Note to marker:

I was asked to include copies of the following e-mail messages with this assignment.

Paul --> Sonia --> Birmingham:

TS 00/01 specifies IELTS or the TOEFL tests but would it be acceptable to use the Test of English for International Communication instead? I'd prefer to use this as it ties in much better with my work at the university. Students there take it regularly, so reading material probably wouldn't be a problem and being well informed about the test would make me a bit more useful (i.e. indispensable).

Bob at Birmingham --> Sonia --> Paul:

I can't see why this need be a problem, assuming that he can otherwise fulfill the requirements of the question as set with reference to TOEIC. Could you please forward this e-mail to him, however, and ask him to attach a printed copy of it to his assignment when he submits this (under the 'cover sheet').

This is just to ensure that there is no misunderstanding on the part of anyone on the marking team, who might otherwise be tempted to assess on the basis of "hasn't answered the question".

The Test of English for International Communication: necessity, proficiency levels, test score utilisation and accuracy.

Contents

List of tables

List of appendices

1 Introduction

This paper examines the four issues raised in the original task title, but with respect to the Test of English for International Communication (hereafter 'the test' or 'the TOEIC® test') rather than the International English Language Testing System test. Firstly, whether the test is necessary in principle and/or practice is approached from three perspectives. Secondly, since it relates to the workplace context, what should the required proficiency level for any given work position be? Thirdly, do test-users apply policy relating to these required proficiency levels consistently, or do other factors intervene? If so, what are those other considerations? Finally, how accurate a test is it, not only in terms of its validity and reliability but also with respect to its fairness?

Although the author presents his position on these issues in turn, he will devote a larger proportion of the paper to the matter of accuracy because it seems necessary to demonstrate an understanding of the concepts relating to a test's accuracy and an ability to critique a particular test in these regards. Some implications for test-users and test-takers are also forwarded, with particular reference to the Japanese context, as well as some suggestions for future research in currently neglected areas.

2 Background and overview

Developed by the Educational Testing Service (ETS), the test was first administered in Japan on December 2nd, 1979 to 2,710 examinees (Woodford, 1982: 5). Currently available in 39 countries, with 1.5 million tests annually (Chauncey Group International Ltd., 2001), Korean and Japanese examinees made up 94% of the total in 1997 (Sharron, 1997: 26). Used by government agencies, language schools, academic institutions and over 4000 corporations world-wide (Chauncey Group International Ltd., 1999: 4) for a wide variety of uses (discussed in section 3.2), it is promoted as the world's leading commercially available test of English ability for business purposes.

The test uses a multiple-choice format and comprises two equally weighted sections, one for listening, the other for reading, divided into four and three parts respectively (appendix 1) and takes about two hours to administer. Examinees mark their answers on pre-printed test books which are then machine-scored, facilitating a turn-around of results in as little as 48 hours (TOEIC® Service International, 1999: 3). These results take the form of three 'equated', numerical scores: one for each section, plus the combined 'total score', which ranges from 10-990

points. Section 4 exemplifies what these scores reportedly mean in terms of actual English ability. Though it only tests listening and reading directly, it also claims to measure speaking and writing abilities indirectly (Woodford, 1982: 9-16), a claim further examined in section 6.

3 Necessity

If a test cannot be justified on the grounds that it is necessary, then there seems little point taking the time, effort or financial and material resources required to develop it. By examining three differing perspectives, those of the sceptic, the test-user and the examinee, the author argues in favour of the test's necessity, both in principle and in practice.

3.1 A sceptic's view

A sceptic might argue that the test is unnecessary, firstly in principle, because candidates' self-assessment of their English abilities is equally valid, reliable and fair and is also more cost-effective. Secondly, that it is unnecessary in practice because it is essentially self-serving in that the Chauncey Group International Limited's [1] (hereafter 'the test's managers') only motivation for providing the test is to make a profit. In support, he might say that to this end, it deliberately plays on the (particularly Japanese) desire to approach native-speaker fluency, by expressing results as numerical scores, rather than as categorical pass grades (e.g. A, B and C) or a *potentially demotivating* fail. Examinees are thus enticed to continually retake the test to improve their score, so maximising the test manager's profits. Indeed, their own research shows that 32% of Japanese examinees have taken the test three times or more (Chauncey Group International Ltd., 2000: 12).

