An Over-reliance on Discrete Item Testing in the Japanese Business Context

Mark Chapman, Hokkaido University

Introduction

The Test of English for International Communication (TOEIC) is an increasingly used English language proficiency test both in Japan and internationally. It was originally intended as a test of the communicative ability of Japanese businessmen, but is now used far more widely. Since its launch in Japan in 1979, the TOEIC has become one of the most taken English tests in the world. There are now more than 3 million candidates every year, according to Educational Testing Service (ETS), the maker of the TOEIC. The main feature of the TOEIC, according to ETS, is that it is a test of English used in business and commerce. Despite the popularity of the TOEIC it has attracted little independent research and in comparison to other standardized English proficiency tests such as TOEFL or IELTS it is relatively unknown outside of Asia (see Chapman, 2003a for a full discussion of the lack of research into TOEIC). The little independent research into the TOEIC (Douglas, 1992; Childs, 1995; Hirai, 2002) has been largely critical and unsupportive of the claims made for the test by ETS.

ETS suggest a range of uses for the TOEIC (ETS, 2003), including recruitment, promotion, selection for overseas assignments and language training courses. If TOEIC scores are used as criteria for hiring and promotion, the test is high stakes, meaning the test scores will have a significant impact on the lives of the examinees. The importance of TOEIC scores, coupled with its popularity and the lack of existing research into the test are major reasons for this study. The main purpose is to compare claims made by the test maker, ETS, and administrator in Japan, The Institute for International Business Communication, with the results of an independent study carried out at a major Japanese corporation. The main validity claims for the TOEIC are empirically based around reliability estimates and correlations with other established tests of speaking, listening, reading, and writing. The correlations act as the basis for the claims that the TOEIC is a valid test of communicative competence. The aim of this study is to independently investigate the claims made by ETS and to analyze any discrepancies arising between the two sets of results. It should be noted that the studies discussed in this paper were undertaken in the 1980's before the ascendancy of Messick's (1989) unitary validity model and later theories of test validity (Kane, 1992; Kane, Crooks and Cohen, 1999; Lynch, 2001). The current study employs a concurrent validity framework which, while admittedly unfashionable today, is necessary for

replicating the analysis of the initial TOEIC validity reports.

The TOEIC is a norm-referenced multiple-choice test of English that consists of two main sections: Listening Comprehension and Reading. There are forty-five minutes for the listening section and seventy-five minutes for the reading section. Scaled scores are reported numerically from 0 – 990.

	LISTENING COMPREHENSION	
Part I	One picture, four spoken sentences	20 items
Part II	Spoken utterances, three spoken responses	30 items
Part III	Short conversation, four printed answers	30 items
Part IV	Short talks, four printed questions and answers	20 items
	TOTAL	100 items

	READING COMPREHENSION	
Part V	Incomplete sentences	40 items
Part VI	Error recognition - underlines	20 items
Part VII	Reading comprehension - passages	40 items
	TOTAL	100 items

The first open TOEIC test was administered in Japan in December 1979 and was taken by 2,710 people and the average score was 578. Since then there has been a dramatic growth in the number of people taking the test. In 2003 there were approximately 3,400,000 examinees in 60 countries. Japan accounted for 1,423,000 candidates or 41.8% of the total (Chapman, 2004). The number of people taking the TOEIC has doubled worldwide since 1999 and in Japan alone more than 11 million candidates have sat the test since its introduction in 1979.

ETS claim that TOEIC scores “indicate how well people can communicate in English with others in the global workplace” (ETS, 2003: 4). ETS also state that TOEIC scores give an objective assessment of English as it is used in the working world and is highly reliable, with scores always being accurate and consistent. The Institute for International Business Communication (IIBC), who administer the test in Japan, claim that the TOEIC is designed to “provide information on an individual’s overall communication ability, including speaking and writing skills, by means of objective measuring of listening and speaking skills” (IIBC,

2003:7). The Chauncey Group (1999), a subsidiary of ETS, provides support for these claims, stating that “studies with large samples of non-native speakers of English from around the world have confirmed a strong link between TOEIC results and oral proficiency. Smaller studies have shown a similar link with writing skills,” (The Chauncey Group, 1999:8). The following section provides an analysis of the research into the TOEIC by both the test maker and independent researchers.

Previous research into the TOEIC

1 Research performed by ETS

The initial validity study into the TOEIC (Woodford: 1982) was based on the first administration of the TOEIC in 1979 to a group of 2710 examinees in Japan. The final reliability estimations were:

Listening Comprehension	0.92	SEM=25.95
Reading Comprehension	0.93	SEM=23.38
Total Reliability	0.96	SEM=34.93

Woodford (1982:8) claims these figures are “well within the generally accepted limits for measurement of individual achievement.” The same study reported that the correlation between the listening section and the reading section was 0.77. Woodford states that this figure of 0.77 indicates that “each score provides somewhat different information about the examinee and justifies reporting separate scores.” The mean score for all examinees in this first administration was 578 with approximately sixty-eight percent of the candidates scoring between 400 and 745.

Woodford also reported on later validation exercises carried out after the first administration of the TOEIC. These exercises attempted to measure “the degree to which performance on the TOEIC corresponded to performance on more direct measures of four language skills.” One hundred candidates at five differing levels of ability as defined by TOEIC score were selected to take four separate tests of listening, reading, writing and speaking. For details of the separate tests please refer to Woodford (1982). The correlation coefficients are given below:

TOEIC Listening + Listening Test	0.90
TOEIC Reading + Reading Test	0.79
TOEIC Listening + Speaking Test	0.83

TOEIC Reading + Writing Test	0.83

The figure of most concern would seem to be the result for the Reading section of the TOEIC and direct test of reading. The figure of 0.79 seems to be rather low, as the TOEIC is itself a direct test of reading. It is surprising that the correlation coefficients for speaking and writing are higher than that for reading. Given that the TOEIC is only an indirect measure of speaking and writing, it would be expected that these figures should be lower than those of listening and reading. The correlation coefficient of 0.79 led Woodford to claim that:

> *The high degree of similarity of performance by the examinees on both the TOEIC Reading section and the Direct Measure of Reading suggest that the TOEIC Reading Test provides a good indication of the examinee's ability to read English with understanding.*

This claim was made despite Woodford having previously stated that the correlation of 0.77 between the listening and reading section indicated that they provided different information about the examinees. There is a clear inconsistency here in how Woodford is interpreting the results of the study. It is difficult to see how the claim that the two tests of reading show a high degree of similarity of performance can be supported.

