Useful Information Teachers and Administrators Should Know About the TOEIC

Brian D. Bresnihan
University of Hyogo

Reference Data:
Bresnihan, B. D. (2013). Useful information teachers and administrators should know about the TOEIC. In N. Sonda & A. Krause (Eds.), *JALT2012 Conference Proceedings*. Tokyo: JALT.

In recent years, a total of over half a million TOEIC Institutional Program (IP) Tests and TOEIC Bridge IP Tests are administered to college and university students in Japan yearly. This paper covers various issues concerning the use of the TOEIC test in schools, such as using the TOEIC, a norm-referenced test, for criterion-referenced test purposes, which results in misuses of the scores, as does using the scores without reference to their standard errors of difference. It also explains why it is unreasonable for schools to expect that most of their students' TOEIC scores will increase, even after a semester or year of studying, and points out that using the TOEIC test tends to hinder the teaching, practicing, and learning of certain English language abilities, namely those that are not specifically addressed on the TOEIC test.

近年、総計５０万以上のTOEICインスティテューショナルプログラム(IP)テストとTOEICブリッジIPテストが、日本国内の大学生に対して執り行われている。この論文では、ノームレファレンステストであるTOEICを、クライテリオンレファレンステストの目的に使用することや、スタンダードエラーズオブディファランスを参照せずにスコアを使用することにより起こる、スコアの間違った解釈等、様々な問題点を取り上げている。また、何故、半年や1年の学習後でさえ、ほとんどの学生のスコアが上がることを学校が期待するのが不合理なのか説明し、TOEICの使用が、ある種の（TOEICで特に取り上げていない）英語能力の指導や練習や学習の妨げになっていることを指摘している。

Before the mid 1990s, the TOEIC test, produced by the Educational Testing Service (ETS), was practically unknown outside Japan's business community, for which it was created. Now it is the most sat for test of its kind in the world. In Japan, students of all ages are now sitting for it, or for the TOEIC Bridge test, which is a version for those with lower English language abilities. Its use has become particularly pervasive in Japan's colleges and universities, increasing every year. Over 440,000 TOEIC (IIBC, n.d.b, p. 8) and over 100,000 TOEIC Bridge (IIBC, n.d.a, p. 7) tests were administered in 2011, the last date for which statistics are available.

The standard TOEIC, the concern of this paper, tests listening and reading. Speaking and writing are not tested. (The TOEIC Speaking Test and the TOEIC Writing Test are totally separate tests.) Test takers receive a Listening score, a Reading score, and a Total score, which is simply the first two scores added together. There are two types of administrations: Secure Program (SP) and Institutional Program (IP). People who take the TOEIC test at their place of work or study are taking an IP test. Those who take the test at an official testing site are taking an SP test.

Certain issues arise when schools use the TOEIC test and its scores. In this paper, the following will be addressed:

- The TOEIC test was not created to test students' English abilities, how much students learned, or how well students performed in a class, but, according to Wilson, one of the leading TOEIC researchers for ETS, to measure workers' "English-language proficiency in the international work environment" (1993, p. 2).
- Even if TOEIC scores are used only to compare students for placement, the standard errors of difference must be used in order to make decisions fairly and correctly.
- TOEIC scores are not as precise as they appear to be, and they are not able to measure English abilities as exactly as some may think.
- It is unlikely that most school programs provide enough classroom hours of English language study for most of their students to be able to increase their TOEIC scores without extensive additional study on their own.
- Requiring the use of the TOEIC test or TOEIC scores tends to inhibit the use of other materials and practices which are beneficial and necessary for students to attain full, well-rounded acquisition of all English language abilities and result in overall competence.

