. He used a self-design test battery, and while the research should not be entirely discounted, the fact that the TOEIC was not compared to an established test needs to be borne in mind.

Other authors (Gilfert, 1996; Eggly et al., 1997; Robb & Ercanbrack, 1999) have also used TOEIC in research projects but the conclusions they draw are either unsubstantiated (Gilfert) or not directly related to the reliability or validity of the TOEIC.

The lack of research into TOEIC is troubling in two ways. Firstly, the great popularity of TOEIC (almost 3 million registered candidates per year) means that it is one of the most taken language proficiency tests in the world. This fact alone should attract independent researchers' attempts to verify the claims made by the test maker. Secondly, the little independent research that has been carried out has been largely critical of the TOEIC. Doubts have been voiced over several claims made for the test by ETS. This combination should be enough to spur further critical discussion into this increasingly important test. Some areas that would be of interest include:

1. Correlations between TOEIC scores and direct, established tests of speaking and writing to establish whether TOEIC is a reliable predictor of these skills. It would be especially useful to investigate subjects with scores around the mean TOEIC score in Japan (approximately 450).

2. The linguistic skills required by the end users of TOEIC. It would be helpful to know what both employees and employers require in terms of linguistic proficiency. Research could help to establish the skills required, which would act as the construct for the TOEIC. If the precise construct is unknown, it is difficult to criticize the validity of the test.

3. The washback effect of the TOEIC. How does TOEIC influence learner motivation and study? Does TOEIC encourage learners to develop skills that are useful to their employers? Does TOEIC affect how teachers run classes for corporations utilizing the TOEIC?

These three areas would help to guarantee the best possible test was being produced for both test takers and the corporations that are frequently paying for the TOEIC. The example of TOEFL shows that extensive critical discussion of a test can lead to consistent development and improvement of the test. TOEIC users would benefit from such a discussion and the time for this to begin is surely imminent.

References

Boldt, R. F., & Ross, S. (1998). *Scores on the TOEIC® (Test of English for International Communication) test as a function of training time and type.* Princeton, NJ: Educational Testing Service.

Childs, M. (1995) Good and bad uses of TOEIC by Japanese companies. In J.D. Brown & S. O. Yamashita (Eds.). *Language Testing in Japan.* (pp. 66-75). Tokyo, Japan: JALT.

Cunningham, C. (2002). The TOEIC test and communicative competence: Do test score gains correlate with increased competence? Unpublished MA thesis: University of Birmingham.

Douglas, D. (1992). Test of English for International Communication. In J. J. Kramer & J. C. Cooley (Eds.). *The Eleventh Mental Measurements Yearbook.* Lincoln, NE: Buros Institute of Mental Measurements.

Dudley-Evans, R. & St. John, M. J. (1996). *Report on business English: A review of research and published teaching materials.* Princeton, NJ: Educational Testing Service.

Educational Testing Service. (2003). *The latest news on TOEIC products.* Princeton, NJ: Author.

Eggly, S., Musial, J., & Smulowitz, J. (1998). The relationship between English language proficiency and success as a medical resident. *English for Specific Purposes, 18* (2), 201-208.

Gilfert, S. (1996). A review of TOEIC. *The Internet TESL Journal, 2* (8). Retrieved on October 4, 2003 from http://iteslj.org/Articles/Gilfert-TOEIC.html

Gilfert, S. (1995). A comparison of TOEFL and TOEIC. In J. D. Brown & S. O. Yamashita (Eds.). *Language Testing in Japan*. (pp.76-85). Tokyo, Japan: JALT.

Hirai, M. (2002). Correlations between active skill and passive skill test scores. *Shiken: JALT Testing & Evaluation Newsletter, 6* (3), 2-8. Retrieved on October 4, 2003 from http://jalt.org/test/hir_1.htm

Jamieson, J. et al. (2000). *TOEFL 2000 framework: A working paper.* Princeton, NJ: Educational Testing Service.

McNamara, T. (2001). The challenge of speaking: research on the testing of speaking for the new TOEFL. *Shiken: JALT Testing & Evaluation Newsletter, 5* (1), 2-3. Retrieved on October 4, 2003 from http://jalt.org/test/mcn_1.htm

Perkins, L. (1987). Test of English for International Communication. In Alderson, C., K. Krahnke, & C. Stansfield, *Reviews of English language proficiency tests.* (81-83). Washington DC: TESOL.

Robb, T. & Ercanbrack, J. (1999). A study of the effect of direct test preparation on the TOEIC scores of Japanese university students. *TESL-EJ, 3* (4). Retrieved from the World Wide Web at http://www-writing.berkeley.edu/TESL-EJ/ej12/a2.html on 7 Oct. 2003.

Shohamy, E. (2001). *The Power of Tests.* Harlow: Pearson Education Limited.

Wilson, K. (1989). Relating TOEIC scores to oral proficiency interview ratings. *TOEIC Research Summaries,* (1). Princeton, NJ: Educational Testing Service.

Woodford, P. (1982). *An introduction to TOEIC : the initial validity study.* TOEIC Research Summaries. Princeton, NJ: Educational Testing Service.

HTML: http://jalt.org/test/cha_1.htm / PDF: http://jalt.org/test/PDF/Chapman1.pdf