Macrothink Institute™

International Journal of Education
ISSN 1948-5476
2015, Vol. 7, No. 1

Evaluating the TOEIC® in South Korea: Practicality, Reliability and Validity

Simon James Nicholson[1,*]

[1]Foreign Languages Education Center, Hankuk University of Foreign Studies, 89 Wangsan-ri, Mohyeon-myeon, Choin-gu, Yongin-si, Gyeonggi-do, South Korea

*Correspondence: Foreign Languages Education Center, Hankuk University of Foreign Studies, 89 Wangsan-ri, Mohyeon-myeon, Choin-gu, Yongin-si, Gyeonggi-do, South Korea. Tel: 82-10-8974-1948 E-mail: simonjamesnicholson@gmail.com

Received: February 23, 2015 Accepted: March 16, 2015 Published: March 26, 2015

doi:10.5296/ije.v7i1.7148 URL: http://dx.doi.org/10.5296/ije.v7i1.7148

Abstract

With the rise of globalization and the reality of English as a lingua franca in international business, it is hard to argue against the need for a practical proficiency test for international communication in the workplace. However, the TOEIC® does not adequately meet this need. This paper critically evaluates the TOEIC® in South Korea and draws on relevant literature to discuss the classic criteria for assessing a test: practicality, reliability and validity. It proposes that though the TOEIC® is practical, its reliability is questionable and that the TOEIC® is inappropriate for its intended purposes as an indicator of language ability as it fails to provide any direct evidence of its validity in order to support its claim of being a true measure of English language proficiency.

Keywords: TOEIC, Test Design, Practicality, Reliability, Validity, South Korea, Proficiency, Language.

1. Introduction

English language tests offer a basis in which to make measurements and comparisons of a candidate's language skills and "contribute to decisions of critical importance in today's world, and yet the growing dominance of these tests is often unquestioned, unchallenged, unmonitored and uncontrolled" (Shohamy, 2007: 524). The Test of English for International Communication (TOEIC®) is such a test and one that the author has been involved with as a preparation course instructor. It is considered a high stakes test as it serves as the largest gatekeeper to professional employment in South Korea, and while it is marketed as a practical, reliable and valid test of business English proficiency, there exists little independent research to verify these claims.

This paper sets out to critically evaluate the TOEIC® test as a measure of communicative language abilities in the workplace. To examine these issues, the paper will first draw on relevant literature to discuss the classic criteria for assessing a test: practicality, reliability and validity (Brown, 2001: 385). The author will then give an overview of the TOEIC® before applying the aforementioned assessment criteria to determine the overall success of the test. The paper will finally propose that though the TOEIC® is practical, its reliability is questionable and that it fails to provide any direct evidence of its validity in order to support its claim of being a true measure of English language proficiency.

2. Assessment Criteria: Practicality, Reliability and Validity

Given the gravity of decisions that are made on the basis of test scores it is of critical importance that systems are set in place to ensure that they are in fact measuring correctly. This is supported by Bachman and Palmer who state that the "primary purpose of a language test is to provide a *measure* that we can interpret as an indicator of an individuals language ability" (1996: 23). This section will look at the essential qualities that make a test an accurate measure of these abilities, specifically reliability and validity. In addition, we will also examine test practicality as this is considered a fundamental requirement and important to the overall success of a test.

2.1 Practicality

Whether a test is practical or not is related to the resource demands of the test set against the available resources (Bachman and Palmer, 1996) and is considered an essential criteria contributing to its success. At the development stage it is common for a test to exceed available resources, but in the long run if a test exceeds the resources that it has available then it will inevitably be deemed unsuccessful. Bachman and Palmer (1996: 37) provide a list of the different types of resources to consider in testing:

- **Human resources**: test writers, scorers or raters, test administrators, and clerical support.
- **Material resources**: space, equipment and materials.

- **Time**: development time and time for specific tasks.

Practicality then refers to the economy of time, effort and money in testing and the consideration of resources is strongly linked to the financial costs involved in developing and administering a test. For a test to be practical it must be practical in terms of financial limitation, time constraints, ease of administration, scoring and interpretation.

