TOEICテストを論評する

ジョージ・ナップマン *

The TOEIC – a critical review

KNAPMAN George Slade

The following report has been compiled from available documentation on the TOEIC. It begins with a discussion of the TOEIC's initial development and purpose, giving an overview of the test as well as recent major changes to it. Following this, there is a discussion of validity, and a critical review of the TOEIC, particularly with reference to construct, concurrent, and consequential validity. The limitations, as well as the strengths of the TOEIC will be examined.

1.1 BACKGROUND

The TOEIC (Test of English for International Communication) was developed in 1979 by Chauncey Group International Ltd.- a subsidiary of the USA based Educational Testing Service. ETS had previously developed the TOEFL for academic settings, but the TOEIC was developed specifically for workplace contexts, in response to a request by the Japanese Ministry of Trade and Industry. The development of the test was therefore a collaboration between t ETS and a Japanese team. The TOEIC was first administered to 2710 test takers in Japan on December 2nd 1979, and since then the number of test takers has increased dramatically. ETS claim that in the period 2004 – 2005, there were over 4.5 million test takers in 60 countries around the world, with Japan and South Korea accounting for a large proportion of these (3.9 million). The test is currently being used by government agencies, language schools, academic institutions and a large number of corporations world-wide, and it is promoted by ETS as the world's leading commercially available test of English ability for business purposes. The successful marketing of the test by ETS is no doubt one of the reasons for its popularity. In addition, good TOEIC scores often provide test-takers with increased employment opportunities or easier entry to academic institutions. This is especially true in Japan and South Korea, where the greatest number of test takers exist.

1.2 THE STANDARD TOEIC TEST

The TOEIC is a two hour norm referenced and fixed response (multiple-choice) test. It is also a proficiency test, therefore measuring English ability in terms of a future criterion (i.e. the workplace). The TOEIC consists of 4 listening sections with a total of 100 questions (45 minutes) and three reading sections with a total of 100 questions (75 minutes). Sections are

* 教養部講師

labelled from part I to part VII. Candidates receive separate scores for listening and reading, on a scale from 5 to 495 points. These provide a total score on a scale from 10 to 990 points. Candidates receive a colour coded certificate of achievement identifying score ranges. These are orange (10–215), brown (220–465), green (470–725), blue (730 – 855), and gold (860–990). According to Gilfert (1996, in Nall 2003), 'A TOEIC score of 450 is frequently considered acceptable for hiring practices... 600 is frequently considered the minimum acceptable for working overseas.'

1.2 THE NEW TEST

After over twenty years of use with no significant modifications, the TOEIC has recently undergone a number of changes. These involve a move away from discreet point items, where a single point or objective is targeted, to items that test more than one point or objective at a time. ETS state that 'Instead of many different conversations with a single question about each, there will be fewer conversations with multiple questions about each' (New TOEIC test Q& A, web page). Other changes are: more authentic English–language reading and listening tasks; a greater variety of accents (e.g. British, Australian, South African, Asian); a reduction in the number of photograph description questions in order to make room for testing longer conversations; the provision of an audio recording for all questions in each part of the listening section; in Part 3 and Part 4, candidates will also be able to read the conversations and short talks in addition to hearing them; some texts in both sections are longer; part V includes 'passage–based, sentence–completion questions'; and in part VI, contextualized passage–based questions replace error recognition. Additionally, some questions and groups of questions refer to more than one passage, measuring skills that involve connecting information across sentences.

ETS claims that these changes are supported by modern theories of communicative proficiency, which emphasize 'the simultaneous engagement of lexical, grammatical, phonetic and pragmatic language abilities' (New TOEIC test Q& A, web page). Although these changes are significant, ETS also states that the new TOEIC has been only slightly altered in structure and that it remains consistent with the old test in several key areas. Importantly, the test will remain a multiple–choice, paper–and–pencil assessment. The new test has already been introduced to Japan and Korea (December 2006) and Australia and New Zealand (September 2007). It will be progressively introduced to the rest of the world, with a minimum cross over period of two years in which both forms of the test will exist (About TOEIC, web page).

1.3 A NEW TOEIC SPEAKING AND WRITING TEST

In addition to the new standard TOEIC (listening and reading), an optional productive skills component (speaking and writing) is currently being introduced, and was offered first in Korea, in December 2006. This test is administered through the internet-based test delivery system (iBT) and will provide students with a separate score for speaking and writing. It is offered as a package, in the same way as the listening and reading test. The new speaking test lasts for 20 minutes, and consists of 11 questions, with a score range between 0 and 200. The writing test lasts 60 minutes and includes 8 questions, also with a score range between 0 and 200. ETS state that the productive skills test will provide an additional measure of test takers' English proficiency for work related environments, but they are careful to avoid any acknowledgment of limitations in the standard TOEIC – an important point for those who will continue to rely on existing scores.