Students as TOEIC Test Takers: The Issues of Validity and Reliability

"The TOEIC test is designed for use by organizations working in an international market where English is the primary language of communication" (CGI, 2000, p. 2). "It measures the everyday English skills of people working in an international environment. The scores indicate how well people can communicate in English with others in business, commerce, and industry" (ETS, 2012, p. 2) "in the global workplace. The test does not require specialized knowledge or vocabulary; it measures only the kind of English used in everyday work activities" (ETS, 2007, p. 2). Hardly any Japanese students have worked in the international business world or have had opportunities to use English in such situations. This lack of background knowledge and experience will cause at least some students to have difficulties understanding the contexts and contents of at least some test items. Examples of these are corporate development, investments, marketing, labor relations, plant management, board meetings, and various technical areas (ETS, 2007, 2012). These shortcomings will result in lower scores because of deficiencies other than those related to English abilities, that is, the students' lack of comprehension of the ways, interactions, contents, and circumstances of the international business world. It will also weaken the TOEIC's validity as a test of these students' English abilities. "A test is valid if it measures what it says it measures" (Kubiszyn & Borich, 1987, p. 278), and it is less valid the more other content or issues affect the test results. This lack of validity is what led to the discrediting, in the eyes of most scientists, if not the general public, of the results of IQ testing: The test results were influenced by factors other than intelligence, which itself is a very unquantifiable capacity (Gould, 1996; Poundstone, 2003, pp. 23-42). Modern standardized testing of such things as learning potential and language proficiency developed out of IQ testing (Poundstone, 2003, pp. 35-36).

Furthermore, as most students in Japan's colleges, universities, and high schools are admitted based on tests that usually include English tests, the students on any one campus, or in any one department, have a much narrower range of English proficiency than does the general population. This suggests that the great majority of their TOEIC scores will also fall into much more restricted ranges, resulting in weaker reliability (Stratton Ray, personal communication, 1 Dec 2008). ETS itself warns that, "If you have a sample of candidates who are very similar to

each other, the reliability of the test within that specific homogeneous group will be quite low. If there is no (or very little) variation among candidates' test scores then, by definition, there can be no accurate estimate of reliability" (CGI, 1998, p. IV.3). "A test is reliable . . . if it consistently yields the same, or nearly the same, ranks over repeated administrations during which we would not expect the trait being measured to have changed" (Kubiszyn & Borich, 1987, p. 291), and it is less reliable the more the rankings vary, which will be the case if a group of test takers have very similar abilities.

Without strong validity and strong reliability, TOEIC scores have little meaning.

Two Types of Tests: Criterion-Referenced and Norm-Referenced

Tests can be separated into two basic types, each of which provides different information about the test takers. Brown (1995) explained the various differences between them. Most tests created for educational purposes, possibly outside of placement, are *criterion-referenced tests*, which try to determine what and how much of certain information or skills a student knows or can perform. If the test is a posttest, the hope is that a great majority of the students will demonstrate mastery of the materials and skills they have studied by scoring highly on the test. For pretests, there is no desire for mastery to be demonstrated. Instead, the purpose is to discover what students already know or can do and what they still need to learn. With this information, the teachers can decide what to teach. The makers of criterion-referenced tests know the details of the individual test takers' abilities well and they create test items that measure precise details of what the test takers will be expected to know and be able to do. In the case of posttests, the students know in detail what the test will cover and are expected to study those specific materials and practice the skills to prepare for the test.

The other type is *norm-referenced tests*, which attempt to measure overall proficiency. If these are also standardized tests, which are administered to large numbers of people in many locations at the same time, as they usually are, the test makers know little if anything about the test takers. The test items of norm-referenced tests must cover a wide range of materials and abilities. Makers of such a test hope that there is no small, identifiable set of materials or precise skills that test takers could study to help them to achieve higher scores. Otherwise, the test would not measure overall proficiency. Of course, the test results cannot provide precise details about what individual test takers know or can do. The test makers also hope that, when all of the scores from one administration are gathered, they demonstrate a normal distribution, that is, that few scores are very high or very low and that most scores fall in a range around the middle of the scale.