2.2 Reliability

While practicality is a crucial element to the overall success of a test, reliability on the other hand is a critical accuracy measurement quality and is defined by Bachman and Palmer (1996) as the consistency of measurement. That is, if a test measures consistently it is said to be reliable (Hughes, 2003). There are two types of reliability:

- **Test reliability**: means that if we assign the same test to the same subjects or matched subjects on two different occasions it would yield the same result.
- **Scorer reliability**: refers to the consistency of scoring by two or more scorers.

Therefore, reliability implies that an individual test taker's score would be very similar whether or not they sat the test on a different day, in a different setting or even a different form of the same test. An unreliable test on the other hand would give different measurements of an individual's score and these fluctuations would not make the test particularly useful. Reliability, then, is concerned with ensuring that a test is in effect measuring consistently the same elements in different settings and across different formats.

2.3 Validity

A further essential component concerned with the accuracy of a test, and needed to justify the interpretations made on an individual's language ability from the scores of a test, is validity. This is because validity is concerned with whether the test measures what it is intended to measure (Hughes, 2003) and for this "we need to provide evidence that the test score reflects the area(s) of language ability we want to measure, and very little else" (Bachman and Palmer, 1996: 21). Presently, validity perspectives are viewed as a 'unitary though multifaceted concept' (Messick, 1989; Bachman, 1990) that includes "an argument supported by multiple kinds of evidence that justify the score interpretations and use of the test for its intended purpose" (AERA [American Educational Research Association], APA [American Psychological Association] & NCME [National Council on Measurement in Education] 1999; cited in Stoynoff, 2009: 3).

Therefore, multiple kinds of validity evidence are used to build the validity argument. However, construct validity has come to refer to the general, overarching notion of validity in language testing (Hughes, 2003) and represents the extent to which a given test score can be interpreted as an indicator of the abilities or constructs that are to be measured (Bachman and Palmer, 1996). Therefore, theoretical categories like 'communicative competence' are constructs and, when preparing tests, need to reflect the definitions of these constructs.

In addition, validity must also ensure that the test's *content* is based on the theoretical construct of the ability(ies) being measured. This involves providing evidence that the contents of a test are matched to its underlying theoretical construct. Hughes (2003) offers several subordinate forms of validity that can be used in support of construct validity and as a means of providing evidence to the unified notion of validity. These include the following:

- **Content validity:** does the test fully cover the theoretical construct; does it cover all the skills it claims to measure.
- **Criterion related validity** does it give the same results as other reliable tests of the test-takers abilities (there are two aspects of criterion related validity)
    - **Concurrent validity**: where the test and criterion are measured at the same time.
    - **Predictive validity**: how well the test accurately predicts the test takers future performance.

In summary of the above testing theories, while practicality is concerned with the resources and costs involved in developing and administering a test, both test reliability and validity are concerned with the accuracy of measurement. It should be noted here that validity and reliability are closely related, often complimentary but that tensions do exist and that trying to maximize all of these qualities "inevitably leads to conflict" (Lawson, 2008: 1).

3. The TOEIC® Test

The following sections will give an overview of the TOEIC® in regards to background, purposes and use, clientele and underlying theories. For the purposes of this paper, the TOEIC® will be discussed in reference to the specific context of South Korea.

3.1 Background

The TOEIC® was designed by the Chauncey Group and first administered in December 1979 in Japan where it was taken by 2,710 people to meet a request by Japanese corporations to measure overall communication skills required for international business (Choi, 2008). Since then, it has seen rapid growth and "has become the de-facto standard measure of English proficiency in many parts of Asia, at least in business contexts" (Chapman and Newfields, 2008: 32). The test is now developed, administered and scored by Educational Testing Service (ETS), a private non profit organization headquartered in Princeton, New Jersey, who report that in 2010 the test was "used by over 10,000 companies, government agencies and English Language Learning programs in 120 countries, and more than 6 million TOEIC® tests were administered last year" (ETS, 2013: 1), making it "the largest and most widely used English-language assessment for the workplace" (ETS, 2012a). Clearly, these figures show that the TOEIC® is playing a predominant and influential role in the lives of many, and this means that an analysis into its claims as being a reliable and valid test are necessary.