2. MODELS OF LANGUAGE AND FRAMEWORKS FOR MEASUREMENT

Theories of language underpin test design (McNamara 2000, p. 10), and although recent changes to the TOEIC reflect some aspects of current language theory, the presiding underlying model of language remains a cognitive one. McNamara states: 'approaches which see the criterion performance as essentially requiring cognitive ability will be pre-disposed to use more indirect language tests, rather than those more communicative and contemporary theories of language which will emphasise social and interactional roles in meeting future criteria.'

Both McNamara and Bachmann (1990, p.82) cite Lado (1961) in explaining early theories of test performance, which were characterised by the influence of structural linguistics, and the testing of 'separate, individual points of knowledge, known as discrete point testing' (McNamara, 2000, p. 14). This "psychometric-structuralist" period of the 1950's, stressed reliability and consistency in measurement of test-takers' ability, and the multiple choice question type (MCQ) was considered most suitable for this purpose. Bachmann (1990, p. 82) further notes the typical focus on a skills components model during this period. The fact that both the standard and currently revised TOEIC (speaking and listening) remains a strictly multiple choice test, and is still divided into separate skills focused sections, clearly indicate that a cognitive and structuralist model of language remains primary.

3. CONSTRUCT VALIDITY

Construct validity relates to whether a test measures what it is supposed to measure.

Similarly, the construct can be defined as 'an ability or abilities that will be reflected in test performance; defined in terms of a theory of language' (Davies et. Al. 1999, p. 31). The TOEIC is by its own title supposed to be communicative (IC = international communication) and ETS claim a high correlation between TOEIC scores and English communicative ability. Since until recently, the test has contained neither a speaking nor a writing component, it is a claim that has been frequently debated (Cunningham 2002, p.1).

Furthermore, although the new performance test of speaking and writing does appear to more closely represent a communicative construct, it is not yet known to what degree it provides 'real world tasks in realistic contexts' (McNamara 2000, p. 6). This is especially relevant, considering the nature of iBT (internet based test delivery system). Also, it is not apparent to what degree these new tests incorporate communicative features which stress the importance of 'social roles of candidates or the demands of such roles' (McNamara 2000, p. 16). Finally, the optional role of the new speaking and writing test in the TOEIC means that the non-communicative (despite recent changes) standard TOEIC of listening and reading skills will continue to be used to categorise communicative proficiency. The following discussion therefore will concern primarily the standard TOEIC test.

Communicative proficiency cannot be measured by paper and pencil tests, especially those that use only multiple choice items. Alderson et al. (1995, p. 45) point out the obvious - multiple choice questions (MCQ) cannot test a person's pronunciation skills (a key component of communicative ability) in real life. The issue here is that while the standard TOEIC test may give a good indication of students' ability to answer a test question, it provides little guarantee that they will be able to perform in future contexts of communicative interaction (McNamara 2000, p. 29). Cunningham (2002, p.11) makes this very clear, stating that 'real-life interaction does not consist of multiple-choice options'. There is a clear difference between communicative performance tests and paper and pencil tests like TOEIC (McNamara 2000, p. 5), and it is therefore difficult to see the theoretical basis on which ETS make a claim for the TOEIC being an indicator of communicative proficiency in the workplace. The stark contrast between a cognitive model of test design and a communicatively defined construct creates weak construct validity.

4. CONTENT VALIDITY

'Content validity is the extent to which the test incorporates a representative sample of the entire domain being investigated (Hughes 1989, in Sewell 2005, p. 7). ETS claim to have taken

samples from a wide number of spoken and written sources world wide where English is used in the workplace (ETS 1998, in Sewell 2005, p.13), yet it has been stated that 'this approach cannot be validated from a theoretical standpoint and makes no guarantee about the proportionality of the TOEIC's presentation of its language features' (Moritoshi 2001, in Nall, 2003). Elsewhere however, Moritoshi (2001 p.16), claims that the TOEIC has high content validity as a measure of reading and writing skills. There appear to be two aspects to content validity that need to be clarified. The first is the relationship between content and construct, and the second is the relationship between test items and the wider sample. The TOEIC test items are generally recognized as representative of the language samples collected by ETS, but there is questionable coverage of the language domain identified by the construct.

