Using TOEIC Scores to Evaluate Student Performance in English Language Courses

Brian D. Bresnihan

Abstract

This paper studies the TOEIC scores of about 1,200 students at a university in Japan and their use in determining those students' final grades in their required English courses. After giving a brief description of the TOEIC and the history of its development, some of its producer's claims concerning its worthiness and criticisms of these claims are considered. Next, the data used in this study is presented and compared with similar data from throughout Japan and from the initial studies of the TOEIC published by its producer. This is followed by comparisons between both average scores and individual scores of students who sat for two administrations of the TOEIC. Finally, reasons why TOEIC scores ought not to be the sole criterion for evaluating student performance in English language courses in schools are discussed.

1. Background Information Concerning the TOEIC

Nowadays, it seems that everyone has heard of the TOEIC, or at least everyone involved both in education and with students in and from non-English speaking cultures who need to study English, as well as many people in non-English speaking countries involved both in business and with the need to engage in work with people from other countries. A decade ago, not many people would have known what this acronym stood for, and almost no one outside of Japan or South Korea would have ever heard of it. Today, the Test of English for International Communication is administered in many countries and is one of the most sat for standardized English language tests in the world, though about 80% of those tests are taken in Japan or South Korea.[1] The perceived importance of this test in the field of English language testing is obvious as TOEIC scores are now included in conversion charts among various other standardized English test scores used in many countries.[2] This is another recent development and attests to the prestige and power the TOEIC

[1] This figure is a rounded estimate based on information from pages 3 and 7 of *TOEIC Newsletter, No. 105* and page 1 of *TOEIC Test Data & Analysis 2009*.

[2] Comparisons of TOEIC scores with other standardized test scores can be found in "The Cambridge, IELTS, TOEFL and TOEIC compared for equivalencies," page 2 of "Can-Do Levels Table," page 1 of "Can-Do Levels Table (2)," Taylor, "TOEFL Equivalency Table," and "TOEFL Equivalency Table (2)."

now holds in the world of English language testing. Although the government of South Korea is having its doubts about using the TOEIC any longer,[3] TOEIC scores are increasingly being recognized by greater numbers of organizations and schools in many countries.

Briefly, the TOEIC was created, and is still produced and controlled, by one of the largest standardized testing organizations in the world, the Educational Testing Service (ETS), which is also responsible for the Test of English as a Foreign Language (TOEFL).[4] The TOEFL was first administered in 1964. TOEFL scores are often used by schools of higher education in the United States to determine if non-native English speaking students have sufficient English abilities to be successful in their studies, research, and/or work at American universities.[5]

In 1977, a request to ETS for the creation of a different sort of English test was made by Japan's Ministry of International Trade and Industry (MITI), now the Ministry of Economy, Trade, and Industry (METI), and the Japan Federation of Economic Organizations, now the Federal Business Federation, based on the ideas of Yasuo Kitaoka. The desire was for a test of English that would measure English abilities in business-related contexts, would distinguish differences in English abilities at fairly low levels, and would include a number of native English speaker's dialects from various countries. ETS agreed to develop such a test and soon sent people to Japan to do exploratory research on the English requirements of non-native English speaking workers.[6] This is what ETS reports concerning their findings:

> "The studies were revealing. One important finding was that the language of non-native speakers clearly focuses on communication and is delivered with relatively few embellishments. For example, the least proficient person present out of necessity invariably determines the level of English used in meetings. Non-native English speakers use fewer idiomatic expressions. They employ technical terminology only when necessary. Furthermore, they tend to use fewer complex grammatical structures, even

[3] The dissatisfaction of the South Korean government with the TOEIC and how it may be replaced is discussed in Kang and Oh & Kang. Lee gives an earlier criticism of the use of the TOEIC and TOEFL in South Korea.

[4] Details concerning ETS and its tests can be found in "Educational Testing Service," "Tests & Products," and "Who We Are."