In 1989 ETS produced a report investigating to what extent the conversational ability of individuals could be inferred from their TOEIC score. The introduction to the summary of this report (Wilson, 1989: iii) addresses a concern of this paper.

> *A generic problem with norm-referenced second-language proficiency tests is that examinee's scores on the tests do not provide a direct indication of their actual levels of functional ability to use a target language as demographically comparable native speakers can be expected to use it.*

Wilson's study compared the TOEIC scores of candidates with their Language Proficiency Interview (LPI) scores and this data provides the support for the test maker's claim that TOEIC scores are a valid measure of oral proficiency. The LPI was developed by the US government and is an extensively used test to measure oral ability. It is scored on a scale of 0 to 5 with 5 being equal to an educated native speaker and 0 indicating no ability. The study consisted of 285 Japanese candidates, 56 French candidates, 42 from Mexico and 10 from Saudi Arabia. The main findings of the study are as follows: (Wilson, 1989:51)

- TOEIC Listening/LPI correlations were higher than TOEIC Reading/LPI correlations. The former correlated in the mid- .70's and the latter at 0.70.
- TOEIC total/LPI correlations were approximately the same as those for the TOEIC Listening/LPI correlations, but slightly lower in some instances.

These figures are significantly lower than those reported in the earlier 1982 ETS study, which claimed that the TOEIC Listening section correlated with a separate speaking test at 0.83. This later report (1989) indicates that a separate speaking test and the TOEIC will provide somewhat different information about examinees. However, the results of this 1989 investigation are the basis for ETS claiming that the TOEIC is a valid measure of oral proficiency. The correlation of 0.83 between the TOEIC Reading section and an independent writing test in Woodford's 1982 investigation is the basis for The Chauncey Group's claim that the TOEIC is a valid test of the ability to produce written English. These two sets of data and resulting correlation coefficients are the main concern in the current study, as to the authors' knowledge there has been no independent corroboration of the findings.

2 Independent research into the TOEIC

The most widely known independent assessment of the TOEIC remains that contained in the broad review; *English Language Proficiency Tests* (Alderson, Krahnke and Stansfield: 1987). The reviewer makes the point that the TOEIC Listening section is not a pure listening test but "is an integrative test . . . the subject reads the options in English, choosing the correct answer based on what was heard on the tape." The reviewer then goes on to quote the statistics given above from Woodford (1982), regarding reliability coefficients and the correlation with a speaking test. From the evidence the reviewer concludes that "the TOEIC is a standardized, highly reliable and valid measure of English, specifically designed to assess real-life reading and listening skills of candidates who will use English in a work context." Little issue could be taken with this conclusion; however, the reviewer also goes on to suggest that:

> *Empirical studies indicate that it is also a valid indirect measure of speaking and writing. The items assess major grammatical structures and reading skills and, in addition to being an integrative test, the TOEIC also appears to tap communicative competence in that the items require the examinee to utilize his*

or her sociolinguistic and strategic competence.

(Alderson, Krahnke and Stansfield: 1987: 82)

The reviewer refers to “empirical studies” but is quoting Woodford’s (1982) paper. This has been updated by ETS with the more comprehensive 1989 report (Wilson), which shows the TOEIC to be a less reliable predictor of spoken English than Woodford suggested.

A second review of the TOEIC (Douglas, 1992) was rather more critical. Douglas questioned the relevance of many TOEIC items, reporting that “only about 40% of the 200 items on the test can be said to be direct measures of the English language skills required in international commerce and industry.” Douglas was not supportive of all claims made by ETS about the TOEIC. Douglas’ conclusion was that the TOEIC tested a narrower range of skills than ETS claim:

> *Overall, users should be aware that the TOEIC requires primarily a knowledge of English grammar and vocabulary, with an overlay of commerce-oriented subject matter. Very little else of language knowledge—textual, illocutionary, or sociolinguistic knowledge—is being tested.*

A review of the listening section of the TOEIC was conducted by Buck (2001). Buck provides a balanced account of the merits of this part of the TOEIC; being critical of the breadth of items tested but supportive of the quality brought to a relatively narrow construct. Buck claims that the TOEIC does not attempt to assess inferences, “such as indirect speech acts, pragmatic implications or other aspects of interactive language use” (p214). TOEIC does not test discourse or sociolinguistic processing according to Buck. Given the requirements of employees functioning internationally, there are further limitations raised (p.216), “the test is not assessing many of the oral characteristics that make spoken language unique; there is very little fast speech, no phonological modification, no hesitation and no negotiation of meaning between the interlocutors.” These are aspects of spoken language that employees working overseas are likely to experience on a regular basis. While Buck is critical of the narrow construct employed by the TOEIC, he does praise the listening items as effective for the construct measured (p.216).

> *It mainly requires processing sentences on a literal semantic level, and might be best described as a test of general grammatical competence through the oral mode.*

Both Buck and Douglas were critical of the construct measured by the TOEIC, with Douglas believing it is testing grammar and vocabulary and Buck proposing an even narrower construct; simply grammatical competence.

