The TOEIC: Reliability and Validity Within the Korean Context

'Testing'

Question TS/05/05:

Describe an English language test with which you are familiar and discuss how valid and reliable the test appears to be. (If possible, include illustrative examples from the test itself.) Describe any procedures you would use to establish its validity and reliability. (You should not carry out these procedures unless they are quick and simple to complete.)

H. Douglas Sewell

September 15th, 2005

Words: 4439

The TOEIC: Reliability and Validity Within the Korean Context

Section 1: Introduction

1 Introduction 3

Section 2: Test Considerations

2.1 Reliability 3
2.1.1 Factors Affecting Reliability 3
2.1.2 Estimating Reliability 4
2.1.3 Acceptable Reliability Coefficients and True Score 5
2.2 Validity 5
2.2.1 Construct Validity 6
2.2.2 Content Validity 6
2.2.3 Criterion-Related Validity 7
2.2.4 Face Validity 7
2.3 Reliability-Construct Validity Tensions 8
2.4 Backwash and Test Wiseness 8

Section 3: The TOEIC

3.0 The TOEIC 9
3.1 TOEIC Test Preparation 10

Section 4: The TOEIC and Reliability in Korean Context

4.1 Reliability in Terms of Test Facets and Random Factors 10
4.2 Reliability in Terms of Personal Characteristics 11
4.3 Establishing and Reestablishing TOEIC's Reliability 12

Section 5: The TOEIC and Validity in The Korean Context

5 The TOEIC and Validity in The Korean Context 12
5.1 Construct and Content Validity 13
5.2 Criterion-Related Validity 13
5.3 Face Validity 14
5.4 Establishing and Reestablishing TOEIC's Test Validity 15

Section 6: Conclusion

6 Conclusion 16

References 17

Appendix - Sample TOEIC Test 18

1 Introduction

Upon arriving in Korea, one thing many English teachers find surprising is the large number of high stakes tests, in particular English tests, which are required by educational, governmental or corporate entities. One of the most common such tests, and one I have become deeply involved with as a preparation course text book writer and part time preparation course teacher is the TOEIC produced by Educational Testing Service (ETS).

Upon first learning of this test, my initial impression was positive as I felt using a test was a quite equitable means of making decisions about people. However, as I discuss in this paper, I have come to suspect some aspects of the TOEIC's reliability and validity, especially in light of the ubiquitous TOEIC test preparation courses available in Korea. In light of these concerns, I will also consider in this paper some ways in which the TOEIC's validity and reliability could be reestablished in light of the challenges facing it in these areas.

2 Test Considerations

2.1 Reliability

In discussing the same candidate writing the same test at a different time, Hughes notes, "the more similar the scores would have been, the more reliable the test is said to be" (1989:29). While this view of reliability issues being the result of random measurement error is quite common (e.g. ETS 1998a:VI-2, Lewis 1999), Bachman uses a broader definition encompassing not only “how much of an individual's test performance is due to measurement error" (1990:160), but also how much is due "to factors other than the language ability we want to measure” (1990:160).

2.1.1 Factors Affecting Reliability

As diagramed below in Table 2.1.1, Bachman (1990:165) discusses three factors that may affect test reliability. The first are Test Method Facets, which are broken down into the five categories of testing environment, test rubric (organization), input (format of and nature of), expected response (format of, nature of, and restrictions on), and the relationship between input and expected response (reciprocal, non-reciprocal, adaptive) (Bachman 1990:119). The second are

Personal Attributes and include aspects such as age, gender, cognitive style and, background (Bachman 1990:164). The third are Random Factors and can include aspects such as tiredness and a candidate's emotional condition, as well as random differences in the testing environment (Bachman 1990:164).