[5] Information concerning the TOEFL can be found in "About the TOEFL Test," "ETS Premieres World's First Internet-Based English-Proficiency Test," "Research and Design," and "Test Content (TOEFL)."

[6] More details about the initial negotiations prior to the creation of the TOEIC can be found in pages 14 to 16 of Bresnihan, page 8 of Chapman (2004), McCrostie (2009), page 2 of McCrostie (2010), page 6 and 18 of *TOEIC Newsletter, No. 105*, and page 2 of *TOEIC User Guide: Listening & Reading*.

> though the more capable speakers in the studies were capable of speaking quite impressively.
>
> The language specialists also noted that the business people seldom need to read very long narratives. Instead, the international business community receives much of its English-language exposure from letters, and memoranda, and other short texts."[7]

Based on this research and the design of the TOEFL, ETS quickly created the TOEIC for the following stated purpose:

> "The TOEIC test measures the everyday *English skills of people working in an international environment*. Test scores indicate how well *people can communicate in English with others in the global workplace*. The test does not require specialized knowledge or vocabulary beyond that *of a person who uses English in everyday work activities*."[8]

The phrases I have italicized in the above quotation indicate that the TOEIC is meant to measure the English abilities of adults whose work requires them to interact and communicate with others using English. It is clearly not intended for young people who have not yet entered the work force and do not have any work experience. To me, this indicates that it is inappropriate to use TOEIC scores as a major factor in, and certainly not the excusive means of, measuring the English language abilities of school-age children or even college and university students due to the differing contexts, settings, and experiences between workers in the international business world and students. This disadvantage will likely affect their scores in unpredictable ways due to factors not related to their English language proficiency.

The TOEIC was first administered in Japan in December of 1979 to a few thousand examinees. For some years, there were few test takers, and those responsible for the test in Japan were worried. However, two changes concerning the TOEIC occurred in the early 1980s that led to increases in the number of examinees sitting for the test.[9]

---

[7] This quotation is from page 2 of *TOEIC User Guide: Listening & Reading*.

[8] This quotation is on page II-1 of *TOEIC Technical Manual*. Italics are added.

[9] Details about the first few years of the TOEIC can be found in pages 17 to 20 of Bresnihan, McCrostie (2009), page 3 of McCrostie (2010), pages 2, 3, 6, and 7 of *TOEIC Newsletter, No. 105*, page 1 of *TOEIC Test Data & Analysis 2009*, and pages 5 and 6 of Woodford.

The TOEIC now has two versions, the TOEIC SP (Secure Program) Test and the TOEIC IP (Institutional Program) Test. The former is the original, and the latter was first administered in 1981.[10] It is less expensive for test takers to take the latter than the former, and scores are reported more quickly for the latter than the former. Also, the TOEIC IP Test is administered at, by, and when any institution wishes and makes arrangements to do so rather than according to the fixed TOEIC SP Test locations and schedule determined by ETS and its promoters and handlers. The downside of the TOEIC IP Test is that it is not as secure and, as it is produced from already-used TOEIC SP Tests, and because those who take the institutional program tests have been preselected by the organizations, the schools or companies, they belong to in some way, its results are not as statistically robust as those from the TOEIC SP Test.[11] The second change was that in January of 1982, the TOEIC was administered in South Korea for the first time.[12] The result of these two changes was that within a few years, there were a great many Japanese and Koreans taking the TOEIC. Currently, the worldwide total is about 5,000,000 examinees per year, and the test is offered in about 90 countries, though, as mentioned earlier, about 80% of the tests are taken in Japan or South Korea. Cumulatively, the TOEIC has been sat for about 20,000,000 times since the test was first administered.[13]

The TOEIC taken by most people has two sections of multiple choice questions with most answers chosen from four possible choices, though one part provides only three possible choices to choose from. The first section of 100 questions measures the ability to understand recorded spoken English scripted as if taking place in a variety of business-related situations, including conversations, announcements, and speeches. The second section of 100 questions measures the ability to understand various types of written texts, such as letters, advertisements, and reports, also related to business in some way. All of the discourse is meant to mimic language used in circumstances that workers might find themselves facing at, related to, or as a result of their jobs. The Listening section takes about 45 minutes, determined by the length of the recording. Test takers are given 75 minutes to complete the Reading section. In May of 2006, some changes were made to the TOEIC SP Test concerning the format of certain questions, and an increased number of

[10] The first administration of the TOEIC IP Test is noted in pages 2, 3, and 7 of *TOEIC Newsletter, No. 105* and page 1 of *TOEIC Test Data & Analysis 2009*.