An independent investigation into the TOEIC (Childs, 1995) examined whether it is a suitable test to measure progress in English. He investigated a group of 113 new employees in a Japanese company. The new employees underwent an intensive one-week English course followed by a half-day of English study once a month for the next four months. The new employees were given a TOEIC test before the start of the intensive course, at the end of the intensive course, after the second half-day class and finally after the last half-day English class. There were a total of four TOEIC tests administered. After analyzing the results of all the administrations of the test, Childs concluded the following:

- The TOEIC is reasonably effective at measuring overall group gains in proficiency.
- The TOEIC is not as effective at measuring the progress of individual learners in the short term. Childs states that "the use of TOEIC for gauging individual learning is, in general, inefficient or wrong." The reason for this conclusion was the standard error of the total scores was in the range of the expected individual gains. The SEM's in Childs' report are somewhat higher than those given in the TOEIC's initial validity study (Woodford, 1982). 43 points for the total score SEM on Childs' investigation, against Woodford's figure of 34.93.
- It is not possible to explain the reasons for students' progress using the TOEIC. This is again due to the standard error of the scores. Childs dismisses TOEIC's ability for use as a diagnostic test.
- The TOEIC can be an effective tool for comparing the performance of different language schools or programs. Childs qualifies this conclusion with the comment that care is needed as TOEIC is not especially effective for measuring individual gains. Presumably, if a company sends a large number of employees to different schools or programs, then the group gains can be reasonably confidently compared.

Childs' (1995:75) closing comments reported that:

> *"Company education directors and language schools should be warned that short-term TOEIC results cannot be substituted for more specific measures of*

learning achievement. Test users await a series of criterion-referenced tests complementary to the norm-referenced TOEIC."

And later:

"Education directors who incorporate TOEIC into their testing programs should do so thoughtfully. They should understand that the long-term solution to many of their needs will be not TOEIC but a series of tests that are in tune with the specific goals and methods of their English education programs."

Although Childs' study is of direct relevance to the validity and reliability of the TOEIC some words of caution are required in interpreting his findings. Firstly, the participants underwent only 53 hours of formal training. Childs points out that as much as 200 hours of formal training would be required to accurately gauge the improvement in TOEIC scores for individuals. The sample population is also different from that in Woddford's (1982) report. The mean TOEIC score in Child's study is 269, well below the mean TOEIC score in Japan of 451. Childs' sample did form a normal distribution, but all the participants were recent graduates and young employees of a Japanese manufacturing company and do not constitute a random sample. In addition, given that reliability and standard error are always sample dependent (e.g. Thompson, 2003), caution is required in interpreting Childs' comparison of his findings with those of earlier studies.

Other authors (Gilfert: 1996, Eggly et al: 1997, Robb & Ercanbrack: 1999) have also used the 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 reviews of Douglas and Buck, along with the investigation of Childs begin to raise questions about the construct measured by the TOEIC and the accuracy of the claims made by ETS. Considering how widely used the TOEIC is (over three million candidates per year), the quantity of independent data verifying the claims of the test maker is startlingly low. With a view to providing more independent data on the correlation between productive skills and TOEIC scores, the authors conducted two studies: the first compares the scores of a company-internal interview test and the TOEIC, while the second compares the scores of the Business Language Testing Service (BULATS) Writing Test, which is part of a four-test suite developed by the University of Cambridge Local Examinations Syndicate (UCLES), and the TOEIC. The basis of all validity claims about the TOEIC by ETS is the correlations with direct tests of speaking and writing. This is the reason the authors are attempting to

independently investigate these correlations and not attempting to assess the validity of the TOEIC in any broader fashion.

Materials and Method

1) Participants

In order to study the correlations between speaking skills and TOEIC scores, the authors collected the test scores of 475 students enrolled in an intensive, total immersion business English program (divided into Intermediate and Advanced levels, which students self-select) at the foreign language training center of a major Japanese company between October 1999 and September 2001. To investigate the correlation between writing skill and TOEIC scores, the British Council and the authors jointly administered the BULATS Writing Test to a total of 100 employees, 90 of whom were students of either the Intermediate or Advanced course between September and November 2001. Ten individuals, who were not students at the corporate language training center at the time, took the test in April 2001. The participants made up a homogeneous sample: they were all graduates of the Japanese education system and held at least an undergraduate degree from a Japanese university. They had six years of English as a foreign language education in school, with an emphasis on reading and translation. In the Japanese university system students typically receive three semesters of foreign language education and as almost all the participants were from a science and engineering background, it is unlikely they will have had more than this minimum level of English education at the university level. They ranged in age from 24 to 46 with an average age of 29 and had been working in the corporation for an average of six years. English proficiency varied from elementary (TOEIC 255) to advanced (TOEIC 935) with an average score of 559. Although some of the participants would have had prior experience of using English as part of their professional duties, the majority of the employees attending the foreign language training center came with little experience of communicating in English.

2) Instruments

The authors employed the following three tests:

- A company-internal interview test, which, based on a series of questions, evaluate the test taker's aural comprehension, grammatical accuracy, vocabulary, pronunciation, and fluency. The test lasts between 10 and 20 minutes and is conducted face-to-face with a native English speaking instructor.

- The BULATS Writing Test, which evaluates the test taker's ability to produce short, well-organized business-related memos/letters/e-mail with linguistic accuracy and appropriateness. The test lasts 45 minutes and consists of two tasks.
- The TOEIC, which is designed to evaluate the test taker's ability to read and aurally comprehend "real-life, business-type" English. The listening and reading parts have 100 items each, and the total test time is 120 minutes.

Each student took an interview test at the beginning and the end of the intensive course. The scores reported throughout this study are the exit test scores. The interview test starts with a warm-up, where the examiner attempts to put the examinee at ease with questions about the candidates' job and interests. The interview builds from the information supplied by the candidate and no two interviews will be identical. The examiner will pursue work-related issues with the examinees and later in the interview will challenge opinions produced to see whether examinees can defend a point of view. The exit test (relevant to this study) contains questions based only on information generated by the candidate or a topic covered in the previous intensive course. All the topics are business related unless they regard a particular interest raised by the candidate in the warm-up. As the interview tests focus on the candidates' capacity to discuss their jobs and work-related issues the authors believe the test is a valid concurrent criterion against which to check the validity of the TOEIC, which itself claims to be a test of English as used in business and commerce and how well an individual can communicate in a business environment (ETS, 2003). The interview tests are performed exclusively be the three full-time in-house instructors and intra-rater reliability was .95 or higher for all the instructors. Inter-rater reliability was consistently above .9 for the three instructors. Each student's most recent TOEIC score before the start of the program was recorded, and all Intermediate level students took the TOEIC at the end of the program.