Such criticisms however lack substance. Candidates' self-evaluation of their English abilities are generally considered insufficient because they are subjective and non-systematic (Owen, 1997: 4), i.e. potentially inaccurate and inconsistent. Hughes also notes:

> ....we have to recognise the need for a common yardstick, which tests provide, in order to make meaningful comparisons.
>
> (Hughes, 1989: 4)

In other words, tests permit the accurate (i.e. valid, reliable and fair) comparison between candidates which self-assessment does not. Further, the test's scores take a numerical form for the

[1] the subsidiary of the Educational Testing Service responsible for managing the TOEIC® test.

reason detailed in section 4. Finally, though the test manager is a for-profit organisation, much of its income is reinvested for the ongoing enhancement of the test's content (Woodford, 1997: 13), which evidences a desire to provide a quality product.

3.2 Necessity from the test-users' standpoint

The two main contexts in which the test is used are the corporate and language education settings.

3.2.1 Corporate use

Each corporation uses the test to solve its own unique problems, but the TOEIC® Steering Committee highlights its more common uses:

> ....test scores are used by corporations, language schools, government agencies, etc., for the purpose of hiring; choosing persons for overseas posts; assigning to, or promoting employees within departments where English is needed; identifying employees with sufficient English ability to benefit from overseas training programs; determining the effectiveness of English language training programs; and assigning to or determining placement within English language-training programs. A leading factor contributing in the importance of the ..... test to Japanese companies is the movement towards placing manufacturing facilities overseas.
>
> (TOEIC® Steering Committee, unknown: 4).

This wide range of uses evidences the main argument in favour of its necessity in practice: it enables businesses to solve many of their personnel issues in the cost-effective, rapid, accurate and convenient manner which they require (Chauncey Group International Ltd., 1999: 8) and to do so based on independent information which facilitates fair comparisons (Hughes, 1989: 10). It has become a powerful management tool, without which, well-informed personnel decisions might be more problematic. Its necessity is further emphasised in light of reports indicating that applicants and employees often lack the English skills which their positions demand (TOEIC® Steering Committee, unknown: 1; Chavanich, 1989, in Gelb, 1989: 1; Nakatsu, 2000).

If, as the test's managers point out, companies world-wide recognise that English language proficiency is central to their success in the international marketplace (Chauncey Group International Ltd., 1999: foreword), then they need the TOEIC® test, or at least an equivalent instrument, to assess employees' English abilities. The implication is that without it, they risk making poor personnel choices that may subsequently adversely affect corporate performance.

3.2.2 Language education

Recently, language education institutions have also started using the test, mainly for four purposes:

a). as a placement test (TOEIC® Service International, 1999: 3),
b). end-of-course assessment (Hemingway, 1999: 3),
c). pre- / post-testing, i.e. measurement of achievement or proficiency gain over a course (Geis and Fukushima, 1997: 19; Gelb, 1990a: 3) and
d). helping students to find suitable employment (Chauncey Group International Ltd., 1998a: section 2).

The author has serious reservations concerning the use of the TOEIC® test for purposes a-c above, but elaboration is beyond the scope of this paper. It is sufficient to illustrate here that such institutions perceive firstly, that a test of some kind is necessary, both in principle and practice and secondly, that the TOEIC® test fulfils this need more appropriately than other available tools.

3.3 Necessity from the test-takers' standpoint

Although the test provides motivation for students to study (Woodford, 1993: 3, 1994: 4; Chauncey Group International Ltd., 2000b) and improves job applicants' confidence during the screening process (Woodford, 1994: 4), the main reasons given by the test's managers for taking the test are that:

- [it] will enable you to:
- verify your current level of English proficiency.
- qualify for a new position and / or promotion in your company.
- enhance your professional credentials.
- monitor your progress in English.
- set your own learning goals.
- involve your employer in advancing your English ability.

(Chauncey Group International Ltd., 1996: 6)

These points help to highlight the fact that most test-takers are either company or government employees or students studying the English language (Chauncey Group International Ltd., 1998a: 4). Unsurprisingly, some of these points also pertain to improving examinees' English either for the purpose of securing employment or enhancing existing job prospects and

indeed Anacker (1993, in Woodford, 1993: 1) observes how evidencing proficiency in English can "*open doors to broader opportunities.....and for further promotion*". The TOEIC® Steering Committee (unknown: 4-5) also reports that examinees understand the test's potential in this respect:

> .....college students have become aware of the importance companies place on TOEIC test results and they now make up 40% of the [Secure Program] administration examinees, while business people account for more than 50% of the examinees.
>
> TOEIC® Steering Committee (unknown: 4-5)

Since the labour market in many countries, including Japan, is currently depressed and looks set to remain so for some time, employers can and are demanding a wider range and higher level of skills from applicants and those seeking advancement. In response to this, many job-seekers and employees seem to understand the growing necessity to provide employers with a recognised measure of their English abilities as a standard part of their *curriculum vitae*. What was once an advantageous inclusion in a resume is increasingly becoming more of a prerequisite, particularly in qualification-oriented cultures such as Japan. The implication is that those seeking employment or promotion without such qualifications in support, instead preferring to rely upon technical knowledge or experience alone, may become increasingly disadvantaged, particularly in multi-national corporations.