[11] Explanations of the differences between the TOEIC SP and IP Tests can be found in page 8 of Chapman (2004), "Differences between SP group application and IP," and "Group Application."

[12] This date is given in pages 2 and 7 of *TOEIC Newsletter, No. 105*.

[13] Increases in and numbers of test takers can be found in pages 2, 3, 6, 7, 8, 18, and 19 of *TOEIC Newsletter, No. 105* and page 1 of *TOEIC Test Data & Analysis 2009*.

dialects of native-language English began to be used in the Listening section. These changes were implemented in the TOEIC IP Test in April of 2007. The details mentioned so far, however, remained the same. What follows includes these changes.

The Listening section has four parts and consists of a recording plus some of the questions and possible answers written in the test booklet. Part 1 contains ten questions. Each question consists of one printed photograph and four one-sentence descriptive oral statements, from which the description that best fits the photograph is to be chosen. Part 2 has 30 questions, each of which is made up of one oral statement or question followed by three oral responses. The response that would be best to use following the initial statement or question is to be chosen. Part 3 has 30 questions. There are ten short oral conversations in this part. Each short conversation has three written questions about it. Following each question, there are four possible written answers, from which to choose the best answer. Part 4 is made up of 30 questions based on 10 short oral talks, each of which has three associated questions. The questions, with four possible answers each, from which the best answer is to be chosen, are all written.

The Reading section has three parts. All of the texts, questions, and possible answers are contained in the test booklet. Part 5 has 40 questions. Each question is one written sentence with a word or phrase replaced by a blank. Four possible written words or phrases are given for each blank, from which the best grammatical and semantic fit is to be chosen. Part 6 consists of 12 questions based on four written texts. Each text contains a number of sentences of which three have a word or phrase replaced by a blank. Four possible written words or phrases are provided for each blank, from which to chose the one that fits the best grammatically and semantically. Part 7 contains 48 questions in two slightly different formats. The first 28 questions consist of seven to ten single written texts followed by two to four written questions each. Each question has four possible written answers from which the best answer is to be chosen. The final 20 questions are based on four pairs of written texts, each pair of which has five associated written questions. Each question is followed by four possible written answers, from which to choose the best.[14]

After taking the TOEIC, examinees receive three reported, or scaled, TOEIC scores; Total, Listening, and Reading. The possible scores vary from 10 to 990, from 5 to 495, and from 5 to 495, respectively. The reported Total score is the sum of the reported Listening score and reported

[14] Details concerning the format of and changes to the TOEIC discussed in this paragraph and the two preceding paragraphs are explained in pages 32 to 34 of Chapman & Newfields, "New TOEIC test premieres in Japan and Korea; all TOEIC versions are equally valid worldwide," page 4 of Powers, Kim, & Weng, "Sample (TOEIC)," "Test Content (TOEIC)," and pages 3 and 4 of *TOEIC User Guide: Listening & Reading*.