Table 2.1.1 Factors Affecting Test Reliability (Adapted from Bachman 1990:165)

Major Factor	Some Specific Concerns
Test Method Facets	Testing environment Test rubric Input Expected response Relationship between input and expected response
Personal Attributes	Age Gender Cognitive style Background
Random Factors	Tiredness Emotional condition Random differences in the testing environment

While Bachman notes that unlike the third category random factors, the first two categories are systematic, test method facets in that every candidate is exposed to the same facets, and personal attributes in that each candidate will be effected by the same attributes regularly (1990:164). However it should also be noted that some personal attributes such as the background aspect of test wiseness and acquired test taking strategies may be more subject to modification than some other personal attributes.

2.1.2 Estimating Reliability

One way of investigating reliability is through Classical True Score measurement theory (CTS) (Bachman 1990:167). CTS suggests that an observed test score is the sum of the candidate's real score and error score, and is based on the premise that sources of test error are both random and uncorrelated with the candidate's real score (Bachman 1990:167). The implication of this being that if it were possible for an individual to take the same test without interference a sufficiently large enough number of times, the mean of their observed scores would approach their true score.

CTS allows for the investigating of reliability and calculation of reliability coefficients in three domains. The first is internal consistency and "is concerned with how consistent test takers' performances on different parts of the test are with each other" (Bachman 1990:172). Internal consistency is especially valuable as it may allow for estimates of overall test reliability in certain situations through a 'split half' methodology in which a single group of candidates take a single test only once (Bachman 1990:173). The second reliability coefficient domain is test score stability over time (Bachman 1990:181), while the third is equivalence of different test forms (Bachman 1990:182). While CTS seems more in line with Hughes' (1989:29) tighter definition of reliability than Bachman's (1990:160) more general one, a related but more involved theory, Generalizability Theory, has provisions for looking at what Bachman considers non-random factors such as test method facets and personal attributes (Bachman 1990:188).

2.1.3 Acceptable Reliability Coefficients and True Score

Reliability coefficients can range between zero and one, with a higher coefficient indicating greater reliability (Hughes 1989:31). Lado (1961 in Hughes 1989:32) suggests that good grammar, vocabulary and reading tests can have reliability coefficients in the 0.90-0.99 range, while listening tests can be in the 0.80-0.89 range and speaking tests in the 0.70-0.79 range. Hughes however, suggests that if appropriate steps are taken, speaking and writing tests can have reliability coefficients as high as 0.9 (1989:87).

While reliability estimates are important, they do not give direct information about the accuracy of individual test scores (Bachman 1990:197). As elaborated by Bachman (1990:170-1), the range of a candidate's true score, the standard error of measurement, can be estimated by using a test's reliability coefficient in conjunction with the variance in test scores among a large enough population of test takers. This standard error of measurement could then be reported as a band score to a certain confidence level (Bachman 1990:200).

2.2 Validity

In defining validity, Hughes states, "a test is said to be valid if it measures accurately what it is intended to measure" (1989:22), while Bachman on validity notes,

> In examining validity, we look beyond the reliability of the test scores themselves, and consider the relationships between test performance and other types of performance in other contexts. (1990:236)

Bachman further quotes the standards of educational and psychological testing as in part saying, "validity always refers to the degree to which the evidence supports the inferences that are made from the scores" (APA 1985:9 in Bachman 1990:237).

Bachman goes on to point out that reliability and validity are not independent entities, but are complementary in that reliability is a necessary condition for validity (1990:160). Thus, a certain grammar test may be highly reliable, yet not considered a valid test of spoken English, while a poorly assessed spoken test may have low reliability but may otherwise be considered valid for its purpose.

2.2.1 Construct Validity

A construct is a definition of an ability that permits one to make specific hypothesis about how that ability interacts with other abilities (constructs), and how that ability is manifested in observed behaviour (Bachman 1990:255). Demonstrating construct validity therefore is about showing that "the test score reflects the area(s) of language ability we want to measure, and very little else" (Bachman and Palmer 1996:21), or more simply demonstrating construct validity means showing a suitable relationship between what we are testing and what we wish to assess.