The BULATS writing test has two parts. In part one the candidate is required to write a short reply (50 – 60 words) to a short text such as a letter, memo, or advert. In part two the task is to compose a report or letter (180 – 200 words) after receiving brief instructions. The candidate has a choice of two different prompts. Although the BULATS may not be the ideal choice of an English-writing assessment in a business context, (it does not include any lengthy or complex writing tasks and the tasks are not specific to the jobs of the employees enrolled in the intensive courses) it does test the writing skills required by the students at an intermediate level, who are the main participants in this study. The employees taking the intensive courses are generally not at an advanced level of proficiency and are unlikely to

have to write in English beyond the level of an e-mail or a short report. Hence, the BULATS is seen as a sufficient concurrent criterion against which to investigate the validity of the TOEIC as a test of written English proficiency in a business context. BULATS reliability is reported as .93 for the computer based test and .86 for the paper and pencil test (Geranpayeh, 2001). The paper and pencil test was used in the current study. For details of the differences between BULATS and the writing test employed by ETS in their 1982 study, please refer to the discussion section. The correlation was calculated between the BULATS Writing Test scores and either the actual TOEIC scores on exiting the program (Intermediate students) or the most recent TOEIC scores (all others).

Results

Figure 1 presents univariate descriptive statistics for all measures including breakdown by ability range as commented on later in this section and the discussion. These data demonstrate that the reported scores form normal distributions. The only score of concern, in terms of forming a normal distribution, for calculating correlation coefficients was that of the BULATS which had kurtosis of -0.54 and a standard error of 0.25.

Fig. 1 Univariate descriptive statistics for all measures

	TOEIC (T)	TOEIC (730 and above)	TOEIC (below 730)	TOEIC (LC)	TOEIC (R)	BULATS	Interview (Int.)	Interview (Adv.)
N	475	60	415	475	100	100	383	92
Mean	558.60	783	526.2	293.4	304	2.51	47.10	64.20
Minimum	255	730	255	145	155	1.3	22.0	50.50
Maximum	935	935	725	495	455	4.0	80.0	80.0
Maximum (possible)	990	990	725	495	495	5.0	100	100
SD	122.51	44.01	92.51	63.95	69.09	0.66	10.52	6.40
SEM	24.50	8.80	18.50	18.09	18.28	0.25	3.33	2.02
Skewness	0.45	0.96	-0.02	0.44	-0.10	0.11	0.72	0.10
Kurtosis	-0.26	0.98	-0.49	-0.24	-0.42	-0.54	0.37	-0.52

When the scores of the Intermediate course students and the Advanced course students were combined, the interview scores were found to correlate fairly well with the TOEIC Total

and Listening scores, as shown in Figures 2 and 3. The correlation coefficients of 0.78 for the Total and 0.73 for the Listening score were in line with the correlation coefficient of 0.75 between the direct speaking measure and the TOEIC Listening score as reported by ETS (Wilson, 1989). When the two groups were examined individually, the correlation coefficient between the interview score and the TOEIC Total score dropped significantly (0.49 for the Intermediate course students and 0.65 for the Advanced course students), as shown in Figures 4 and 5. Figure 4 also reveals a distribution pattern for the Intermediate course students that is markedly different from that of the entire sample.