The TOEIC test is a norm-referenced standardized test. Therefore, it cannot provide details about exactly what a test taker has learned in a class or what a test taker knows or does not know, can or cannot do. Instead, it gives information about how an individual's English proficiency compares with others who took the same test.

Interpreting TOEIC Scores

The possible TOEIC Total score range is from 10 to 990, and the possible Listening score and Reading score ranges are from 5 to 495. If an individual takes the TOEIC test twice (at times A and B), one cannot just subtract the A scores from the B scores to determine if any increases in the scores indicate true increases or by implication demonstrate probable increases in English language abilities. Instead, the standard error of difference must be used with each score to find out if the differences in the scores, when subtracted, are wide enough to confidently state that the test taker's B scores are truly higher than the A scores.

The same method needs to be used to determine if the scores of any two test takers are the same or different (PsyAsia, 2013a; CGI, 1998, p. IV.6-IV.7). Tests "are always associated with some degree of error. . . . An *obtained* score has a *true* score component (actual level of ability, skill, knowledge) and an *error* component (which may act to lower or raise the obtained score)" (Kubiszyn & Borich, 1987, p. 304). If the difference in two scores is less than the error or confidence band created using the standard error of difference, then neither score can truthfully be said to demonstrate higher ability. Scientists, mathematicians, and testing researchers are aware of this, and so created the practice of using error or confidence bands to make the measurements they gather more precise. Most people, however, do not know about these statistical procedures or the importance and necessity of using them.

Unfortunately, ETS does not publish the standard error of difference for the TOEIC Total score, which is the score many administrators, teachers, and students are concerned about. However, ETS does say that the standard error of difference for both the Listening score and the Reading score is about +/-35 points. This allows for a comparison of scores with 68% confidence. To be 95% confident in one's decisions, two standard errors of difference, or +/-69 points, need to be used (CGI, 1998, p. IV.6-IV.7). "Why bother with the 95-percent level . . . . If you are going to make important decisions about a student, a conservative approach appears warranted. . . . If you are concerned about the effects of a 'wrong' decision (that is, saying a real difference in achievement exists when it is really due to chance), take the conservative approach" (Kubiszyn & Borich, 1987, p. 321).

Table 1 presents the TOEIC IP Test scores for 10 of 25 freshman university students who were studying together in three English language classes, a convenience sample. Each class met once a week for 90 minutes, 15 weeks per semester. The students were placed in the classes in the Japanese equivalent of alphabetical order. They took the TOEIC test on campus twice, with 6 months between the two administrations, which included a summer break of 2 ½ months. These scores are typical examples of the approximately 1,200 students' scores from which they were selected, though there are also many students' scores with less variety and some with more. They were specifically chosen to demonstrate the points that will be made. The last three columns on the right give the changes in scores, that is, the amount each student's scores were higher or lower on the second test than the first. Taking a few minutes to look over and consider these scores may make following the discussion somewhat easier.

Table 1. Student TOEIC Scores on Two Administrations

	Total Score		Listening Score		Reading Score		Change in Score		
Student	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2	T2-T1	L2-L1	R2-R1
A	405	515	240	255	165	260	110	15	95
B	480	390	240	220	240	170	-90	-20	-70
C	395	495	240	270	155	225	100	30	70
D	500	585	280	350	220	235	85	70	15
E	530	490	295	315	235	175	-40	20	-60
F	460	450	270	230	190	220	-10	-40	30
G	445	480	210	270	235	210	35	60	-25
H	440	470	290	250	150	220	30	-40	70
I	635	625	305	270	330	355	-10	-35	25
J	385	465	205	225	180	240	80	20	60

In comparing scores for individual students in Table 1, there are only four changes in scores that demonstrate an increase

in Listening or Reading score with 95% confidence, that is, of more than 69 points: student D's change in Listening score and students A, C, and H's changes in Reading scores. Student B had the second highest score in Test 1, but the change in Reading score of more than 69 points indicates a truly lower score on Test 2 than on Test 1. No true increase or decrease in the other 15 Listening and Reading scores can be claimed with 95% confidence, including those of student I, who has the highest Total score on both tests. One thing to note is that students A and C, two of the three students whose Listening scores demonstrate true increases, had two of the three lowest Total scores on the first administration.