Hughes discusses "gross constructs" such as reading ability and writing ability as well as more specific ones such as control of punctuation and sensitivity to demands on style (1989:26), with presumably some such specific constructs as these underlying certain gross constructs. Constructs and demonstrating construct validity therefore requires investigating items and interactions among items on numerous levels. It thus seems that the more direct a test is, the easier it should be to demonstrate construct validity, and it seems in part for this reason that Hughes suggests that direct testing, when possible, is recommended (1989:27).

2.2.2 Content Validity

It is generally impractical, if not impossible, to test all of a candidate's knowledge in a certain domain, and as such a test must necessarily sample from the domain and extrapolate the results as necessary. Content validity is the extent to which the test incorporates a representative sample of the entire domain being investigated (Hughes 1989:22), and is evaluated by comparing the test specifications to the test contents (Hughes 1989:22).

For a grammar test this would mean that items were selected from the full range of grammar points specified, and presumably that if different test forms were available, the entire range of grammar points would eventually be tested. Unfortunately, as Hughes notes, "too often the content of tests is determined by what is easy to test rather than what is important to test" (1989:23), the result of which may lead students and educators to focus on only specific parts of the domain in question.

2.2.3 Criterion-Related Validity

The first aspect of criterion-related validity is concurrent validity, which is the extent to which the results of the test in question agree with another independent, highly dependable second assessment method (Hughes 1989:23). The relationship between the scores can be calculated and expressed as a validity coefficient ranging, as with reliability coefficient, from zero to one with a higher coefficient in this case indicating greater concurrent validity. Hughes (1989:24) notes that the acceptable level of agreement will vary depending on the situation, and perhaps being as low as 0.7 for a relatively low stakes test.

The second aspect, predictive validity, considers how well a test is able to predict a future result (Hughes 1989:25), such as employment or educational success. Hughes notes a problem here is in determining what constitutes future success or failure considering the myriad of other factors that affect the outcomes, and suggests that a validity coefficient of 0.4 is the most that can be expected (1989:26).

2.2.4 Face Validity

Face validity is viewed as the extent to which a test appears on the surface to accurately assess what it is meant to assess (Hughes 1989:27). Problems with face validity could result in a test not being accepted or used (Hughes 1989:27), and would possibly interact with a candidate's performance on the test. To avoid issues with face validity, Hughes (1989:27) notes the need for novel testing methods, in particular indirect measures to be introduced slowly and with sufficient explanation.

2.3 Reliability-Construct Validity Tensions

While reliability is a necessary condition for validity (Bachman 1990:160), both Bachman and Palmer (1996:23) and Hughes (1989:42) note there is also a tension between them. This could be exemplified in a situation where a desire to have a highly reliable test of writing could result in the use of a multiple choice error recognition test as a measure, however this test would likely be considered to have less construct validity than a less reliable but more direct test of writing ability.

Reliability and validity are not simple have/don't have conditions, but instead an equation in which it is desirable, but often impossible to maximise both. This is reflected in Bachman and Palmer's concept of usefulness in which a test's usefulness is defined as:

> Usefulness = Reliability + Construct Validity + Authenticity +
> Interactiveness + Impact + Practicality (1996:18)

While a full discussion of this is beyond the scope of this paper, the concept of usefulness and the tension between reliability and validity must be considered when looking at testing decisions.

2.4 Backwash and Test Wiseness

Backwash is the term used to describe how knowledge of a test's characteristics affect the nature of any learning preceding the test (Hughes 1989:1), and depending on its form, may be considered positive or negative (Hughes 1989:1). While backwash may influence what language content and skills students are exposed to, it may also allow for the teaching of test specific strategies and techniques that allow students to achieve higher test marks without improving their overall English ability.