5. CRITERION RELATED VALIDITY

5.1 Concurrent validity

Concurrent validity is determined by 'comparing results from one test format with those of another instrument which is assumed to be testing the same thing' (Nall 2003, web page). It has been claimed that the listening comprehension tests of the type used in TOEFL and TOEIC 'have been found to correlate highly with proficiency as measured by larger batteries of tests' Boyle (1987 pp. 277–288). The problem here is that correlating TOEIC scores to other measures of proficiency does not provide adequate support for a notion of general validity.

In The TOEIC Technical Manual (ETS 1998, in Sewell 2005), ETS state that 'the most common form of test validation is correlation with other, established methods'. This is a clear reference to concurrent validity, and highlights the weakness of construct validity for the TOEIC. If validity should be simply dependent on comparison with other known tests, then validity becomes a 'spiral of concurrent relatedness' (Bachman 1990, in Nall 2003) with no fixed point of reference.

ETS has carried out extensive concurrent validity testing on the TOEIC, using it to validate the test in general (Chauncey Group International, 1999 in Nall, 2003), and has found that the highest consistent correlations relate to the receptive skills of listening and reading. However, out of the eight tests chosen by ETS for correlation with TOEIC, only one productive skills test shows any significant correlation (Chauncey Group International, 1999, in Nall, 2003) and Sewell points out (2002, p.14) that the correlation within a Korean sample of test–takers on the same test was much lower. Other independent speaking tests have produced some

correlations with the TOEIC as low as 0.49 (Hirai, in Sewell 2002). Hirai further states that 'a TOEIC score is practically meaningless as a measure of writing skill'. To conclude, while there is reasonable evidence supporting the concurrent validity of the standard TOEIC alongside other measures of listening and reading skills, there is little concurrent validity in the productive skills of speaking and writing. The new TOEIC speaking and writing test will require further investigation in terms of concurrent validity.

5.2 Predictive validity

Since the TOEIC is a proficiency test, a discussion of predictive validity seems fairly important. A proficiency test is forward looking (McNamara 2000, p.70) to some 'future situation of language use,' whereas achievement tests relate to what students have 'learned as a result of teaching'. To a large extent, predictive validity has been covered in the discussion of construct validity because the trait to be measured (the construct), is conceptualized in terms of a future criterion. Accordingly, the TOEIC cannot be thought of as an accurate predictor of workplace communicative competence, but it seems a reasonable predictor of listening and reading skills (as indicated in 5.1, above).

6. CONSEQUENTIAL VALIDITY

Consequential validity is concerned with a tests impact 'on individuals, on educational systems and on society in general' (Davies et. al. 1999, p.79). This includes the test's effect on existing teaching practice (washback) and syllabus content. Washback will be affected by the consequence a test has for test takers, because a high stakes test will produce increased student demand for tuition specific to it. Cunningham (2003, p. 4) discusses how the effect of the TOEIC on teaching and on course content in Japan, is largely fuelled by the 'high stakes situation for learners', created by companies and universities that extensively adopt ETS benchmark descriptions and cut off scores without question. Students frequently try to rote learn answers to multiple choice questions, or to develop test taking strategies: "test-wiseness" (Alderson et. al. 1995, p. 45) that will help them to achieve a higher score. Taking advantage of this situation, ETS has published a huge amount of TOEIC preparation material, and there is an ongoing demand for schools to teach it. Because of this, there is some confusion about the status of the TOEIC as a proficiency test.

Discussing the TOEIC's close relative the TOEFL, Hamp-Lyons (1998, p. 332) classifies ETS preparation material in the following way:

'Because the books are built around the model of the test and because the test is not intended

to reveal or reflect a model of language in use...teacher and learners find themselves teaching–and trying to learn–discrete chunks of language rules and vocabulary items without context or even much co-text'.

7. FACE VALIDITY

Face validity is the extent to which a test is recognised publicly as being valid. Since so many companies, institutions and governments, currently accept and use the TOEIC, it could possibly be seen to have high face validity. However face validity is a highly subjective quality, as Nall points out (2003, web page), and 'can be unduly influenced by such things as effective advertising, or peer evaluations'. The subjectivity of face validity is noted by Sewell (2005, p.14), who writes about the difficulty of finding any students in Korea, who feel that the TOEIC is a good test of English, in part because students may feel the best way to get a high mark is to learn the 'tricks of the test'.