As for comparing scores between students, only the Listening scores of students G and J on the first administration are more than 69 points lower than any of the others' Listening scores, and so can be claimed to be truly lower with 95% confidence. All of the rest are statistically the same. On the second administration, all but one of the other students' Listening scores can be said to be truly lower than student D's, the highest, with 95% confidence. These eight scores are the same statistically as each other, and five of them are statistically the same as the second highest Listening score, student E's. Similar comparisons can be made for the Reading scores.

These kinds of comparisons and considerations ought to be carried out whenever TOEIC scores are being used, in order to understand what the scores are indicating and to use them fairly when judging and comparing test takers' English abilities. No matter how TOEIC scores are used, if standard errors of difference are not considered, it will lead to unfair and unwise decisions and practices, such as giving a higher score to one student than another based on scores that are within two standard errors of difference of each other.

TOEIC Scores as Precise Measurements of English Language Abilities

Some schools use TOEIC scores or gains in TOEIC scores as at least a partial measure for grading students in classes. Some schools use achievement of a certain TOEIC score as a criterion for awarding certificates of completion for courses. The latter usage may be justified, if it is in combination with other criteria, and if the administrators believe it is legitimate to use the TOEIC test to measure their students' English language abilities. Yet it may lead to misuse, for example, if students are denied being awarded a certificate in areas or fields not related to international business just because they did not achieve a high enough TOEIC score. The former usage, however, is definitely not intended by ETS and is not supported by norm-referenced tests, and so is erroneous. ETS publishes information on how to interpret and use TOEIC scores (CGI, 1998; ETS, 2007, 2012), yet it seems that this information may not be well known or understood by many administrators and teachers.

In the present age, people want numbers and measurements to support claims and ideas in all fields. This is considered as providing scientific proof. TOEIC scores seem to provide this proof. The scores appear to be precise measurements of test takers' English language abilities. However, Cameron (1963, p. 13) stated, "Not everything that can be counted counts, and not everything that counts can be counted," a claim seemingly so innovative, radical, and yet correct that it is often accredited incorrectly to Albert Einstein. This is the case with language abilities, which have no physical aspects, though language itself is manifested physically when we write or speak. Language abilities are aspects of our thinking, our will, and our feelings. They are part of our inner being and inner self, not part of our physical bodies, even though we use our physical bodies to make use of them. They cannot be assigned meaningful, precise, numerical scores, just as IQ cannot. Therefore, when a school

uses something like TOEIC scores, it needs to do so with great caution and with careful attention to the fairness and truthfulness of the usage.

Classroom Study and Increases in TOEIC Scores

Many students, teachers, and administrators would like to know how much time it takes to improve foreign language abilities enough to be demonstrated in increased test scores. In an attempt to answer this question, Saegusa (1985) generated multiple correlations and regression equations using pairs of TOEIC scores from workers who had been studying English in courses arranged by their companies. He then used these and standard errors of measurement to determine how much classroom study time would be needed to expect most learners to improve their TOEIC scores by certain amounts. He concluded that, "less than 80 hours of (English language) instruction is not very effective. In such classes, a majority will make little or no progress. If effectiveness is given top priority, at least more than 100 hours of instruction, and ideally 200 hours of instruction, as a unit should be recommended" (p. 174). He also determined that approximately 400 classroom hours of English language instruction would be needed for most students to raise their TOEIC Total scores from 450 to 600 or from 600 to 730 (p. 181).