Test wiseness is Bachman's term for those personal characteristics that a candidate develops to assist them in writing a test, and includes aspects such as guessing strategies and test pacing (Bachman 1990:114). As Bachman considers test wiseness a personal characteristic (1990:114), and also considers personal characteristics a reliability consideration (1990:164), this paper will consider test wiseness in relation to reliability concerns.

3.0 The TOEIC

Originally developed by Educational Testing Service (ETS) in 1979 at the request of the Japanese Ministry of International Trade and Industry (Gilfert 1996), the TOEIC is a 200 question norm referenced multiple choice listening and reading comprehension test with a focus on business communication. As diagrammed in Table 1 below and exemplified in Appendix 1, the TOEIC consists of 4 listening sections with a total of 100 questions in 45 minutes and three reading sections with a total of 100 questions in 75 minutes. Results are reported as test form equated scores ranging from 5-495 on both sections with an overall score ranging from 10-990 also reported (ETS 1998a:II-3). To help in the interpretation of these scores and introduce an element of criterion referencing to the test, ETS has attempted to correlate test scores to a matrix of language abilities expected for candidates with certain scores through a document entitled TOEIC Can-Do Guide (ETS 1998b).

Table 3.0 TOEIC Test Structure

Listening Section 45 Minutes	Reading Section 75 Minutes
Photograph Selection - 20 Items	Incomplete Sentences - 40 Items
Question-Response - 30 Items	Sentence Error Recognition - 20 Items
Short Conversations - 30 Items	Reading Comprehension - 40 Items
Short Talks - 20 Items	

The TOEIC is written by over 2 million candidates yearly, 72% of whom are Japanese and 15% of whom are Korean (ETS 2004:3). In the 2002-3 testing period an incredible 99% of the Korean candidates claim to have previously written the test, with a international total of 37.8 percent having taken the test four or more times (ETS 2004:10). Although no mention of it is made in the source material (ETS 2004), too few first time test takers in Korea suggests that many Koreans who indicated they previously took the test, may have done so informally. Also of interest is that while first time test takers have a mean performance of 624, repeaters have mean scores rising continually from 399 for first time repeaters to 525 for four plus time repeaters (ETS 2004:10).

3.1 TOEIC Test Preparation

As an acceptable test mark, often in the 700 to 800 out of 990 range in Korea is required before a job application will even be considered, there is a strong desire among many Koreans to achieve high TOEIC scores, a desire which is the basis for a large publishing and preparation course industry. I would suggest that this industry is far more developed than ETS would care to admit, as exemplified by the paid test takers I have met who regularly sit TOEIC exams to record the listening sections on MP3 players and memorize test questions. This authentic test material then forms the basis of books and courses aimed at helping candidates achieve high scores.

While the TOEIC purports by definition to be a test of communicative ability, such preparation books and courses do not generally teach English in any communicative or interactive way. Instead they focus on having students memorize structures and vocabulary items commonly found on different parts of the test, memorize grammar rules to dissect parts five and six, and learn where to find the answers to listening and reading passages without having to listen to or read the entire passage. Additionally, such books and courses teach a multitude of general test taking strategies, and provide numerous practice tests.

4 The TOEIC and Reliability in Korean Context

4.1 Reliability in Terms of Test Facets and Random Factors

One often noted benefit of the TOEIC is its claim to be a highly reliable test, estimated by ETS to be within 25 points for one Standard Error of Measurement (ETS 1998a:IV-5). As exemplified in the sample test in Appendix 1, when looking at reliability in terms of Bachman's (1990:164) concern with test method facets, it appears that the TOEIC is successful in minimizing reliability issues in these areas. This is done in part by having the test clearly organized, including clear instructions and by attempting to provide non-culturally based inputs.

Where possible ETS also ensures reliability in terms of random factors by the objective marking possible with a computer scored multiple choice test, the effort put into developing equivalent test forms that are unbiased to candidates worldwide and then reviewed for fairness (ETS 2003:9), and in the equating of these test forms (ETS 1998a:II-4). The highly standardized testing environment and procedures also contribute to an increase in overall test reliability.