In South Korea, the TOEIC® test has been administered since 1982 as a multiple-choice test

of Listening and Reading (L/R). In 2011, the L/R test was administered more than 2.1 million times (ETS, 2012b; cited in Thomson, 2012: 19) whereas the Speaking and Writing (S/W) test, introduced in 2006, was administered fewer than 200,000 times (ETS 2012c; cited in Thompson, 2012: 19). Though the S/W test is gaining in importance, the L/R test is still generally considered as the standard TOEIC® as it has for over two decades. For this reason the focus of this paper is to critically examine the TOEIC® L/S test, herein known as the TOEIC®.

3.2 Test Purpose and Use

The purpose of the TOEIC® is to measure language proficiency. According to ETS, the TOEIC® "measures the everyday English skills of people working in an international environment…. scores indicate how well people can communicate in English with others in the global workplace" (ETS, 2007: 2). As a measure of everyday English skills or communicative abilities, its use is "to determine the proficiency levels of employees, or potential employees, for human resource planning and development in the contexts of business, industry, and commerce" (ETS, 2007: 2). However, its uses in South Korea at times fall beyond the scope of the purpose for which it was originally designed to measure, and it is often misused as an entrance examination, a requirement for earning credits and for placement purposes in universities (Choi, 2008).

3.3 Clientele

The TOEIC® seems to enjoy a remarkably high level of respect among companies. Many leading corporations in South Korea (such as Asiana Airlines, Daewoo Shipbuilding and Marine Engineering, Dongkuk Steel Mill Co Ltd, Hynix Semiconductor, Hyundai Heavy Industries Co Ltd, Hyundai Motor Company, Korean Air, LG Electronics, S-Oil Corporation and Renault Samsung Motors) use the TOEIC® as a criterion for recruiting, promotions and measurement of English training progress (ETS, 2009). This is largely due to the extremely well-targeted marketing strategies employed by ETS which marketed the TOEIC® to companies first, thereby establishing it as 'the company English test', before promoting it to broader individual and institutional markets. As Choi (2008) notes, this has led to the TOEIC®'s establishment as "practically the one and only EFL test for hiring employees at major Korean corporations" (44); and, while scores from other tests are accepted to demonstrate a candidates English skills, "84 of the 93 employees who joined the company [Hyundai Mobis] in 2010 submitted a score from the TOEIC® test" (Chung, 2010: 6) It is therefore not surprising that in South Korea the majority of TOEIC® test takers are of university age (ETS, 2012b: 14) and that the main motivating factor behind taking the TOEIC® is due to its importance in regards to employment opportunities. Clearly in the context of South Korea, the TOEIC® has become a powerful and influential tool for employment which has led to a rise in the number of people taking the test as "most Korean college and university students have had no choice but to prepare for this test" (Choi, 2008: 44). However, considering how widely used the TOEIC® is, the quality of independent data verifying the claims of ETS are startlingly low.

3.4 Underlying Theory

The TOEIC® only tests and measures listening and reading skills directly. Yet, ETS claim that it is an indirect measure of speaking and writing abilities thus implying that reading, listening, speaking and writing abilities are unitary, integrated skills. As Miyata states, "this would imply that the TOEIC® is constructed upon the theory that an individual's productive language abilities are proportional to his/her receptive abilities" (2004: 61). However, studies have shown doubts about these claims. Both Hirai (2002) and Cunningham, (2002) contend that the TOEIC® is an unreliable predictor of spoken and communicative competence which lies in direct opposition to ETS's claims that the purpose of the test is to measure communicative abilities. However, since the test does not measure any productive skills, serious questions must be raised about the theories that underlie the test, specifically in regards to validity.

4. Evaluating the TOEIC®: Practicality, Reliability and Validity

In this section we will analyze the practicality, reliability and validity of the TOEIC® to determine whether it can be considered the successful test of communicative language abilities that it claims to be.

4.1 TOEIC® Practicality

One of the TOEIC®'s greatest strength lies in its practicality. The answer cards are easy to mark in that the answers are either correct or incorrect and they are machine scored by a computer. This ease of marking gives quick turn around of results and ensures that it is relatively cheap to administer and less expensive than tests which directly measure all four language abilities, such as the TOEFL. Test takers and companies are drawn to the cost advantages that the standard TOEIC® offers as a measure of English proficiency (Chapman and Newfields, 2008). However, achieving such high practicality comes at a cost, namely in the areas of validity as the ease of marking can only be realized through the scoring of multiple choice responses which is limited to the testing of receptive skills.