Reading score.[15] This is the explanation given by ETS for reporting these three scores and not each of the seven part scores separately:

> "The internal consistency score reliabilities for the seven parts of the TOEIC test have ranged from a low of 0.67 to a high of 0.87; these subpart scores are not reported to candidates because they are not sufficiently reliable for use in making decisions about candidates' English language abilities. In compliance with testing standards, scores with reliabilities over 0.90 are considered to be adequate for reporting and usage (cf. *Standards for Educational and Psychological Testing*, 1985)."[16]

The raw scores from the correct answers to the 100 Listening questions and from the correct answers to the 100 Reading questions are converted to the respective reported or scaled scores by a method called equating. There is no a simple linear conversion of number of correct answers to scaled scores. There are many forms of the TOEIC based on the one format. Forms, here, means the contents of the tests; their specific texts, questions, and possible answers. Each form has a different conversion scale based on the results of a large number of test takers' performances. Statistical analyses of these results are used to create a unique conversion scale for each form of the test that produces scores equivalent to the scores produced by all other conversion scales of the TOEIC. Through this method, ETS claims, scaled scores from various forms of the test are reporting examinees' performances on the various forms of the test equivalently. So, according to ETS, it does not matter which form of the test an examinee takes, as their English abilities will be measured in the same way and they will receive very much the same score regardless of the form. Test takers do not know which form of the test they sit for.[17]

In addition to the standard TOEIC that includes a Listening section and a Reading section, the TOEIC Speaking Test and the TOEIC Writing Test were introduced in Japan and South Korea in December of 2006 and are now available elsewhere. They are administered in one sitting, with the former taking 20 minutes and the latter taking 60 minutes.[18] There is also a test similar to the

---

15 These scores ranges are mentioned in page 5 of Woodford and pages I-1 and II-3 of *TOEIC Technical Manual*.

16 This quotation is from page IV-2 of *TOEIC Technical Manual.*

17 Equating is explained in page 10 of Chapman (2004), "Frequently Asked Questions About the TOEIC Listening and Reading Test," and pages II-4 to II-5 of *TOEIC Technical Manual*.

18 Information about the TOEIC Speaking and Writing can be found in "About the TOEIC Speaking and Writing Tests," "ETS Europe UK Launches new TOEIC Speaking and Writing tests May 22, 2008," "Registration Opens in Japan for

TOEIC for examinees with lower English abilities called the TOEIC Bridge, which was first administered in Japan and South Korea in November of 2001. It contains two sections of 50 questions each. The Listening section comes first and lasts about 25 minutes, according to the length of the recording. This is followed by the Reading section, which takes 35 minutes. All questions are multiple choice with four possible answers given for most questions. Some of the Listening questions only have three possible answers to choose from.[19]

2. What ETS Claims about the TOEIC

Though ETS has not published nearly as many studies on the results of TOEIC administrations as it has of results from some of its other tests, such as the TOEFL,[20] these published studies indicate that the TOEIC is a valid and reliable test of English language proficiency. A test is said to have strong validity if measures what its makers and users claim it measures. A test is said to be highly reliable if it measures consistently across various administrations and versions of the test. ETS carried out its first validity study for the TOEIC using scores from the first administration correlated with the scores from subsequent tests taken by some of those first examinees chosen, and then retested, in the following manner:

> "When score distributions were obtained for the first administration of TOEIC, 500 examinees were selected to take TOEFL. The 500 were selected on the basis of their scores on the TOEIC. One hundred examinees were selected at each of five approximate score levels: 950, 765, 580, 315, 45. A smaller group of 20 examinees from each group of 100 was selected. To these examinees a series of direct measures of language ability were administered."[21]

TOEIC Speaking and Writing Tests," "South Koreans Take First TOEIC Speaking and Writing Tests," and TOEIC Speaking and Writing Tests Launched in the UK.

19 Information about the TOEIC Bridge can be found in "About the TOEIC Bridge" and "Sample (TOEIC Bridge)."

20 Chapman (2003) writes on page 2, "ETS has released 69 research reports into TOEFL, with an additional 17 technical reports on this exam between 1977 and 2002. However, for the TOEIC there are only three full research reports. In addition there was an initial validity study in 1982 and one technical manual." The same details are reported in Chapman (2005). Based on a similar search, Bresnihan writes on page 18, "From the ETS TOEFL website, under TOEFL Research, on March 3, 2010, the present author was able to gain access to 197 research and technical reports and summaries and 17 data reports that could be downloaded and 1 explanation of a mapping of the TOEFL and the Common European Framework of Reference (CEFR), of which a study could be requested. From the ETS TOEIC website, under Research, on the same day, the present author was able to find 15 research reports and summaries, 8 of which were classified as TOEIC-related, and 2 publications, 1 of which led to a report, that could be downloaded. The *TOEIC Technical Manual* was not one of them, but it was still available online."