4.2 Reliability in Terms of Personal Characteristics

Bachman's (1990:164) consideration of personal characteristics a reliability issue gives an interesting insight into the TOEIC's reliability. On a purely intuitive level, it seems that in a test-retest situation, or a retest using a statistically equated test form, candidates would score higher in the second testing simply due to what Bachman calls test wiseness (1990:114). Interestingly, while I was able to locate numerous statistics relating to the TOEIC, it seems that no information is available for untrained first time candidates in a test-retest situation. Paralleling test wiseness, it further seems intuitive that test preparation courses focusing on teaching test strategies would also have a positive effect on a candidate's score. Despite claims made to this effect by numerous preparation schools and textbook publishers, research in this area with respect to the TOEIC also seems lacking.

One line of evidence that may support reliability issues relating to test wiseness is that the mean TOEIC scores of those who retake the test between two and four times rises continually (ETS 2004:10). While this would be expected due to an improvement in English skills, a further study, elaborated on below under criterion referenced validity, showed that Korean TOEIC scores significantly over estimated Korean's Language Proficiency Interview (LPI) scores compared to TOEIC-LPI estimates for non-Koreans, a result the authors attributes to a relative lag in speaking skills among Korean candidates (Wilson 2001).

If it is hypothesized that sitting the TOEIC numerous times, along with taking extensive test preparation courses which focus on developing test taking strategies was able to increase candidates' test wiseness and therefore test score without significantly improving their overall English ability, then it could be suggested that for a test which candidates did not prepare for and had relatively little familiarity with, in this case the LPI, the results would show a more accurate measure of their overall English ability. I therefore suggest that Korean LPI scores could be overestimated not due to a lag in speaking skills, but due to a lack of reliable TOEIC scoring resulting from students who are excessively test wise.

As it may be possible for candidates to noticeably improve their TOEIC scores without improving their overall language ability, I would suggest this constitutes unreliability based on a personal characteristic, and is thus a threat to the TOEIC's overall reliability. This is especially important in light of the commonness and sophistication of test preparation courses and materials in Korea, and the incredibly large number of Korean candidates who retake the test.

4.3 Establishing and Reestablishing TOEIC's Reliability

Considering the above discussion, it appears that while the TOEIC's reliability is reasonably established in many aspects, in terms of personal characteristics it clearly needs to be reestablished. The first step in this procedure would be to conduct a series of experiments exploring the test-retest characteristics of the TOEIC, and including candidates with initially little TOEIC experience, as well as those familiar with the test. Such experiments would need to consider the subsequent test results when candidates received extensive strategy training, but minimal English training, between testings.

If the results of this first procedure showed little effect from test wiseness or test preparation programs, then this would likely support the TOEIC's overall reliability claims. However, if the results showed a significant effect, then further work, perhaps theoretical and statistical, would need to be done to develop methods of estimating the effect of test wiseness before claims of reliability could be supported.

Perhaps however the TOEIC's reliability is most threatened in terms of personal characteristic by many candidates' desire to get the highest possible score in the easiest possible way. In this light it is hard to imagine how the TOEIC in its present form, or any indirect test of a language ability could be robust enough to withstand the capacity of the test preparation industry to undermine a test's reliability.

5 The TOEIC and Validity in The Korean Context

In considering the TOEIC's validity, it is important to remember Bachman's assertion (1990:160) that reliability is a necessary condition for validity. From this it should be clear that until the test is shown to be reliable, despite threats from candidates' personal characteristics, it will be impossible to fully establish its validity.