Issues of face validity may also 'jeopardize the public credibility of a test' leading to a need to produce direct tests 'out of a concern for face validity' (Davies1996, in McNamara 2000, p.106). Considering the importance of this issue for ETS, it is no doubt one of the reasons for the recent introduction of the direct speaking and writing test.

8. RELIABILITY, PRACTICALITY AND FAIRNESS

A test will be reliable if scores on one examination are consistent with scores on subsequent examinations of the same test–takers. ETS has measured the TOEIC's reliability as '0.95 for the Total score and between 0.91 and 0.93 for the Listening Comprehension and Reading Comprehension sections (Chauncey Group International 1999, in Nall 2003, web page). While the TOEIC is generally seen to be a reliable test even by its critics, it is worth mentioning that a reliable test will not be very useful if its construct validity is weak – if it does not measure what it is supposed to measure (Nall 2003, web page).

It is the TOEIC's practicality that appears to be its most positive feature. The TOEIC is available in a large number of countries, easily administered, easily and quickly scored (by machine), and gives 'reasonably fair and objective comparisons between examinees' (Nall, 2003, web page). The indirect MCQ format enables large groups of students to be tested at the same time. Increasingly, controlled testing of direct performances is becoming feasible due to ongoing improvements in internet communication technology, and this is reflected by the new iBT TOEIC speaking and writing test. However, the large number of raters required by a possible increase in demand for these tests may result in future practical difficulties.

In terms of fairness, ETS has been careful to address several issues. These relate to cultural equality, equal representation of different nationalities, and avoidance of culturally specific names or situations (Chauncey Group International Ltd., in Moritoshi, 2001 p.16). The test is therefore widely considered to be fair. However the fairness of excluding job applicants from positions on the basis of their TOEIC scores remains somewhat dubious.

9. DISCUSSION

The TOEIC has only recently undergone significant modifications, and although ETS claim that these are in line with current language theory, the continued use of the MCQ format means that a cognitive model of language remains dominant (despite more integrative test items). It seems likely that ETS will continue to use the MCQ format for practical reasons, despite the apparent weakness of construct validity represented by it. The dilemma faced by ETS as with other test designers, is that 'as a test becomes more authentic, it also becomes less practical and potentially unfair across a wide base of test takers' (McNamara '00, p. 29). The difficulty of identifying the authentic language of communicative competence is highlighted by Savignon (1983, in Bachmann 1990, p.83) who states that 'communication takes place in an infinite variety of situations'. Considering the scope of both the standard TOEIC receptive skills test and the new speaking/ writing test (with only 11 questions in the speaking section), we might justifiably be concerned that the TOEIC does not adequately represent the domains which it claims to, and that scores are potentially misleading as a basis for employment or for entry into academic institutions.

The strong wash back effect of the TOEIC, means that its purpose as a proficiency test is in question – since achievement tests rather than proficiency tests aim to measure how well students have covered the material. But English programs are under increasing pressure to teach to the TOEIC test rather than their own syllabus, leading to possible problems. For example, in a communicative curriculum, teaching to the TOEIC test will almost definitely lead to a conflict of interest, since the nature of the TOEIC is not communicative.

10. RECOMMENDATIONS AND CONCLUSION

It has been suggested that a code of practice for the TESOL profession should address the issue of ethical test preparation material and practices (Hamp-Lyons1998, p. 332), and this seems justified in the light of the tension between market interests on the one hand and educational interests on the other. Confusion about test function (proficiency or achievement), as well as conflict between test preparation materials and existing

curriculums has a potentially negative impact on teaching practice and perhaps even on society. The TOEIC may have far-reaching implications for individuals, education practice, and governments, in a large number of countries around the world, and the danger is that such high stakes tests maybe become 'devices for the institutional control of individuals' (McNamara 2000, p.4).

If the standard TOEIC was used and interpreted more narrowly, or if ETS were to redefine their construct more narrowly, then a stronger validity would be maintained. However if this were to happen, the test might then be seen as an inappropriate test of workplace English communicative proficiency, but may nevertheless be useful for other purposes, such as a measure of reading and listening skills (specifically cognitive). It is perhaps in part because of the dependence of ETS as an industry, on the number of test takers, that such a redefinition of construct has been avoided. If businesses and academic institutions can continue to be convinced of the relevance of, and need for high TOEIC scores, then large numbers of people will continue to enroll for the test. In this light, the TOEIC appears to have become somewhat of a financial juggernaut.

Despite the issues raised in this report, the TOEIC is generally seen as having high face validity (although this might depend on who you ask), and will most likely continue to be widely used because of its practical benefits and high demand. It will be interesting to see how the new speaking and writing test progresses, and whether it too is destined to become an industry standard.