Fig. 2 Correlation between Interview scores and TOEIC composite scores

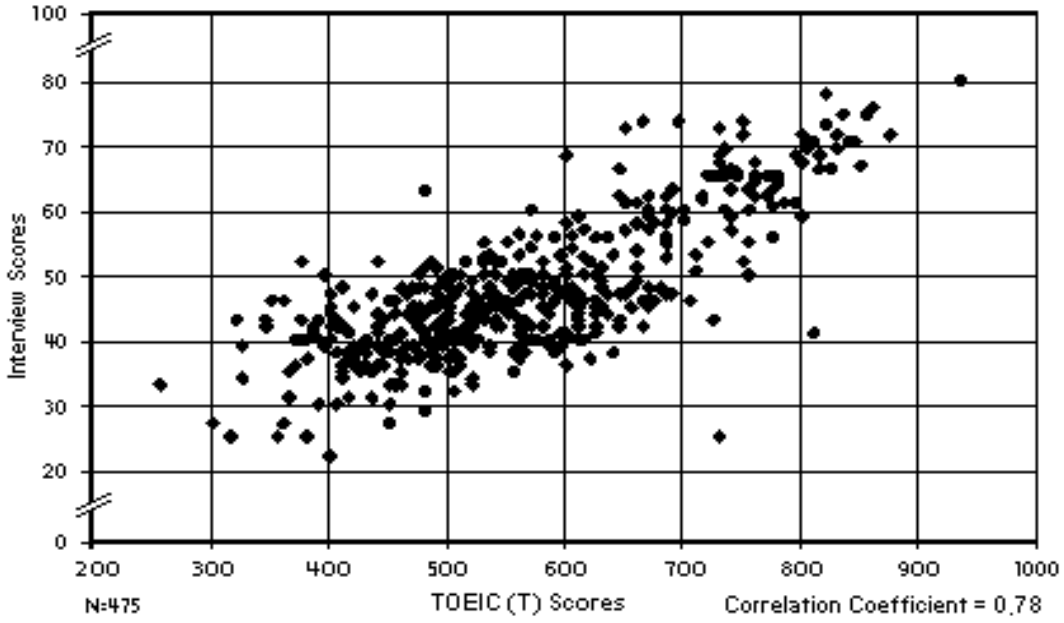


Fig. 3. Correlation between Interview scores and TOEIC (L) scores

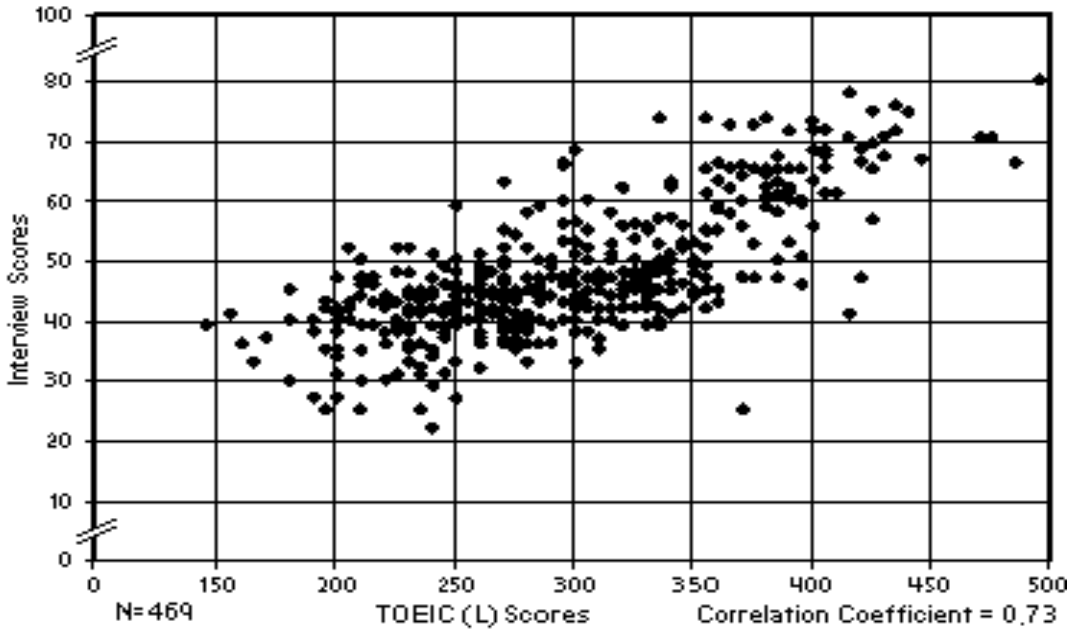
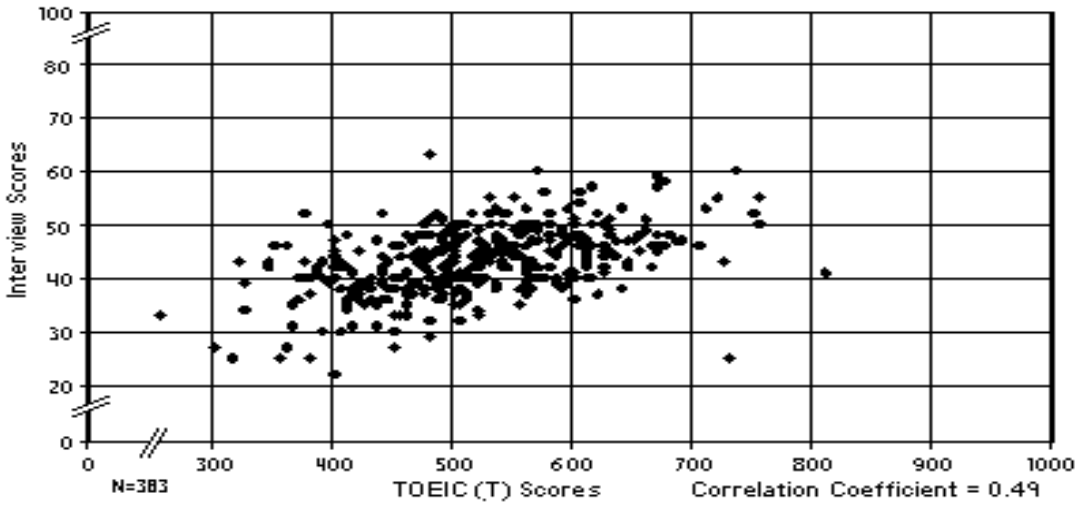


Fig. 4. Correlation between Interview scores and TOEIC (T) scores among intermediate students before starting an intensive English program



BULATS Writing scores were found to correlate a little more loosely than overall interview scores with the TOEIC Total and Reading scores, as shown in Figures5 and 6. The correlation coefficient was found to be 0.66 for the Total and 0.59 for the Reading score, both considerably lower than the correlation coefficient of 0.83 as reported by Woodford (1982, p.15) and The Chauncey Group International (1998, pp. 1-2).

Fig. 5. Correlation between Interview scores and TOEIC (T) scores among advanced students before starting an intensive English program