4.2 TOEIC® Reliability

Reliability is an essential quality of any test as, without it, the TOEIC® would not be valid. The creators and administrators of the TOEIC® test claim a high reliability regarding their test (ETS, 2007: 5) and while on the surface these claims do appear to have some merit, specifically in relation to test method facets, "in terms of personal characteristics the TOEIC® reliability clearly needs to be re-established" (Sewell, 2005:12). The following two sections will explore test method facets and personal characteristics in greater detail and their relationship to TOEIC® reliability.

4.2.1 Test Method Facets

The TOEIC® is a norm referenced test and is designed to discriminate between high and low achievers on a standardized scale of 0-990. The test is a multiple choice exam, with two

equally weighted sections: listening (divided into four parts) and reading (divided into three parts). The listening section is administered via compact disc and lasts 45 minutes. The reading section allows 75 minutes. As mentioned above (4.2), all answers are machine marked by a computer.

It is clear that TOEIC® procedures are highly standardized thus giving the appearance of a reliable test. Test takers respond to each test question by marking the letter (A), (B), (C) or (D) on a separate answer sheet. The testing time is approximately two hours. Books, dictionaries, papers, notes, rulers, calculators, watch alarms, mobile phones, listening devices, recording or photographic equipment, highlighters or aids of any kind are not allowed into the testing room. Furthermore, test takers may not mark, underline words or make notes in the test book or on the answer sheet and twenty eight days should lapse before a test taker retakes the TOEIC® test.

These well-standardized administrative procedures give a certain degree of reliability to the test. Scores are unlikely to vary due to different test settings or subjective marking and this gives the appearance of consistency. In addition, answers are computer-marked, thereby creating high scorer reliability (see 2.2). However, standard settings and objectivity are not the only factors that can effect variations in test scores. Personal attributes are also a factor in measuring reliability.

4.2.2 Personal Attributes

The greatest threat to the reliability of the TOEIC® test lies in the degree to which score improvements can be accredited to test familiarity. There are two ways of increasing test familiarity: taking test preparation courses and sitting the test multiple times. In the 2002-03 testing period an incredible 99% of South Korean candidates claimed to have previously written the test (ETS, 2004:10). Furthermore, there is an abundance of TOEIC® test taking preparation courses on offer in Korea where the focus is not on improving communicative competence but on developing test-taking strategies in order to improve TOEIC® scores. That the TOEIC® is so formalized and "allows students to employ test taking strategies to get a high score without knowing how to use the language" (Thomson, 2012: 18) demonstrates that these personal characteristics threaten the reliability of the TOEIC® test as a consistent and true measurement of language abilities.

4.3 TOEIC® Validity

There exists very little evidence of validity to the TOEIC® as an indicator of communicative language skills. This section will look at some of that evidence and answer the questions: does the TOEIC® fulfill the purpose for which it was designed and does it actually test for the skills it claims to measure?

4.3.1 Construct Validity

As discussed in section 2.3, construct validity refers to the overarching notion of validity in language testing. The TOEIC® claims to indicate English communication abilities with others in the global workplace (ETS, 2007: 2). One of the ways that ETS could demonstrate

construct validity is to provide a definition of the theoretical construct of communicative abilities that the test purports to measure. However, ETS fail to provide such a definition and what is claimed to be measured by the exam remains tenuous at best. Though further investigation into the specific language areas required to show a communicative language construct is needed, the fact that the TOEIC® remains a multiple choice test in a limited response format (receptive language skills) and seems largely to test for knowledge of mostly grammar and vocabulary clearly demonstrates the test's lack of a communicative construct thus rendering its validity weak. For the TOEIC® to validate its claims it needs to test for not only grammar, but also discourse, sociolinguistics and illocutionary competence and it fails to do so (Douglas, 2000). Instead of showing a construct of the communicative abilities that the test claims to measure, ETS have tried to support construct validity through criterion and concurrent validity evidence, which will be discussed in greater detail in section 4.3.3.