As most Japanese college and university English language classes meet for just 90 minutes once a week for 15 weeks per semester, students would need to attend a minimum of five such classes, with nine being preferable, in order for most of them to be expected to raise their TOEIC scores. Such a schedule is likely to be found only in programs in which students major in English. With this information, and the information concerning standard errors of difference, the scores in Table 1 seem much less unusual. It is almost expected that few students would demonstrate improvement by achieving truly higher scores, as they did not spend enough time studying, unless they also studied English extensively outside of their classes. The seemingly large variability of many of the individual students' scores would also be expected, as "jumping around is in the nature of TOEIC scores" (Childs, 1995, p. 73), due, at least partially, to the sizes of the standard errors of difference.

In addition, Saegusa (1985, p. 167) explained that, generally, the company classes consisted of about 10 people per class, met for 2 hours two or three times a week (for a total of 50 to 200 hours during a period of 3 to 6 months), and were taught by native English speakers. Attendance was 80%. The English language study requirements at most colleges and universities in Japan do not meet most of these conditions. Therefore, it is possible that the estimates of classroom English study time needed for most students to raise their TOEIC scores by the amounts suggested by Saegusa would prove to be too low. Also, because Saegusa used standard errors of measurement, which are used for determining the range in which a test taker's true score falls based on a single obtained test score (PsyAsia, 2013b), where he should have used standard errors of difference, his estimates of the number of classroom hours of English language study needed for most students to achieve the gain scores he spoke of are probably about 30% too low (Bresnihan, 2010, p. 213-214).

Teaching for the TOEIC

ETS's initial head TOEIC researcher, Woodford, wrote, "The way in which we test can inform the manner in which we teach" (1982, p. 2). It can also distort the way we teach. It is not unusual for English language classes at Japanese colleges and universities to use TOEIC-like materials, which are the focus of a great many textbooks, and to have students do drill work with them. When asked about the practice of teaching the TOEIC test in an interview, an ETS representative seemed confused by the question and finally replied, "The student needs to be motivated to learn English and NOT simply to pass the test. . . . TOEIC is a test not

a language, so *teaching TOEIC* is not really an option. The best thing to do is to teach English focusing on proficiency rather than rules or vocabulary" (Wood, 2010, p. 44). If a class focuses on TOEIC-like drills, the students will not be encouraged to study or practice English in other ways or with other materials. On paper, these classes may appear systematic, efficient, and rigorous. In reality, though, such styles of teaching are stifling and ignore a great many other ways of learning and acquiring language and other aspects of language usage that students need to learn, not to mention materials to use. For example, there seems to be no reason to be able to read or understand novels or poems or to learn how to speak or write clearly and accurately, because these things are not on the TOEIC test. Yet, such abilities will surely benefit anyone who is interested in using English.

ETS has produced charts (for example, CGI, 2000; ETS, n.d.) that give expected speaking and writing abilities related to TOEIC Listening and Reading scores, although at the same time explaining that these are general claims and cannot be used as definitive statements about any particular person's abilities. The charts are based on research published by ETS (for example, CGI, 1998; Liao, Qu, & Morgan, 2010; Wilson, 1989, 1993; and Woodford, 1982), even though Liao, Qu, and Morgan concluded their study of the standard TOEIC test and the TOEIC Speaking and Writing tests by stating, "The results . . . confirm that there are four separate language skills measured by the TOEIC tests. It is natural that different language skills are correlated with each other to a certain degree; however, each test measures distinct aspects of English language proficiency that cannot be adequately assessed by the other tests. Examinees should take all of the TOEIC tests in order to gain a full understanding of the complete spectrum of their language proficiency skills" (p. 13.11). Hirai (2013), in his comparisons of TOEIC Total scores with direct tests of both speaking and writing, also found ETS's claims based on the standard TOEIC scores to overestimate abilities in the productive skills, stating that "Japanese people's business speaking/writing skills . . . are substantially lower than the levels the general public might expect of them from their TOEIC [Total] scores" (p. 124). These findings suggest the possibility that these other abilities may not be fostered as much as some claim by practicing only listening and reading.