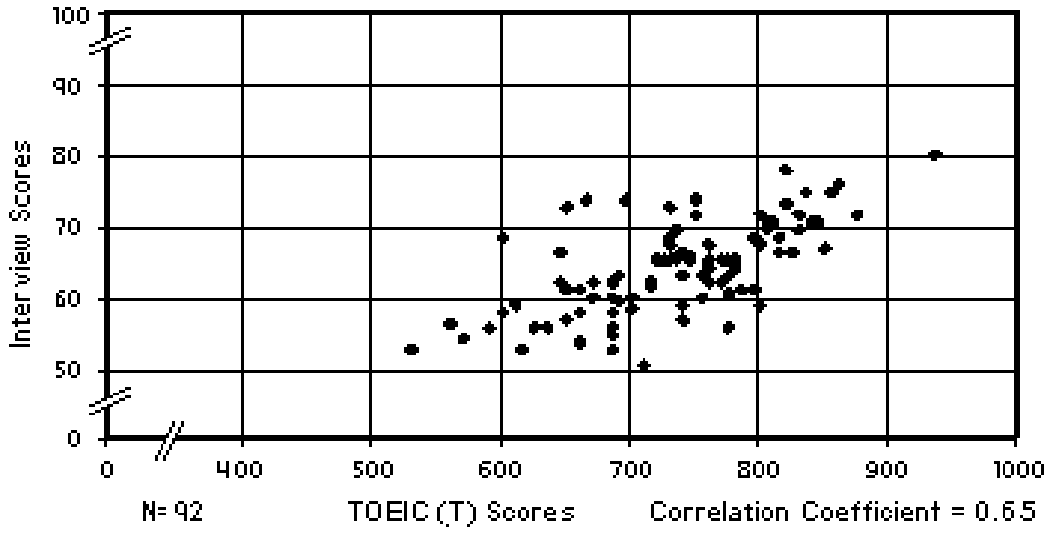


Fig. 6. Correlation between BULATS Writing levels and TOEIC (T) scores

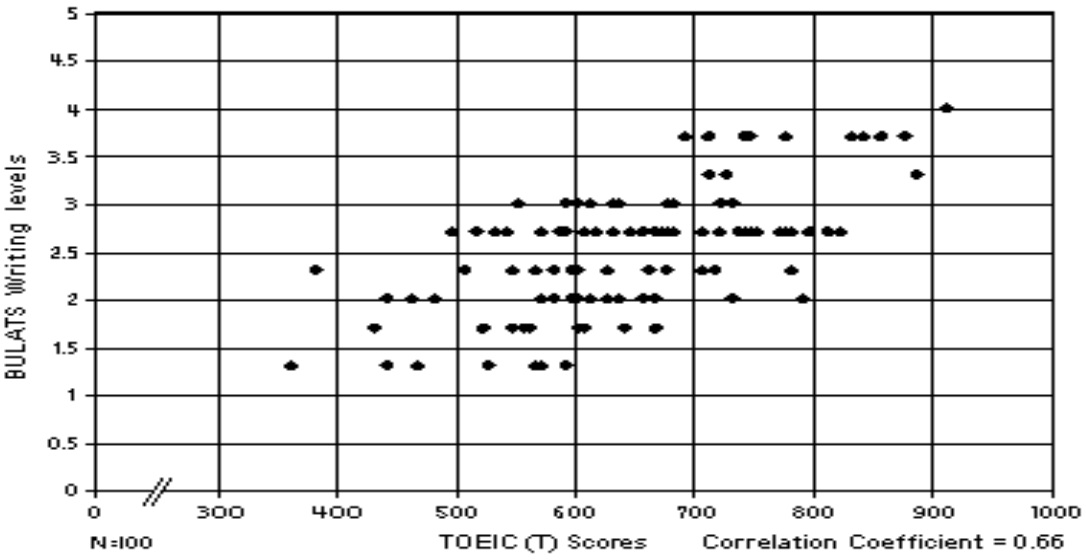
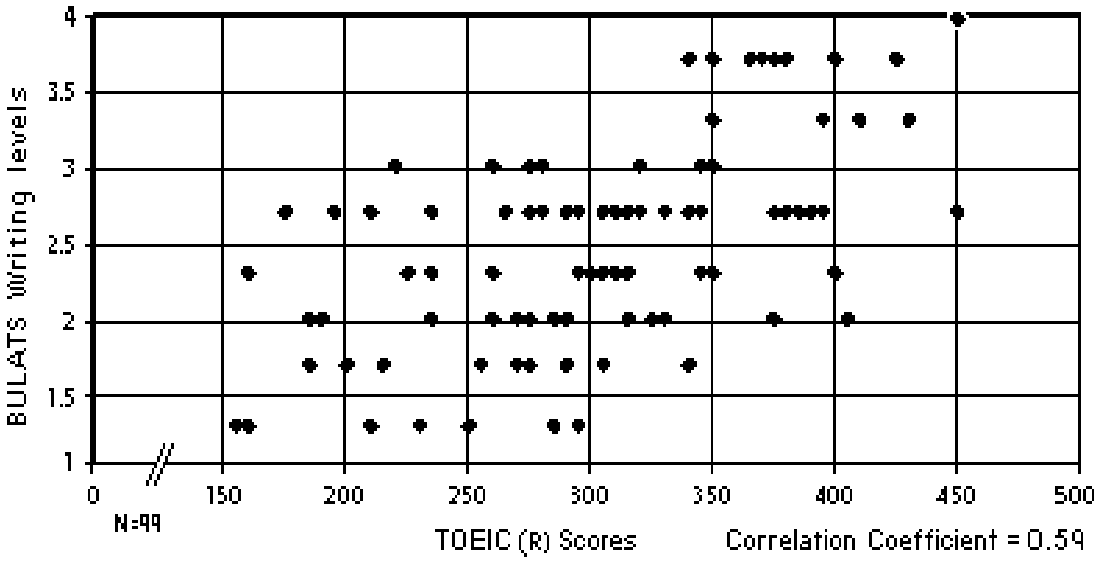


Fig. 7. Correlation between BULATS Writing levels and TOEIC Reading scores



Discussion

i) Interview Test and TOEIC Scores

With the scores of the Intermediate and Advanced course students combined, the distribution of the two sets of scores (interview and TOEIC) yielded relatively high correlation coefficients of 0.78 (for TOEIC total) and 0.73 (for TOEIC Listening). Since the subjects in this study had a wide range of general English ability (with TOEIC scores of 255 to 935), they can be considered representative of the range of English learners in Japanese business environments. Superficially, therefore, the results point to a fairly high degree of correlation between speaking skills and TOEIC scores, at least among this group of Japanese businesspeople. In terms of general applicability, however, a few words of caution are in order.

In interpreting statistical data, it is essential to check the degree of meaningfulness or reliability of the data in statistical terms, such as the sample size and the range or scope of the sample. Also, if the range of the sample does not match that of the population, then the statistical data does not accurately represent the characteristics of the population. In the present study, splitting the entire sample into two subgroups, where one group is made up of individuals with TOEIC (Total) scores of 730 or above and the other group is those with TOEIC (Total) scores of less than 730, reveals statistical features that are significantly different from those of the original sample. Both subgroups exhibit much lower correlation coefficients (0.45 and 0.63, respectively). As would be expected, restricting the range results

in lower correlation coefficients.

Likewise, it is of critical importance to check the nature of the data. Studying the data of an inherently biased sample often leads to an interpretation which is different from, or even contradictory to, that for the total population. In the present study, the Intermediate course students' interview scores flattened in the range of TOEIC 700 and above, as shown in Figure 3. This flattening effect should be attributed to the fact that the Intermediate course attracts employees with no or limited speaking experience, regardless of their TOEIC scores. The students came to the Intermediate course with an inherent bias toward poor speaking proficiency. As a result, the correlation coefficient was as low as 0.49. Note that the restricting of the sample's range to the Intermediate level was another contributing factor here. In contrast, the students of the Advanced course, which assumes experience in an intermediate-level course and/or prior exposure to an English speaking environment, exhibited a slightly higher correlation coefficient of 0.65.