4.3.2 Content Validity

According to Brown and Hudson (1998: 658) TOEIC® is an example of a selected response test as it does not require test-takers to use any productive language skills. The content of the TOEIC® is open to further criticisms as a measurement of communicative language abilities in that as Chapman and Newfield state, "over half of the questions in this test still focus on sentence level comprehension rather than discourse level input" (2008: 2). In addition, the test also purports to use "international English, the language used most often to conduct global business" (ETS, 2011: 4). However, the international language used in the test is still limited to the accents spoken by inner-circle countries (Kachru, 1982) and is unlikely to represent the actual international English language accents most Koreans would be exposed to in real business life settings, i.e., varieties of Asian English. As Chapman and Newfields (2008) point out, "the accents [used in the TOEIC®] represent a narrow sample of the range of varieties that are spoken worldwide, particularly in the context of Asia" (cited in Booth, 2012: 44).

Content validity as discussed in section 2.3 is "the extent to which a test's content is *proportionally representative* of all the construct's features" (Morotoshi, 2001: 9). Content validity is therefore directly related to construct validity, as without a theoretical construct of the features we wish to measure, it is impossible to measure the proportion of its content and how it relates to the construct. Indeed, ETS make no claim that the test's content is proportionally represented of its construct; as previously stated, there exists no operational definition of its construct. Without content validity the test is unlikely to be an accurate measure of what it purports to measure.

Furthermore, as a direct test of listening skills the content validity of the listening section has come under criticism (Buck, 2001; Douglas, 1992; Hirai, 2002). As Booth (2012) argues:

> [The TOEIC®] fails to assess essential aspects of listening comprehension required in real-life communication. This includes: indirect speech acts, pragmatic implication, or other aspects of interactive language use, including natural hesitations, phonological shifts and negotiations for meaning between interlocutors (Booth: 2012: 29).

Further content validity is questioned by Lee, Yoshizawa and Shimabayashi (2006) who point out that the TOEIC® does not measure a specific business English domain as suggested by the construct and as Chapman and Newfields (2008) point out it still does not employ authentic methods of testing reading comprehension. Therefore, even within the realms of listening and reading, TOEIC®'s content validity is still weak.

4.3.3 Criterion Related Validity

Concerning the two forms of criterion related validity, predictive and concurrent, ETS have relied exclusively on the latter. An external aspect of validity and one that would strengthen ETS validity claims is that of how well the TOEIC® scores directly relate to communicative job performance. However, the difficulty in dealing with predictive validity is that many other factors are involved in determining future performance and this makes it very difficult to measure. Almost exclusively by establishing concurrent validity to other established listening, speaking, reading and writing tests (including the LPI [Language Proficiency Interview], TOEFL [Test of English as a Foreign Language], and OPI [Oral Proficiency Interview]), ETS have relied upon concurrent validity as a means of validation to their claims that the standard TOEIC® test is a basis for predicting oral proficiencies and therefore a measure of English communicative abilities.

However, correlating the TOEIC® with other tests is inappropriate as, in the case of TOEFL, the test constructs are quite different, i.e. that of English in the global workplace and English for academic purposes respectively (Chapman and Newfields, 2008; Niall, 2004). Therefore, the claims to establish validity through concurrent correlations are erroneous, and as Bachman states, should be dismissed on the grounds that it "simply extends the assumption of validity to these other criteria, leading to an endless spiral of concurrent relatedness" (1990: 249). It would seem that ETS are beginning to address these concerns with recent statements regarding issues with concurrent validity. As Liao, Qu and Morgan (2010), researchers for ETS, state:

> [A]lthough it is natural to assume that different language skills are correlated with each other to a certain degree...each test measures distinct aspects of language proficiency that cannot be assessed by the other tests. (2010: 11)

In other words, ETS are now indirectly acknowledging that correlating the TOEIC® with other tests is incongruous with the overall validity of the test.