Conclusion

Unfortunately, despite, or perhaps because of, its widespread usage, the TOEIC test is often used in ways that testing experts, even the makers of the TOEIC test, do not support. The TOEIC test can only measure general listening, reading, and overall English language proficiency, and only for those who are familiar with the settings, circumstances, and basic content of the test items. Although ETS has now added students as target users of the TOEIC test in its promotion materials, the contexts and contents of the test are still aimed at people who use English in international business situations. As a norm-referenced test, the TOEIC cannot determine what a student has learned in a class, what class materials a student knows and does not know, or what functions a student can or cannot perform well. Even if the scores are being used only to find out how students compare with each other, the standard errors of difference must be used along with the scores. In addition, the English abilities of Japanese college and university students on a given campus or in a given department are more similar to each other than desirable for TOEIC scores to be strongly reliable measures of English language ability.

Even if all of the above problems with TOEIC score usage were rectified, it is unreasonable and unfair for most Japanese colleges and universities to expect their students' TOEIC scores to increase during a semester or even a year because they do not offer nearly enough classroom hours of English language study for this to happen. It is depressing and demotivating for students and teachers when it appears that most students'

scores do not increase and many go down, due to not taking the standard errors of difference into consideration.

In any case, measurements of English proficiency are only estimates. TOEIC scores fool us into thinking otherwise and distract us from engaging in more beneficial practices and setting our sights on more useful goals. Using TOEIC scores for evaluative purposes, or even just requiring the TOEIC test to be taught or taken, has a very restrictive effect on what and how teachers teach and what and how students study and learn. Choice and motivation become connected with and distorted by the idea of increasing TOEIC scores rather than improving English language ability.

Quite opposite to what administrators might hope, using a test like the TOEIC in place of classroom-based tests "minimizes the possibilities that their program will look good" (Brown, 1995, p. 18). It also minimizes the possibilities that teachers and the students will look good. The wellsprings of teaching and learning are self-motivation and freedom. The limiting and conforming tendencies of using the TOEIC test in schools work against these impulses.

Bio Data

Brian Bresnihan teaches EFL at University of Hyogo. Before this, he did the same plus administrative and supervisory work in Temple University Japan's intensive English language program in Tokyo and at a small school that no longer exists in Hiroshima. Between those two positions, he spent 4 years studying at Teachers College, Columbia University (including 2 years coordinating the in-house ESL program) and teaching ESL part-time in a few programs in New York City. <brian@econ.u-hyogo.ac.jp>

References

Bresnihan, B. D. (2010). *Possible reliability problems affecting use of TOEIC IP Test scores*. Kobe: Institute for Policy Analysis and Social Innovation, University of Hyogo.

Brown, J. D. (1995). Differences between norm-referenced and criterion-referenced tests. In J. D. Brown & S. O. Yamashita (Eds.), *Language testing in Japan* (pp. 12-19). Tokyo: JALT.

Cameron, W. B. (1963). *Informal sociology: A casual introduction to sociological thinking*. New York: Random House.

CGI (The Chauncey Group International). (1998). *TOEIC technical manual*. Princeton, NJ: Author.

CGI (The Chauncey Group International). (2000). *TOEIC can-do guide: Linking TOEIC scores to activities performed using English*. Princeton, NJ: Author. Retrieved from http://www.ets.org/Media/Research/pdf/TOEIC_CAN_DO.pdf

Childs, M. (1995). Good and bad uses of TOEIC by Japanese companies. In J. D. Brown & S. O. Yamashita (Eds.), *Language testing in Japan* (pp. 12-19). Tokyo: JALT.