References

About TOEIC . (n.d.) Retrieved September 04, 2007, from http://www.pro-match.com/toeic/toeic.htm

Alderson, JC, Clapham, CM & Wall, D (1995) Item writing and moderation, Language test construction and evaluation (pp. 40 - 60), Cambridge University Press.

Boyle, P (1987) Intelligence, Reasoning, and Language Proficiency . The Modern Language Journal, Vol. 71, No. 3. (pp. 277- 288).

Bachman, LF (1990) Communicative language ability, Fundamental considerations in language testing (pp.81-110). Oxford University Press.

Chapman, M (2003) "TOEIC® : Tried but undertested". Shiken: JALT Testing & Evaluation SIG Newsletter Vol. 7 No. 3. Autumn 2003. (pp. 2 - 5)

Chen, (Shu-Fen) (2005) Co-operative learning, multiple intelligences and Proficiency. Retrieved September 04, 2007, from http://dlibrary.acu.edu.au/digitaltheses/public/adt-acuvp120.25102006/02whole.pdf

Cunningham , CR (2002) the TOEIC test and communicative competence. Retrieved September 04, 2007, from http://www.cels.bham.ac.uk/resources/essays/cunndiss.pdf

Davies, A, Brown, A, Elder, C, Hill, K, Lumley, T &McNamara, T (1999, p. 31) dictionary of language testing. Cambridge University Press.

Douglas, D. (2000). Assessing languages for specific purposes. Cambridge, England: Cambridge University Press.

ETS press release (n.d.) Retrieved September 04, 2007, from http://www.europe.etseurope.org/fileadmin/free_resources/Europe%20website/TOEIC_sales_volumes_0906.pdf

Everyone TOEIC and Answer Guide (2006) Retrieved September 04, 2007, from http://datum.studyget.com/sh/200607/20060724_25453.shtml

Hamp-Lyons, L (1998) Ethical Test Preparation Practice: The Case of the TOEFL. TESOL Quarterly, Vol. 32, No. 2. (pp. 329-337).

Hungerlang, R (n.d.) language proficiency testing, a critical survey. Retrieved September 04, 2007, from http://arts-srv.arts.mun.ca/tesl/tesl%20canada%20national%20conference.ppt#1

McNamara, T (2000) Language Testing. Oxford University Press

Moritoshi, P (2001) The Test of English for International Communication (TOEIC): necessity, proficiency levels, test score utilisation and accuracy. The University of Birmingham, Retrieved September 04, 2007, from http://www.cels.bham.ac.uk/resources/essays/Moritoshi5.pdf

Nall, T.M (2003) TOEIC: A Discussion and Analysis. Retrieved September 04, 2007, from http://www.geocities.com/twocentseltcafe/teach/toeic.htmlNew TOEIC® Test Launch Announced for China (2007) Retrieved September 04, 2007, from http://www.ets.org/portal/site/ets/menuitem.c988ba0e5dd572bada20bc47c3921509/?vgnextoid=1e49df90d0ff2110VgnVCM10000022f95190RCRD&vgnextchannel=1b67d898c84f4010VgnVCM10000022f95190RCRD

New Toeic test Q&A, (n.d.) Retrieved September 04, 2007, from http://www.enclub.cn/html/exam/TOEFL/20070629/4957_2.html

Sewell, HD (2005)　The TOEIC: Reliability and Validity Within the Korean Context. Retrieved September 04, 2007, from http://www.cels.bham.ac.uk/resources/essays/Sewell%20Testing.pdf

Registration Opens in Japan for TOEIC® Speaking and Writing Tests, (2006) Retrieved September 04, 2007, from http://www.ets.org/portal/site/ets/menuitem.c988ba0e5dd572bada20bc47c3921509/?vgnextoid=187c45aa19a6f010VgnVCM10000022f95190RCRD&vgnextchannel=1b67d898c84f4010VgnVCM10000022f95190RCRD

Sparks, R L (2006) Native Language Predictors of Foreign Language Proficiency and Foreign Language Aptitude. Retrieved September 04, 2007, from 2006 http://findarticles.com/p/articles/mi_qa3809/is_200606/ai_n17186266/pg_3

Toeic,(2007) Retrieved September 04, 2007, from http://en.wikipedia.org/wiki/TOEIC

The New TOEIC Speaking and Writing Test at a Glance (2006) Retrieved September 04, 2007, from http://www.etscanada.ca/teachers/qa_toeic.php

(Received March 31.2008)