5. Discussion

It is clear from the analysis in previous sections that the TOEIC® is a practical and somewhat reliable test. However, in the current situation there is no evidence to show that it is the valid indicator of communicative ability in international business contexts that it claims to be. The introduction of the TOEIC® Speaking and Writing test in December 2006 does help to promote greater construct validity by making the TOEIC® a more comprehensive and communicative measure of L2 ability, though arguably, speaking answers into a computer

fails to engage the test taker in true aspects of communication. Nevertheless, this is a step in the right direction. However, that it remains an optional, more expensive and less practical measure, taken by vastly less numbers, leaves the standard TOEIC® still functioning as the sole measure of communicative ability in the workplace for most employment candidates and corporations. It is the author's belief that for the TOEIC® to be marketed as a more reliable measure of communicative language proficiency, the Speaking and Writing Test should be made an integral part of the entire test package rather than an optional element. Furthermore, greater evidence in regards to the test's construct, content and criterion related validity is needed to ensure that the TOEIC® is an accurate indicator of communicative competence and fulfilling the purpose for which it was designed. However, until these issues are addressed, the standard TOEIC® will continue to function wrongly as the so-called measure of communicative ability as it has done now for over two decades.

6. Conclusion

With the rise of globalization and the reality of English as a lingua franca in international business, it is hard to argue against the need for a practical proficiency test for international communication in the workplace. However, the TOEIC® does not adequately meet this need as it fails to be a reliable and valid measurement of English language proficiency. Despite this, due to its practicality and misleading reliability and validity, it remains as one of the leading gatekeepers of advancement in South Korea.

This paper has shown that the TOEIC® is inappropriate for its intended purposes as an indicator of language ability. It is of no wonder, then, that despite the importance that companies and test takers place on the test, and the concentration and focus put upon attaining competitive TOEIC® scores, that many cannot communicate in English even with high TOEIC® marks (Choi, 2008). This is the test's greatest failure as it lets both test takers and companies down, for in their efforts to attain the desired high TOEIC® score, communicative language skills are overlooked. Until the TOEIC® becomes a more accurate measure of true communicative competence, TOEIC® fever will continue to be a detriment to communicative language teaching and English language acquisition in South Korea.

Acknowledgement

This work was supported by Hankuk University of Foreign Studies Research Fund 2015.

References

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

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

Bachman, Lyle, F., & Palmer Adrian, S. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Booth, D, K. (2012). 'Exploring the washback of the TOEIC in South Korea: A sociocultural perspective on student test activity.' Phd. The University of Auckland. Retrieved 20 March 2013 from https://researchspace.auckland.ac.nz/handle/2292/19379

Brown, J., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly*, *32*(4), 653-675. http://dx.doi.org/10.2307/3587999

Buck, G. (2001). *Assessing Listening.* Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511732959

Brown, H. D. (2001). *Teaching by Principles* (2nd edition). New York: Longman.

Chapman, M., & Newfields, T. (2008). The 'new' TOEIC. *Shiken: JALT Testing & Evaluation SIG Newsletter, 12*(2), 32 – 37.

Choi, I.C. (2008). The impact of EFL teaching on EFL education in Korea. *Language Testing, 25*(1), 39-62. http://dx.doi.org/10.1177/0265532207083744

Chung, H, W. (2010). Language instruction and promotion requirements upgraded to foster English skills. *TOEIC Newsletter.* Vol. 58. KOREA TOEIC Committee.

Cunningham, C. (2002). *The TOEIC test and communicative competence: Do test score gains correlate with increased competence?*. Unpublished Masters Dissertation. Retrieved 6th January from http://www.bhamlive.bham.ac.uk/Documents/college-artslaw/cels/essays/matefltesldissertations/Cunndiss.pdf

Douglas, D. (1992). Test of english for international communication. In J. Kramer, & J. Conoley, (Eds.), *The eleventh mental measurements yearbook* (pp. 950-951). Lincoln, NE: Buros Institute of Mental Measurements.

Douglas, D. (2000). *Assessing Languages for Specific Purposes*. Cambridge, England: Cambridge University Press.

Educational Testing Service. (2004). *TOEIC report on world test takers worldwide 2002-03*. Princeton NJ: Educational Testing Service.

Educational Testing Service. (2007). *Examine handbook listening and reading*. Princeton NJ: Educational Testing Service. Retrieved 20 March 2013 from http://www.ets.org/Media/Tests/TOEIC/pdf/TOEIC_LR_examinee_handbook.pdf

Educational Testing Service. (2008). ATA celebrate 10 years of the *TOEIC*® test in China, 2012. *JALT Testing & Evaluation SIG Newsletter, 12*(2), 32 - 37.

Educational Testing Service. (2009). *You are in good company when you use the TOEIC tests*. Princeton NJ: Educational Testing Service.