Thus, in response to the question "*Is B's final observation correct?*", i.e. whether the large-scale testing of candidates is necessary in principle, the author maintains that it is, both in principle and practice: from both the test-users' and examinees' perspectives it fulfils various essential roles in the business and language education contexts, which self-assessment cannot adequately fulfil.

4 Proficiency levels

The question "*What should the required proficiency level be?*" is not really a test- specific issue because it relates to test-users' predetermined standards, independent of whichever tool is used to measure candidates' proficiency. The author concedes that a meaningful, definitive answer, if possible at all, is beyond the scope of this paper because test-users come from such a very wide range of industries (Educational Testing Service, 2000: 17) and use test scores for a wide variety of purposes (TOEIC® Steering Committee, unknown: 4). However, examination of

test-related publications, particularly 'functional descriptions' (brief statements of what candidates within a particular test score range should be able to do in English) and reported 'standards charts' (which detail the test scores companies require for various types of work), has provided an insight into how companies often derive their English language proficiency levels and from this, it is possible to make a slightly more meaningful, though unsurprising generalisation about them.

In line with the view that commercial test developers should facilitate test score interpretability (Woodford, 1982: 3; Hughes, 1989: 10; Chauncey Group International Ltd., 1998b: 2), the Educational Testing Service (1997) provides an advisory proficiency scale (appendix 2). When deciding proficiency levels for a particular position, 52.5% of the 758 companies surveyed responded that they use this scale as the sole basis for setting English proficiency standards (Educational Testing Service, 2000: 10). By selecting the functional description which best describes the English abilities that a job-type demands, personnel officers can work backwards to find the corresponding test score range which they will require of applicants. For example, level B on the scale (730-860 points) describes the English abilities one might expect of someone taking an overseas assignment. For such work, Matsushita Electric Industrial Company in Japan requires employees to obtain 750 points (Makino, 1992, in Woodford A., 1992: 4) and Anam Industrial Company Limited of Korea requires 770 points (Gelb, 1990b: 4). A similar procedure can also be used with 'standards charts' which individual companies create 'in house' after consideration of their specific needs. It is this flexibility of use that argues in favour of the more meaningful numerical scores over categorical pass (A, B, C etc.) or fail grades which the sceptic's viewpoint advocated.

Unsurprisingly, the more prominent the position in terms of exposure to, and productive use of English, the higher the proficiency level employers demand. For example, translators and managers need greater competence (typically 805-990 points) than drivers or carpenters (around 205-270 points) (Gelb, 1996: 3). It could be said then that the proficiency level any job requires is directly proportional to its potential for affecting overall corporate performance, particularly in the international marketplace.

Since developing foreign language skills to high levels takes time and effort, the implication is that those seeking more prominent positions need to take a more proactive stance as early as possible in their careers, to develop their foreign language competence. However, as

the next section shows, though predetermined proficiency levels help to set standards, other considerations sometimes intervene in their consistent application.

5 Consistent application of proficiency levels

Owen puts the question:

> Does it often happen that the supposed required proficiency level is disregarded in the face of supervening economic or social considerations?
>
> (Owen, 1997: 127).

As with section 4, this issue is not test-specific, but relates to how consistently policies pertaining to proficiency levels are applied and what other considerations may influence that application.

In the transcript (Owen, 1997: 10), 'A' notes that "*[universities] can't afford to turn [students] down*". This may be particularly true for many private educational institutions worldwide, which by their very nature are not only seats of learning but also corporate enterprises of a sort. For example, in recent years educators in Japan have begun to lament the fact that, for demographic reasons, there are fewer students now than previously, resulting in increased competition among private institutions to attract students. Predictably, these institutions do sometimes lower their predetermined admission standards in order to remain financially viable. One implication of this phenomenon is that syllabus designers may have to account not only for students' lower starting levels but also for groups that may be more heterogeneous with respect to students' foreign language abilities. Another is that test designers may need to create or select tests that are more sensitive to distinctions at lower ability levels. Michigan English Placement Test scores, for example become uninterpretable below the 25% mark (Hemingway, 1999: 5), as do the TOEIC® test scores, because examinees have a good chance of guessing that many answers correctly due to the 4-way multiple-choice format.