21 This quotation is from page 9 of Woodford.

> "The examinees who underwent the direct measures were divided into five groups for the purpose of analysis. Examinees were grouped according to their part scores on the TOEIC. For both Listening and Reading, Group I had TOEIC part scores below or equal to 100; Group II had TOEIC part scores between 100 and 205; Group III had TOEIC part scores between 205 and 300; Group IV had scores between 305 and 400; and Group V had scores at 405 or above."[22]

For those examinees who took the TOEIC and another English proficiency test as part of the just mentioned analyses concerning validity, ETS reports that the TOEIC Listening section scores and scores on another test of listening correlated "very highly--0.90," and the TOEIC Reading section scores showed a "high degree of similarity" with scores on another reading test, "(t)he correlation between the two . . . (being) 0.79." Also, the relationship between the TOEIC Listening section scores and scores on a direct test of speaking was reported as correlating at "0.83 . . . (a) high degree of correlation," and the TOEIC Reading section scores and scores on a direct test of writing "correlated 0.83," which was considered a "high correlation."[23]

In the same initial study, correlations concerning estimated reliability internal to the TOEIC itself were also reported. For the Listening section scores it was 0.916, for the Reading section scores it was 0.930, and for the Total scores it was 0.956. "These reliabilities are well within the generally accepted limits for measurement of individual achievement." It was also mentioned that "(t)he correlation between the sections was 0.769 . . . . This would indicate that each score provides somewhat different information about the examinee and justifies reporting separate scores."[24]

Concerning validity, the *TOEIC Technical Manual* reports all of the above validity figures along with others from a number of additional studies. There are correlations between TOEIC Listening section scores and scores on other tests of listening and scores on tests of speaking and between TOEIC Reading section scores and scores on other tests of Reading and scores on tests of writing. All of these correlations are reported to be above 0.65, with most being above 0.75, and all are judged there to indicate strong or very strong relationships between the two measures in each

[22] This quotation is on page 12 of Woodford.
[23] These explanations and quotations can be found in pages 12 to 15 of Woodford.
[24] These figures and quotations are on page 8 of Woodford.

case.[25]

Concerning reliability, the *TOEIC Technical Manual* does not explain what studies were used to generate the reported correlations. However, the figures are very similar to those reported in the initial study referred to just above. The estimated reliabilities of the Listening section scores are reported to be from 0.91 to 0.93. Of the Reading section scores, they are reported to be from 0.92 to 0.93. The Total scores' estimated reliabilities are reported to vary from 0.95 to 0.96.[26]

Three other important measures reported in the *TOEIC Technical Manual* are the standard errors of measurement, the conditional standard errors of measurement, and the standard errors of difference of the reported scores for each section. These standard errors attempt to compensate for fluctuations in each test taker's scores that might occur due to factors unrelated to the test taker's true abilities supposedly being measured by the test. The expected possible difference between a test taker's actual test score and her/his supposed average test score or true ability as measured by the test is the standard error of measurement. For each of the reported scores of the two sections of the TOEIC, the standard error of measurement is about +/-25 scaled points. So, a test taker's true Listening score is expected to be between her/his reported Listening score plus 25 points and her/her reported score minus 25 points, with about 68% certainty. The situation is the same for her/his true Reading score. (Perhaps the test taker's true Total score would be between her/his reported Total score plus 50 points and her/his reported Total score minus 50 points, with about 68% certainty, but this was not mentioned.) To find the probable range within which a test taker's scores would fall with 95% certainty, then the standard errors of measurement would need to be almost doubled, to +/-49 scaled points for the Listening score and for the Reading score (and perhaps to +/-98 scaled points for the Total score). The conditional standard errors of measurement are more precise standard errors of measurement for each of the different scores within the whole range of scores on a given test administration. The conditional standard errors of measurement for a score are smaller the further the score is from the average of all the scores.[27]