ETS (Educational Testing Service). (2007). *TOEIC user guide: Listening & reading*. Princeton, NJ: Author. Retrieved from http://www.ets.org/Media/Tests/Test_of_English_for_International_Communication/TOEIC_User_Gd.pdf

ETS (Educational Testing Service). (2012). *TOEIC examinee handbook: Listening & reading*. Princeton, NJ: Author. Retrieved from http://www.ets.org/Media/Tests/TOEIC/pdf/TOEIC_LR_examinee_handbook.pdf

ETS (Educational Testing Service). (n.d.). *TOEIC can-do levels table*. Princeton, NJ: Author. Retrieved from http://www.amideast.org/sites/default/files/otherfiles/hq/advising%20testing/toeic20can-do-20levels20table1.pdf

Gould, S. J. (1996). *The mismeasure of man* (Rev. ed.). New York, NY: W. W. Norton.

Hirai, M. (2013). Correlations between BULATS Speaking/Writing and TOEIC scores. In R. Chartrand, S. Crofts, & G. Brooks (Eds.), *Literacy: SIGnals of emergence. Proceedings of the 11th Annual JALT Pan-SIG Conference* (pp. 118-125). Hiroshima: Hiroshima University. Retrieved from http://www.pansig.org/2013/JALTPanSIG2013/Proceedings/The2012Pan-SIGProceedings.pdf

IIBC (The Institute for International Business Communication). (n.d.a). *TOEIC Bridge data & analysis 2011*. Tokyo: Author. Retrieved from http://www.toeic.or.jp/toeic_en/pdf/data/TOEIC_Bridge_DAA2011.pdf

IIBC (The Institute for International Business Communication). (n.d.b). *TOEIC Test data & analysis 2011*. Tokyo: Author. Retrieved from http://www.toeic.or.jp/toeic_en/pdf/data/TOEIC_DAA2011.pdf

Kubiszyn, T., & Borich, G. (1987). *Educational testing and measurement* (2nd ed.). Glenview, IL: Scott, Foresman.

Liao, C. W., Qu, Y., & Morgan, R. (2010). *The relationships of test scores measured by the TOEIC Listening and Reading test and TOEIC Speaking and Writing tests*. Princeton, NJ: ETS. Retrieved from http://www.ets.org/Media/Research/pdf/TC-10-13.pdf

Poundstone, W. (2003). *How would you move Mount Fujii? Microsoft's cult of the puzzle: How the world's smartest companies select the most creative thinkers*. New York: Little, Brown and Company.

PsyAsia (PsyAsia International Pte.). (2013a). *Standard error of difference*. Singapore: Author. Retrieved from http://www.psyasia.com/support/Knowledgebase/Article/View/28/0/standard-error-of-difference

PsyAsia (PsyAsia International Pte.). (2013b). *Standard error of measurement*. Singapore: Author. Retrieved from http://www.psyasia.com/support/Knowledgebase/Article/View/27/0/standard-error-of-measurement

Saegusa, Y. (1985). Prediction of English proficiency progress. *Musashino English and American Literature, 18*, 165-185. Tokyo: Musashino Women's University.

Wilson, K. (1989). *TOEIC research report, No. 1: Enhancing the interpretation of a norm-referenced second-language test through criterion referencing: A research assessment of experience in the TOEIC testing context*. Princeton, NJ: ETS. Retrieved from http://www.ets.org/Media/Research/pdf/RR-89-39.pdf

Wilson, K. (1993). *TOEIC research summaries, No. 1: Relating TOEIC scores to oral proficiency interview ratings*. Princeton, NJ: ETS. Retrieved from http://www.ets.org/Media/Research/pdf/TOEIC-RS-01.pdf

Wood, J. (2010). TOEIC materials and preparation questions: Interview with an ETS representative. *The Language Teacher, 34*(6), 41-45. Tokyo: JALT. Accessed from http://jalt-publications.org/tlt/articles/109-readers-forum-toeic-materials-and-preparation-questions-interview-ets-representativ

Woodford, P. (1982). *TOEIC research summaries: An introduction to TOEIC: The initial validity study*. Princeton, NJ: ETS. Retrieved from http://www1.ets.org/Media/Research/pdf/TOEIC-RS-00.pdf