However, the situation in other industries may be quite different. Given that, for the sake of corporate performance, employees must be capable of fulfilling their professional duties, it might be argued that companies cannot afford *not* to turn down those with inadequate English proficiency. As noted above, in the current labour market, employers can be more selective, possibly resulting in fewer sub-standard candidates being accepted. However, English proficiency, while very useful, is sometimes not the only criteria used to make personnel decisions. The Educational Testing Service's client survey (2000: 10) shows that 25.1% of

companies questioned use the ETS's advisory proficiency scale in conjunction with other considerations. What those are is not specified, but probably include professional criteria such as candidates' technical knowledge, prior career experiences, other vocational or academic qualifications and seniority and social factors such as personal development, personality, age, gender and even a lack of more linguistically proficient applicants. Some employers may also be prepared to accept sub-standard foreign language proficiency in the short-term, in favour of a more well-rounded candidate, particularly if it is felt that he/she has the potential to obtain the required proficiency level with training at a later date. The implication is of course that if such staff are employed, employers must ensure that appropriate provision exists for subsequent language training.

It seems that private academic institutions and businesses may disregard their own proficiency standards, the former mainly for financial viability, i.e. economic considerations and the latter mainly for professional and/or social considerations which balance their need for people who can use English with their need for people who can do the job. However, organisations should be cognisant of the need for doing so on a well-reasoned and principled basis. It seems more appropriate then to view foreign language proficiency levels only as a point of departure when making admissions or personnel decisions, rather than as an all-exclusive criterion.

6 Accuracy

Whether or not the test accurately measures examinees' receptive and productive English language proficiency can be evaluated by examining the three aspects of any test's accuracy: validity, reliability and fairness.

6.1 Validity

Validity is the extent to which a test measures what it intends to (Hughes, 1989: 22; Brown, 1994: 254). Of the many types noted by Owen (1997: 18), the most relevant here are construct, criterion-related (concurrent and predictive), content and face validities.

6.1.1 Construct validity

Owen (1997: 20) notes the occasional confusion between construct and content validities, so the author will attempt briefly to tackle the issue here. A construct is a conceptualisation, operational definition or description of a phenomenon, such as listening or reading. Construct

validity is the degree to which the test's content '*operationalises*' or reflects the construct as it has been described (Jafarpur, 1987: 199; Hughes, 1989: 26; Richards *et al*, 1992: 80; Brown, 1994: 256-7; Bachman and Palmer, 1996: 21). Content validity however relates to the extent to which a test's content is *proportionally representative* of all of the construct's features (Jafarpur, 1987: 200; Hughes, 1989: 22; Bachman, 1990: 306; Richards *et al*, 1992: 81). The confusion may occur because both types relate to a test's content, though in subtly different ways.

Various publications (Woodford, P., 1978: 2, 1992: 11; Suomi, 1992; Chauncey Group International Ltd., 1996: 1, 1998a: sections 3.1-3.2; 1999: 8; Hemingway, 1999: 3) promote the test as one of general English language proficiency and vaguely imply that reading, listening, speaking and writing abilities are viewed as unitary, integrated skills. However, no explicit operational definitions for these abilities, or for 'general proficiency', have been provided. Even if one accepts Hughes' (1989: 26) argument that this is unproblematic for '*common-sense constructs*' such as reading and listening when tested directly, the claim that the test also measures speaking and writing indirectly would, even Hughes (*ibid.*) concedes, require operational definitions for those constructs. The overall impression given is that the test's managers have tended to skirt around the issue of construct description, so weakening the test's construct validity.

Instead, they rely entirely upon another weaker method for demonstrating construct validity: concurrent correlational evidence. They suggest that, since scores on the test correlate highly with other direct measures of reading, listening, speaking and writing, the TOEIC® test must also be measuring those same constructs (Chauncey Group International Ltd., 1998a: section 3.1-3.2). However, this approach to evidencing construct validity is rather unpersuasive for two reasons. Firstly, correlational coefficients alone are insufficient evidence of high construct validity (Bachman, 1990: 258; Bachman and Palmer, 1996: 135). Secondly, as the following section shows, the correlational evidence suffers from numerous weaknesses.

6.1.2 Criterion-related validity

Of the two types of criterion-related validity, concurrent and predictive, the test's managers concern themselves exclusively with the former. A test has concurrent validity with another if the two measures yield consistently very similar results, expressed as a high positive correlation co-efficient. The previous section showed that concurrent correlation techniques were

used extensively to evidence the test's construct validity and the results of these are provided in tables 1 and 2 below.