Educational Testing Service. (2011). *Learn about TOEIC® test takers worldwide data from the 2010 TOEIC® listening and reading background questionnaire inside*. Princeton NJ: Educational Testing Service. Retrieved 20 March 2013 from http://www.ets.org/Media/Tests/TOEIC/pdf/12338_ToeicWwDatRprt_LR.pdf

Educational Testing Service. (2012a). *ETS, ATA celebrate 10 years of the TOEIC® test in China*. Princeton NJ: Educational Testing Service. Retrieved 20 March 2013 from http://www.ets.org/toeic/news/ets_ata_china

Educational Testing Service. (2012b). *TOEIC® test data & analysis 2011*. Princeton NJ: Educational Testing Service. Retrieved 20 March 2013 from http://www.toeic.or.jp/toeic_en/pdf/data/TOEIC_DAA2011.pdf

Educational Testing Service. (2013). *The TOEIC® tests: the global standard for assessing English proficiency for business*. Princeton NJ: Educational Testing Service. Retrieved 20 Marches 2013 from http://www.ets.org/toeic/succeed?WT.ac=toeichome_succeed_121127

Educational Testing Service. (2012c). *2011 TOEIC speaking test score analysis*. Princeton NJ: Educational Testing Service. Unpublished Report.

Hirai, M. (2002). Correlations between active skill and passive skill test scores. *Shiken: JALT Testing & Evaluation SIG Newsletter*, *6*(3), 2-8.

Hughes, A. (2003). *Testing for Language Teachers.* Cambridge: Cambridge University Press.

Kachru, B. (1982). *The Other Tongue -- English Across Cultures*. Urbana, Ill.: University of Illinois Press.

Lawson, A. J. (2008). Testing the TOEIC: Practicality, reliability and validity in the test of English for international communication. University of Birmingham.: Retrieved 12th January 2013 from http://www.birmingham.ac.uk/Documents/college-artslaw/cels/essays/testing/M6TestingTOEICEssayALawson.pdf

Lee, S., Yoshizawa K., & Shimabayashi S. (2006). The content analysis of the TOEIC and its relevancy to language curricula in EFL contexts in Japan. *JLTA Journal*, *9*, 154-173.

Liao, C., Qu, Y., & Morgan, R. (2010). *The relationships of test scores measured by the TOEIC® listening and reading test and TOEIC® speaking and writing tests. Educational Testing Service.* Princeton NJ: Educational Testing Service. Retrieved 20 March 2013 from http://www.ets.org/research/policy_research_reports/publications/report/2010/itkd

Messick, S. (1989). 'Validity'. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education/Macmillan.

Miyata, R. (2004). A review of TOEIC: A critical analysis of the structure, methods, and purpose of TOEIC, including its effect on teaching and learning English as a second or foreign language. *Bulletin of Beppu University Junior College, 23*, 59-69.

Morotoshi, P. (2001). The test of English for international communication (TOEIC): Necessity, proficiency levels, test score utilization and accuracy. University of Birmingham. Retrieved 12th January 2013 from http://www.birmingham.ac.uk/Documents/college-artslaw/cels/essays/testing/PaulMoritoshisM5Testingassignmentgraded72.pdf

Niall, T. (2004). TOEIC: A discussion and analysis. *The ELT two cents cage.* Retrieved 9th January 2013 from http://www.oocities.org/twocentseltcafe/teach/toeic.html

Rebuck, M. (2003). National standards. The use of TOEIC by companies in Japan. *NUCB Journal of Language, Culture and Communication*, *5*(1), 23-32.

Sewell, H, D. (2005). *The TOEIC: reliability and validity within the Korean context*. University of Birmingham.

Shohamy, E. (2007). The power of language tests, the power of the English language and the role of ELT. In J. Cummins & C. Davison (Eds.), *International handbook of language teaching* (pp. 521-532). New York: Springer. http://dx.doi.org/10.1007/978-0-387-46301-8_37

Stoynoff, S. (2009). Recent developments in language assessment and the case of our large-scale tests of ESOL ability. *Language Teaching, 42*, 1-40. http://dx.doi.org/10.1017/S0261444808005399

Thomson, S. (2012). *The Effects of TOEIC Education in South Korean Universities.* University of Birmingham.