5.1 Construct and Content Validity

ETS (ETS 1998a:I-1) indicates the TOEIC was designed to measure language skills and abilities needed to communicate internationally, and that

> The test questions are developed from samples of spoken and written language collected from various countries around the world where English is used in the workplace. (ETS 1998a:II-1)

This suggests the test is more based on a needs assessment than a theory of language, and as construct validity is based on theoretical/demonstrated relationships between language constructs and language ability, it is thus hard to claim that the TOEIC exhibits construct validity without further work. Furthermore the TOEIC Technical Manual under the title "Construct-related validity" states, "the most common form of test validation is correlation with other, established methods" (ETS 1998a:III-1). This is a clear reference to concurrent validity, and despite the section title, there is no reference to construct related validity in line with the definitions presented in section 2.2.1 above.

Independently establishing the TOEIC's content validity is also a concern. As this would necessitate comparing test forms with test specifications (Hughes 1989:22), yet full specifications for the TOEIC do not seem to be available, and old test forms are confidential, it seems this could only be done in-house by ETS.

5.2 Criterion-Related Validity

The TOEIC's main claims of validity arise through establishing concurrent validity to other listening, speaking, reading and writing tests. In the TOEIC's initial validation (Woodford 1982:12), ETS found a correlation of 0.90 between the TOEIC and an apparently in-house developed direct test of listening ability. Additionally the TOEIC Technical Manuel lists numerous correlations with other listening tests ranging from 0.67 to 0.92 (ETS 1998a:III-3). Similar results to these have been reported for the TOEIC as a reading test, with correlations of between 0.73 and 0.87 (Woodford 1982:13, ETS 1998a:III-4). While it would be favourable to have some of these correlations higher, with few apparent criticisms in the literature to the TOEIC as a listening or reading test, there seems to be reasonable support for considering concurrent validity to have been established in these capacities.

While the TOEIC does not investigate a candidate's speaking or writing skills, ETS has also attempted to establish the concurrent validity of the test for these abilities. For speaking, ETS has focused on correlating the TOEIC to the US Foreign Service Institute's Language Proficiency Interview (LPI). Results from this have shown a correlation of between 0.71 and 0.83 (Woodford 1982:14, Wilson 1993:6), only slightly higher than Hughes' suggested minimum correlation of 0.70 for a relatively low stakes test (1989:24). However as noted in section 4.1 above, within a Korean sample the correlation was much lower at between 0.48 and 0.57 (Wilson 2001). Other correlations between the TOEIC and speaking tests have produced some individual values as low as 0.49 (Hirai 2002). These results mirror a general tenor in the independent literature to see the TOEIC as having only limited validity as a speaking test.

With one exception, the TOEIC's claim of being a valid writing test seems to have had little attention since the TOEIC's Initial Validation Study. That study (Woodford 1982:15) reported a correlation of 0.83 to an apparently in-house developed writing test, a figure which contrasts with Hirai's correlation of 0.66 for the TOEIC against the BULATS test of writing (Hirai 2002). Hirai additionally strongly criticizes the original ETS writing test, and suggests a TOEIC score is "practically meaningless as a measure of writing skill" (Hirai 2002).

5.3 Face Validity

As TOEIC questions are drawn from a range of real world situations and needs (ETS 2003:5), the TOEIC should have relatively high face validity. Unfortunately however, it is very hard to find any Koreans, test takers or otherwise, who feel the test is a good test of English. Considering the extensive test preparation industry noted above, I would suggest that this lack of face validity among students may be due to a feeling that the best way to get a high mark is not to study English, but instead to learn the 'tricks' of the test.

5.4 Establishing and Reestablishing TOEIC's Test Validity.

While the TOEIC does not seem to have a clearly mapped out construct validity argument, it appears from the TOEIC Technical Manual (ETS 1998a:III-1) that ETS feels its construct validity has been sufficiently confirmed by showing concurrent validity. One problem however is that this may result in circular validation in which each test is validated against another, yet with none of the tests in itself shown to have sufficient construct validity, overall validity is still not established (Bachman 1990:249). From this it thus seems a more formal procedure aimed at establishing the TOEIC's own construct validity is needed. Such a procedure would involve ongoing theoretical work and practical research to show how the differing TOEIC sections contribute to the measuring of listening, reading, speaking and perhaps writing abilities.