Table 1 - Pearson Product Moment correlation values between the TOEIC® test listening, reading and total scores

	Listening	**Reading**	**Total**
Listening	1.000	0.822*	0.952*
Reading		1.000	0.957*
Total			1.000

* $p < 0.001$

(Chauncey Group International Ltd., 2000a: 15).

Table 1 shows good internal correlations between listening, reading and total scores at a very high level of significance, suggesting that both sections are measuring aspects of the same thing: 'general proficiency'. However, the data reported in table 2 suffer from various shortcomings, namely: though high on face validity (section 6.1.4), three of the concurrent tests were unvalidated and all were scored subjectively; the level of significance is reported for only one concurrent test, the LPI, making interpretation difficult; the sample sizes, though adequate, are not substantial; and no 'negative evidence' is offered to show that the test is *not* testing other, unrelated abilities, as advised by Bachman (1990: 259). These data should therefore be interpreted with caution.

Table 2 - Pearson Product Moment correlation values between the TOEIC® test scores and other measures of listening, reading, speaking and writing

Language skill	**Concurrent test**	***r* value**	**Level of significance (*p*)**	**Sample size (*n*)**
Listening	A custom-made, direct test of listening comprehension	.90	---	99
Reading	A custom-made, direct test of reading comprehension	.79	---	99
Speaking	Language Proficiency Interview (OPI)	.74	.01	393
Writing	A custom-made, direct test of writing	.83	---	306

(Chauncey Group International Ltd., 1998a: section 3.1-3.2)

Even setting these weaknesses aside, as Bachman explains, the approach's rationale is fundamentally flawed:

> [it] assumes that the criterion behaviour (test or other performance) can be validly interpreted as an indicator of the ability in question. Frequently evidence for the validity of the criterion itself is that it is correlated with other tests, or other indicators of the ability, which simply extends the assumption of validity to these other criteria, leading to an endless spiral of concurrent relatedness....[but] only the process of construct validation can provide this evidential basis of validity.
>
> (Bachman, 1990: 249)

In other words, the concurrent evidence presented is effectively circular, mutually supportive and neglects to make the necessary reference to descriptions of the constructs under examination.

The author concludes therefore that the test's concurrent validity is somewhat shaky. This in turn further weakens the construct validity which relies upon it. Perhaps Woodford's concern over the test's validity is justified:

> Frankly, we were less worried about reliability than we were about validity. Did our multiple-choice questions really measure the linguistic behaviours they were intended to measure ?
>
> (Woodford, 1997: 13)

The case presented here demonstrates that whether or not they do remains contentious.

One final point regarding predictive validity: Given that many corporations use the test scores as the basis for decisions about candidates' suitability for future work positions (Educational Testing Service, 2000: 5), it is surprising that the test's managers provide no evidence for its predictive validity. Research in this area would demonstrate the degree to which the test is suitable for this purpose.

6.1.3 Content validity

Defined above, the content validity for this test is based upon customer research, as the test's managers explain:

> ....needs analysis studies identified certain aspects of English usage that are commonly required in many different countries by multinational companies. TOEIC test specifications are designed to measure performance in terms of these requirements, which are now reflected in the kinds of test questions, sections, and subsections included in the TOEIC test, as well as in the context and setting of test questions.
>
> (Chauncey Group International Ltd., 1998a: section 3.3)

In other words, the needs analysis provided a relatively quick, easy, inexpensive but accurate method for identifying and selecting those features of the general 'target language use domain' (Bachman and Palmer, 1996: 44-45) necessary for inclusion in the test. This is a simple but effective solution for enhancing the content's relevance, exemplified in appendix 3. No claim is made that the test's content is proportionally representative of all the features contained within the theoretical constructs of reading or listening, indeed as previously mentioned, no operational definitions for these are given. Instead, the word "*commonly*" implies that an attempt has been made to make the content proportional with regard to the types of tasks, language content and settings actually prevailing in reality. As such, though it is difficult to evaluate the test's content validity from the theoretical perspective, it does appear high from the practical standpoint. This view is further augmented by the extensive use of unmutilated, authentic target language (Woodford, 1978: 2, 1982:4; Suomi, 1992: 17; Chauncey Group International Ltd., 1999: 4) which the literature notes is an essential component of high quality content (Hughes, 1989: 15; Brown, 1994: 271; Bachman and Palmer, 1996: 23-25; Owen, 1997: 28). It is important to note however that this only holds true for the test as a measure of reading and listening, not for writing and speaking, since these latter abilities are not actually tested.