This observation helps explain the apparent discrepancy between the claim of the test maker that the TOEIC is a valid test of communicative competence and the notion widely held by English educators in Japan that the test is not a reliable measure of productive skills. This observation, therefore, seems valid as far as low to intermediate-level learners are concerned, but should not be extrapolated to speak of the entire range, which has different characteristics. By the same token, the relatively high overall correlation coefficient of 0.78 should not be considered applicable to any subset of the entire range such as low to intermediate levels. In general, statistical indices are valid only within the scope being studied. The data generated in this study provide partial support for the claims made about the TOEIC by the test maker. However, in the case of test utilization, the authors can only urge caution for corporations using TOEIC scores as the sole predictor of oral proficiency where there is only a subset of the entire range of TOEIC scores.

(ii) BULATS Writing Test and TOEIC Scores

The relatively significant difference between the correlation coefficients in the present study and ETS' data can be partly attributed to the difference in nature between the two writing tests. The BULATS Writing Test consists of two tasks, each of which asks the test taker to compose from scratch a short e-mail message, letter, or memo. In contrast, the "direct measure" used by Woodford (1982, pp.10-11) has three tasks: dehydrated sentences, sentence translation, and a short (25-40 words) business letter, with weight factors of 0.3,

0.2, and 0.5, respectively. In the dehydrated sentences task, candidates are given a series of key words and asked to form a full sentence. Woodford (1982) gives the following example: employees/receive/raise/next year/ --> the employees will receive a raise next year. The second task, which is sentence translation (as opposed to passage translation), may test the basic ability to put together words in grammatically correct order but does not test the ability to compose a passage that is acceptable in a business environment. Only the third task requires creative skills, which can be acquired or improved mainly through focused training.

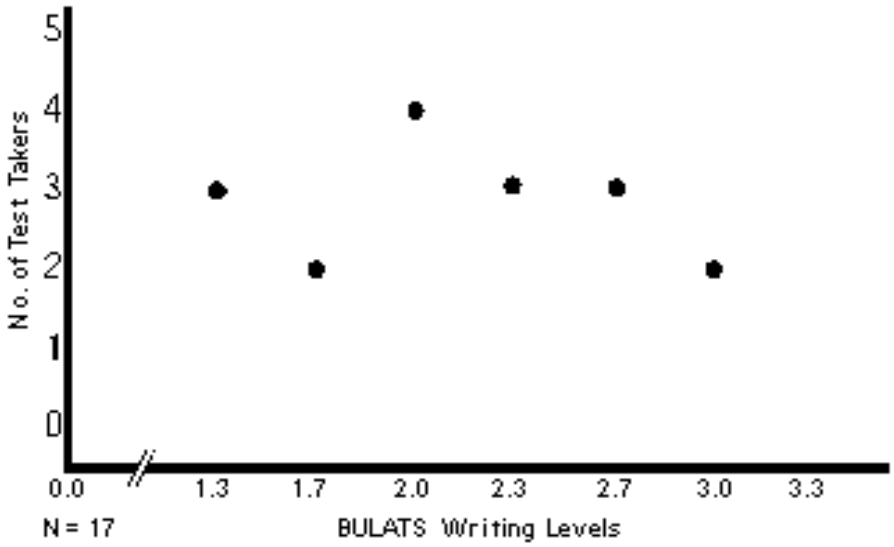
Generally, it is creative writing skill that shows the greatest variance. In fact, Woodford's table summarizing the results of the "direct measure" tests (Woodford, 1982, p.11) shows a relatively large standard deviation of 3.211 against a mean of 5.859 (on a scale of 0 to 14) for the business letter part. On the other hand, the standard deviations for the other two parts were relatively small (7.243 against a mean of 37.824 on a scale of 0 to 50 for the dehydrated sentence part and 9.406 against a mean of 64.033 on a scale of 0 to 75 for the translation part). The standard deviation is a measure of variance in value of one quantity. Therefore, if the standard deviation of one quantity is large, then the correlation coefficient between this quantity and any other quantity is in general relatively small. In Woodford's study, however, the weighting of the three components effectively smoothed out the significant differences in variance among the three test components, producing the relatively high composite correlation coefficient of 0.83.

Taking the letter writing part alone, it should be pointed out that there is a significant difference in elaborateness between the BULATS Writing Test and the direct measure test employed by Woodford. While the BULATS Writing Test gives the test taker two tasks, one between 50 and 60 words in length and the other between 180 and 200 words during 45 minutes, the letter writing part (creative part) of the direct measure test gives one task, to be completed in only 25 to 40 words in 20 minutes. It would be impractical to accurately measure the real writing skill with such a small task, and one should not draw too much significance in a business context from the ostensibly high correlation coefficient of 0.83 reported by Woodford.

While the correlation coefficient is a general indicator of how closely two quantities relate to each other, one should be cautious about the potential pitfall of predicting the value of one quantity (e.g., writing skill level) from that of the other (e.g., TOEIC score) on the basis of the correlation coefficient, unless it is extremely close to ±1. Even for a narrow range of TOEIC

scores, the writing level may vary significantly, if the correlation coefficient is not very close to ±1. For instance, in our sample with an overall correlation coefficient of 0.66, the BULATS Writing Levels of students with inclusive TOEIC scores from 550 to 595 -- one of the most populous score brackets -- spread randomly from 1.3 to 3.0 on a scale of 0 to 5.3. While one can calculate the BULATS Writing Level's standard deviation for this slice of TOEIC score continuum to be 0.56 (against a mean of 2.14), their distribution is far from a normal distribution (see figure 7 below). This fact renders the TOEIC score practically meaningless as a measure of writing skill.

Fig. 8. Distribution of BULATS Writing tests scores for those with TOEIC (T) scores between 550 – 595.