Apart from circular validation, it seems the greatest threat to establishing the TOEIC's concurrent validity is found in the few correlations of the TOEIC to a wide variety of speaking and writing measures. For speaking, ETS appears to have only pursued concurrent validation with respect to the LPI, and little non-ETS sponsored research seems to have been conducted. Procedures for establishing concurrent validity would thus require the TOEIC to be correlated to a wider range of speaking tests, and in particular for this work to be carried out independent of ETS. Similar reasoning suggests that the TOEIC's concurrent referenced validity in terms of writing ability could be enhanced in a similar way.

As noted in section 5.1, the main problem with independently establishing the TOEIC's content validity is access to test forms and detailed test specifications. As test forms are recycled, ETS seems highly unlikely to make them commonly available, thus the only possible procedure would be for ETS to evaluate such materials in-house and publish the findings. While this would be less than ideal, it would at least assist in the establishment of content validity.

Compared to the work needed above, procedures for establishing the TOEIC's face validity would be relatively simple. This would involve educating stakeholders of how the test is able to provide sufficient information, through indirect measures, to accurately assess a candidate's abilities, and how the test is robust enough to withstand challenges to its reliability, especially those noted in section 4.2 above.

6 Conclusion

Clearly reliability and validity are both extremely important test considerations. I suggest that this paper has demonstrated that despite some concerns, the TOEIC seams to be a relatively reliable and valid test of listening and reading comprehension for untrained first time test takers, although its validity as a measure of speaking and writing ability is much more in question for these same candidates. Unfortunately, for those many candidates who have taken TOEIC test preparation classes or have written the test numerous times, I suggest the TOEIC's reliability and hence its validity is much more tenuous. Considering how widespread the use of this test is, it seems clear that these issues need to be addressed as soon as possible.

References

American Psychological Association. (1985) *Standards For Educational and Psychological Testing*. Washington DC: American Psychological Association.

Bachman, L. F. (1990) *Fundamental Considerations in Language Testing*. Shanghai: Shanghai University Press.

Bachman, L. F. and Palmer, S. (1996) *Language Testing in Practice*. Shanghai: Shanghai University Press.

Educational Testing Service. (1998a) *TOEIC Technical Manuel*. Princeton NJ: Educational Testing Service.

Educational Testing Service. (1998b) *TOEIC Can-Do Guide*. Princeton NJ: Educational Testing Service.

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

Educational Testing Service. (2004) *TOEIC Report on World Test Takers Worldwide 2002-03*. Princeton NJ: Educational Testing Service.

Gilfert, S. (1996 July) 'A review of TOEIC'. *The Internet TESL Journal.* Vol. II No. 8. Available: http://iteslj.org/Articles/Gilfert-TOEIC.html (3 Sept 2005).

Hirai, M. (2002). Correlations between active skill and passive skill test scores. *Shiken: JALT Testing & Evaluation SIG Newsletter*. 6(3), 2-8. Available: http://www.jalt.org/test/hir_1.htm (3 Sept 2005).

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

Lado, R. (1961) *Language Testing*. London: Longman.

Lewis, R. J. (1999) 'Reliability and Validity: Meaning and Measurement'. Paper presented at Society for Academic Emergency Medicine, Boston, Ma. Available: www.ambpeds.org/ReliabilityandValidity.pdf (10 September 2005).

Wilson, K. M. (1993) *Relating TOEIC Scores to Oral Proficiency Interview Ratings*. Princeton NJ: Educational Testing Service.

Wilson, K. M. (2001) *Overestimation of LPI Ratings for Native-Korean Speakers in the TOEIC Context: Search for Explanation*. Princeton NJ: Educational Testing Service.

Woodford, P. E. (1982) *An Introduction to TOEIC: The Initial Validation Survey*. Princeton NJ: Educational Testing Service.

Appendix 1: Sample TOEIC Version