6.1.4 Face validity

A test's face validity is the degree to which it is subjectively perceived to test what it claims to (Brown, 1994: 256; Hughes 1989: 27) and depends mainly upon the extent to which a test's topical content, task type(s) and context(s) mimic those of the corresponding real-life language use domain (Bachman, 1990: 315; Bachman and Palmer, 1996: 42). A high face validity is desirable because test-users' and examinees' perceptions of a test can beneficially or adversely affect enthusiasm for using or taking the test, motivation to perform optimally during the test and confidence in the test's scores (Brown, 1994: 256; Bachman and Palmer, 1996: 24). As Jafarpur observes:

> Being objectively valid is not enough for a test; it also requires face validity in order to function effectively in practice.
>
> (Jafarpur, 1987: 205)

The TOEIC® test probably has high face validity as a measure of listening and reading skills because the test's tasks utilise these abilities directly and overtly. The same cannot be said

however for the test as a measure of speaking and writing abilities because it does not sample these skills in any way.

To the author's knowledge, no agency has conducted research into examinees' or test-users' perceptions of the test's face validity. If such research reported a low face validity for the test as a measure of productive language skills, it might provide further incentive for the test managers to clarify their operational definitions and also to provide a non-technical explanation as to why this test of receptive abilities is also a valid measure of the productive skills.

6.2 Reliability

A test's reliability refers to its "consistency of measurement" (Bachman and Palmer, 1996: 19), and concerns two main issues. Firstly, all other variables remaining constant, would examinees re-taking the same test soon afterwards obtain identical or similar scores? The higher this estimate, expressed as a reliability coefficient of between 0-1, the more reliable the test. Secondly, given that an individual's scores may vary slightly, how large is the score band within which his/her 'true score' will fall? The smaller this estimate, expressed as a Standard Error of Measurement (SE) value, the closer test scores approach true scores.

From the test manager's reported data (table 3) the test-retest reliability for the TOEIC® test appears to be acceptably high, since anything above 0.9 is considered "*adequate for reporting and usage*" (Chauncey Group International Ltd., 1998a: section 4.2). However, these data pertain only to measurement of the receptive skills.

Table 3 - Reliability coefficients and Standard Error values

Score	Reliability coefficient	Standard Error
Listening	0.916	25.95 score units
Reading	0.930	23.38 score units
Total	0.956	34.93 score units

(Woodford, 1982: 8, 1997: 13; Chauncey Group International Ltd., 1998a: section 4.2)

As for the second issue, the SE values above can be used to calculate the range bands within which true scores fall, i.e. +/- the SE value 67% of the time, +/- 1.96 times the SE value

95% of the time and +/- 2.54 times the SE value 99% of the time. As to whether these range bands are acceptable is largely a matter of personal opinion, there seem to be no specific rules, but the author feels that these SE values are acceptable as they represent between only 5-7% of the total score range of 990 points.

As Hughes (1989: 3) points out, reliability is derived from various characteristics of the test itself and how it is scored. This test's reliability is high because: administration is well standardised (Chauncey Group International Ltd., 1999: 18-19); content is highly appropriate (section 6.1.3) and subject to continuous assessment (Woodford, 1997: 13); the multiple-choice format and machine-scoring are very objective; and scores undergo an 'equating' procedure to account for differences between test forms (Woodford, 1982: 5, 1997: 13; Chauncey Group International Ltd., 1999: 10).

Hill and Parry (1994: 68, cited by Owen, 1997: 94) point out that machine-scoring yields scores that are not normally distributed because they are not standardised and this is certainly the case with the TOEIC® test. However, given that the test's main use is as a management tool in the business context, examinees and test-users do not usually need to know where individual scores sit on a distribution curve: they are more concerned with whether or not the score in question meets the proficiency requirements. The issue may however cause problems for applications in the language education context.

Despite the author's and others' distrust of the multiple-choice format on numerous grounds (Oller, 1979: 233; Hughes, 1989: 60-62), it does have some advantages. Its objectivity greatly enhances reliability (Owen, 1997: 44) and the ease of marking facilitates swift result returns. There is also tenuous evidence to suggest that one of the chief criticisms of the format, i.e. that candidates can guess the correct answer, does not significantly affect scores (Woodford, 1992: 13).

6.3 Fairness

While a test's design may be valid and reliable in all the above respects, it will still be inaccurate if it favours some examinees over others in some way. Clearly a universally fair test is desirable but inequalities can be introduced through, among other things, topical content or settings which are biased with respect to occupation, personal interest(s) or knowledge, culture or nationality and a predominant use of one type of English over others.