All in all, it is worth pointing out that in interpreting writing test scores, proper attention should be paid to the nature and elaborateness of the test and that, apart from the correlation coefficient, a careful look at the distribution of scores of one test for any given score bracket of the other test would be essential in grasping how well the two sets of scores relate to each other. From the above analysis, the authors maintain that TOEIC scores cannot be employed as a reliable measure of writing skills in business contexts.

Conclusion

This study has produced results that are not fully supportive of the claims made by ETS, IIBC and The Chauncey Group. There is greater support for the TOEIC as a predictor of oral proficiency than of writing skills. The correlations between TOEIC scores and oral interview tests in this study and the 1989 ETS report are similar. However, a more detailed

examination of the results shows that subsets of the entire population have considerably lower correlations. Hence, if a corporation is using TOEIC scores as the sole indicator of language proficiency and the range of ability in terms of TOEIC score is limited, there is a real possibility that the TOEIC will only provide a very rough estimate of an individual's speaking ability. In this situation it would be prudent to employ further, more direct, tests to gain an accurate view of oral proficiency.

With regard to the validity of the TOEIC as a test of written English, this study found less support for the claims made by The Chauncey Group. The correlations between TOEIC scores and an independent test of writing are significantly lower than those generated in ETS' own study. The differences between the writing test employed in this investigation (BULATS) and that used by ETS may account for some of the discrepancy. That said, the BULATS is an established test of writing in a business context produced by UCLES and hence, the fact that TOEIC scores correlate only very loosely raises doubts over the validity of the TOEIC as a measure of writing skills. Again, the results generated here indicate that employing a direct measure of writing proficiency would be preferable to relying on a single TOEIC score.

A limitation of the current study is that it compares the TOEIC with only an internal corporate test and a single test of writing. If other TOEIC users do not accept the BULATS as a valid test of writing in a business context or fail to employ in-house assessment then the criterion-related validity findings reported here are of little relevance. Other criterion tests need to be compared to the TOEIC to gain a better understanding of how TOEIC scores correlate with other independent tests of speaking and writing. The lack of research into the TOEIC is troubling in two ways. Firstly, the great popularity of the TOEIC (more than 4 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 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) The linguistic skills required by the end users of the 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.

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

These 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. Overall, the TOEIC is a high-stakes test for many employees and these people would benefit from a deeper understanding of how TOEIC scores can be reliably utilized.

References

Alderson, J., Krahnke, K. and Stansfield, C. 1987: Reviews of English Language Proficiency Tests. Washington DC: TESOL

Brown, J. and S. Yamashita (eds.) 1995: *Language Testing in Japan*. Tokyo, Japan: The Japan Association for Language Teaching.

Buck, G. 2001: *Assessing Listening.* Cambridge, UK: Cambridge University Press.

Chapman, M. 2003a: TOEIC: Tried but undertested. *Shiken: JALT Testing & Evaluation SIG Newsletter.* 7(3), 2-5. Tokyo, Japan: The Japan Association for Language Teaching.

Chapman, M. 2003b: *The Role of TOEIC in a Major Japanese Company*. 2003 JALT Pan-SIG Conference Proceedings. Tokyo, Japan: The Japan Association for Language Teaching.

Chapman, M. 2004: *Voices in the Field: An Interview with Kazuhiko Saito. Shiken: JALT Testing & Evaluation SIG Newsletter.* Tokyo, Japan: The Japan Association for Language Teaching.

Childs, M. 1995: Good and bad uses of TOEIC by Japanese companies. In Brown and Yamashita (eds.) *Language Testing in Japan*. Tokyo, Japan: The Japan Association for Language Teaching, 66-75.

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

Educational Testing Service. 2003: *TOEIC From A to Z*. Princeton, NJ: Educational Testing Service.

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.

Geranpayeh, A. 2001. *CB BULATS: Examining the reliability of a computer based test using test-retest method.* Cambridge, UK: University of Cambridge Local Examinations Syndicate

Gilfert, S. 1995: A comparison of TOEFL and TOEIC. In Brown and Yamashita (eds.) *Language Testing in Japan*. Tokyo, Japan: The Japan Association for Language Teaching, 76-85.

Gilfert, S. 1996. A review of TOEIC. *The Internet TESL Journal*, Vol. 2, No. 8. [Online]. http://iteslj.org/Articles/Gilfert-TOEIC.html

Hirai, M. 2002: Correlations between Active Skill and Passive Skill Test Scores. *Shiken: JALT Testing & Evaluation SIG Newsletter.* 6(3), 2-8.

Kane, M. 1992: An argument-based approach to validity. *Psychological Bulletin*, 112, 527-535.

Kane, M., Crooks, T., & Cohen, A. 1999: Validating measures of performance. *Educational Measurement: Issues and Practice*, 18(2), 5-17.

Lynch, B. K. 2001: Rethinking assessment from a critical perspective. *Language Testing*, 18, 351-372.

Messick, S. 1989: Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.) New York: American Council on Education and Mcmillan Publishing Company, 13-103.

Perkins, L. 1987: Test of English for International Communication. In Alderson, C., Krahnke, K. & Stansfield, C. *Reviews of English Language Proficiency Tests.* Washington DC: TESOL, 81-83.

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)*. [Online]. www-writing.berkeley.edu/TESL-EJ/ej12/a2.html

The Chauncey Group International, Ltd. 1999: *TOEIC user guide*. Princeton, NJ: The Chauncey Group International, Ltd.

The Institute for International Business Communication, 2003: *Helping Develop Global Human Resources Through TOEIC and Other Programs*. Tokyo, Japan: IIBC

Thompson, B. 2003. *Score reliability: Contemporary thinking on reliability issues*. Thousand Oaks, CA: Sage.

Wilson, K. 1989: *Enhancing the Interpretation of a Norm-Referenced Second-Language Test through Criterion Referencing: A Research Assessment of Experience in the TOEIC Testing Context. TOEIC Research Report Number 1.* Princeton, NJ: Educational Testing Service.

Wilson, K. 1993: *Relating TOEIC Scores to Oral Proficiency Interview Ratings. TOEIC*

Research Summaries Number 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.