The test's managers note that, in line with its purpose of testing general English language proficiency, the TOEIC® test does not require candidates to have specialised or technical knowledge (Chauncey Group International Ltd., 1999: 4) and appendix 3 seems to support this. Suomi (1992: 17) adds that the test uses country non-specific pictures in section 1 to avoid national bias and it does also seem to use a neutral form of English. As the test's managers explain, this high level of fairness is not simply fortuitous:

> Every effort is made to ensure that the test is unbiased and culturally relevant to our many test-takers world-wide. The TOEIC test development team is very careful to:
>
> - Avoid language that is specific to U.S. English.....
> - Avoid contexts that may be specific to one culture, or that may be foreign to test-takers from some cultures
> - Ensure the balanced use of names from different nationalities
> - Avoid the use of locations, people, or events that would be known in only certain regions or countries
> - Avoid situations that are too specific to one occupational area
> - Ensure that different cultures are adequately represented
>
> The finished test undergoes a stringent "fairness review", in collaboration with outside reviewers to be certain that all items are appropriate for use on a global basis."
>
> (Chauncey Group International Ltd., 1998a: section 2.1)

In other words they take the proactive approach to test fairness advocated by Bachman and Palmer (1996: 127).

Though a perfectly fair test, equal in every respect, is probably an unattainable ideal, particularly when testing on this scale, these extensive measures make the test as fair as possible and it is difficult to conceive how more might be reasonably expected.

7 Conclusion

This paper has examined the TOEIC® test with respect to its necessity and accuracy and the wider issues of proficiency levels and their consistent application. The author found that enough cogent literature exists to demonstrate the test's necessity, both in principle and practice, as it services the needs of employers and employees of numerous (particularly multi-national) corporations and those of educational institutions.

Given the scope of this paper and the very wide-range of settings in which the test is used, it was impossible to give a definitive answer to the question of what the required proficiency levels should be. However, insight was gained into how many organisations derive the

proficiency levels best suited to their unique personnel requirements and a statement was made regarding the general nature of such proficiency charts.

Regarding the consistent application of an organisation's proficiency levels, a distinction was noted between educational institutions and business enterprises. While the former do often seem to disregard predetermined proficiency levels chiefly for economical considerations, the latter, if they do so at all, do so usually for other, very different mitigating professional and/or social considerations.

Finally, a lack of operational definitions and 'negative evidence' largely invalidates the test manager's claims of high construct and concurrent validities. Though content and face validities for the test as a measure of reading and listening abilities are high, the same cannot be said for writing and speaking skills. Its claims as an indirect but valid measure of the productive language skills therefore has little proven basis and is highly dubious. Also, no claim was made for, or evidence found in support of its predictive validity. However, persuasive cases exist to demonstrate the test's high reliability and fairness, but only with regard to measurement of the receptive language skills.

Acknowledgements

The author is indebted to the staff of the Institute for International Business Communication (IIBC) in Tokyo for their assistance in providing publications that proved invaluable in the production of this paper. Thanks also are due to the Educational Testing Service for permission to reproduce the TOEIC® published materials contained within this paper.

Appendix 1 The test's format

Section I Listening Comprehension (Total 100 items in 45 minutes) Each item is delivered by audio cassette just one time			
Part I	Photograph	20 items (4-choice)	Statements are not written in the test book.
Part II	Question - Response	30 items (3-choice)	Q-R are not written in the test book.
Part III	Short Conversations	30 items (4-choice)	Questions are written in the test book.
Part IV	Short Talks	20 items (4-choice)	Questions are written in the test book.
Section II Reading Comprehension (Total 100 items in 75 minutes)			
Part V	Incomplete Sentences	40 items (4-choice)	
Part VI	Error Recognition	20 items (4-choice)	
Part VII	Reading Comprehension	40 items (4-choice)	

(TOEIC® Steering Committee, unknown: 2)

Appendix 2 The Educational Testing Service's advisory proficiency chart

Scanned images from the original TOEIC documentation.

(Educational Testing Service, May 1st, 1997 revision).
Reprinted by permission of Educational Testing Service, the copyright owner.

Appendix 3 Sample test questions

Scanned images from the original TOEIC documentation.

(Chauncey Group International Ltd., 1996: 15-34).

9 References

Anacker, M. (1993) 'Excellence is no accident'. In Woodford, A. (ed.) *The Reporter: TOEIC® News International, No. 11.* Chauncey Group International Ltd.

Bachman, L. (1990) *Fundamental Considerations in Language Testing.* OUP.

Bachman L. & A.Palmer (1996) *Language Testing in Practice.* OUP.

Brown, H.D. (1994) *Principles of Language Learning and Teaching* (3rd ed.) Englewood Cliffs, New Jersey. Prentice Hall.

Chauncey Group International Ltd. (1996) *TOEIC® Examinee Handbook.* Chauncey Group International Ltd.

Chauncey Group International Ltd. (1998a) *TOEIC® Technical Manual*.

Chauncey Group International Ltd. (1998b) *TOEIC® Can-Do Guide: Linking TOEIC® Scores to Activities Performed Using English*. Chauncey Group International Ltd.

Chauncey Group International Ltd., (1999) *TOEIC® User Guide*. Chauncey Group International.

Chauncey Group International Ltd. (2000a) *TOEIC® Report on Test-Takers World-wide 1997-1998*. Chauncey Group International Ltd.

Chauncey Group International Ltd. (2000b) What managers and HR specialists are saying. (www) http://toeic.com/oepages/index.htm (21st April, 2001).

Chauncey Group International Ltd. (2001) Test-takers profiles (www) http://toeic.com/testtakerspages/default.htm (21st April, 2001).

Chavanich, K. (1989) 'TAI's recruiting is "Smooth as Silk'. In Gelb, J. (ed.) *The Reporter: TOEIC® News International, No.3* Chauncey Group International Ltd.

Educational Testing Service (1997) *Proficiency Scale.* (May 1st Revision). Educational Testing Service.

Educational Testing Service, (2000) *TOEIC® Client Survey Report's official translation.* Educational Testing Service.

Geis, K. and C. Fukushima (1997) 'Overview of a study abroad course.' *The Language Teacher 21/11: 15-20.*

Gelb, J. (ed.) (1989) *The Reporter: TOEIC® News International, No.3* Chauncey Group International Ltd.

Gelb , J. (ed.) (1990a) *The Reporter: TOEIC® News International, No.4* Chauncey Group International Ltd.

Gelb, J. (1990b) *The Reporter: TOEIC® News International, No. 5.* Chauncey Group International Ltd.

Gelb, J. (1996) *The Reporter: TOEIC® News International, No. 21.* Chauncey Group International Ltd.

Hemingway, M.A. (1999) *English Proficiency Tests: A Comparative Study*. Chauncey Group International Ltd.

Hughes, A. (1989) *Testing for Language Teachers.* CUP.

Makino, S. (1992) 'The Matsushita GOLD Program: English Language Training for the times.' In Woodford, A. *The Reporter: TOEIC® News International, No. 9.* Chauncey Group International Ltd.

Nakatsu, Y. (2000) 'Not good at English but it's necessary.' *Yomiyuri Newspaper.* (Saturday, May 13th, 2000 edition).

Oller, J.W. (1979) *Language Test at School.* Longman.

Owen, C. *et al* (1997) *Testing.* Centre for English Language Studies, Birmingham University.

Richards, J.C. *et al.* (1992) *Dictionary of Language Teaching & Applied Linguistics*. Longman.

Sharron, R.H. (1997) 'TOEIC® today and TOEIC® tomorrow.' In TOEIC® Steering Committee (ed.) *The 58th TOEIC® Seminar*. Educational Testing Service.

Suomi, B. (1992) 'TOEIC® Test Development'. In TOEIC® Steering Committee (ed.) *The 35th TOEIC® Seminar*. Educational Testing Service.

TOEIC® Service International (1999) *The Reporter: TOEIC® News International, No.28* Chauncey Group International Ltd.

TOEIC® Steering Committee. (unknown) *TOEIC® History & Status*. Educational Testing Service

Woodford, A. (1992) *The Reporter: TOEIC® News International, No. 9.* Chauncey Group International Ltd.

Woodford, A. (ed.) (1993) *The Reporter: TOEIC® News International, No. 11*. Chauncey Group International Ltd.

Woodford, A. (ed.) (1994) *The Reporter: TOEIC® News International, No. 15*. Chauncey Group International Ltd.

Woodford, P. (1978) *Script of Presentation entitled 'Test of English for International Communication (TOEIC®)*. Educational Testing Service.

Woodford, P. (1982) *TOEIC® Research Summaries - An Introduction to TOEIC®: The Initial Validity Study*. Educational Testing Service.

Woodford, P. (1992) 'A historical overview of TOEIC® and its mission.' In TOEIC® Steering Committee (ed.) *The 35th TOEIC® Seminar*. Educational Testing Service.

Woodford, P. (1997) 'A historical overview of TOEIC® and its mission.' In TOEIC® Steering Committee (ed.) *The 58th TOEIC® Seminar*. Educational Testing Service.