The amount of difference needed between two test takers' scores, or between two scores by the same test taker from two different sittings of the test, to show a real difference in the scores with about 68% certainty, is called the standard error of difference. For each section of the TOEIC, it is about +/-35 scaled points. (Again, perhaps for the Total score it is about +/-70 scaled points, but this was not stated.) If one wanted to be more certain, 95% certain, if a difference in scores

[25] These correlations and judgments are on pages III-1 to III-4 of *TOEIC Technical Manual*.
[26] These correlations can be found in pages IV-1 to IV-2 of *TOEIC Technical Manual*.
[27] These figures and explanations can be found in pages IV-4 to IV-6 of *TOEIC Technical Manual*.

existed, then these figures would need to be nearly doubled, to +/-69 scaled points (and perhaps to +/-138 scaled points).[28]

Another important point made in the *TOEIC Technical Manual* is that "(f)or groups in which there is a great deal of homogeneity (for example, when candidates are pre-selected . . .), reliability estimates will be lower."[29] "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."[30] This would seem to be an important consideration for many colleges and universities in Japan as their students are selected, at least partially, based on their scores on English language examinations.

3. Doubts Concerning Claims by ETS about the TOEIC

Some researchers have disputed the claims made by ETS, concerning TOEIC scores and their usage, in published studies, which often use the TOEIC scores of workers in Japan participating in company-organized English language classes. Chapman pointed out a few inconsistencies in the first published study of the TOEIC. As was mentioned earlier, in that study, Woodford claimed that the correlation of 0.79 that he found between TOEIC Reading section scores and the scores on another direct test of reading showed a "high degree of similarity of performance . . . (and) provides a good indication of the examinee's ability to read English with understanding."[31] Before this, Woodford stated that the correlation he found between the Listening and Reading sections of 0.769 "indicate(s) that each score provides somewhat different information about the examinee."[32] Chapman stated that "(t)here is a clear inconsistency in the way Woodford is interpreting the results of the study," of his correlations of 0.79 and 0.769, and "(i)t is difficult to see how the claim that the two tests of reading show a high similarity of performance can be supported" based on a correlation of 0.79.[33]

Chapman also questioned certain interpretations in another ETS study done by Wilson in 1989. In that study, correlations between the Listening section scores and scores on a standardized interview test were found to be "typically in the mid-70's."[34] In a very labored discussion over

[28] These figures and explanations are from pages IV-6 to IV-7 of *TOEIC Technical Manual*.
[29] This quotation is on page IV-2 of *TOEIC Technical Manual*.
[30] This quotation is from page IV-3 of *TOEIC Technical Manual*.
[31] This figure and quotation are on page 13 of Woodford.
[32] This figure and quotation are from page 8 of Woodford.
[33] These quotations are on page 4 of Chapman (2006).
[34] This quotation is on page 46 of Wilson (1989).

these findings, Wilson seems to be trying to urge the reader to view his findings as supporting the idea that the TOEIC Listening section scores can be used to infer English speaking ability. We find the following statements concerning the relationship between the two scores:

> "a consistent pattern of concurrent correlation"
>
> "a strong underlying functional linkage"
>
> "examinees with relatively high (low) average levels of TOEIC-assessed ability to comprehend spoken English may be expected to perform relatively well (poorly) in the interview situation"
>
> "the evidence that has been reviewed suggests strongly that the ability to comprehend and produce utterances in English is to some extent "dependent," directly and functionally, upon the ability to comprehend spoken English. Accordingly, it follows logically that level of ability to use English in face-to-face conversation . . . is likely to vary relatively consistently with level of developed English-language listening comprehension"
>
> "likely to be relatively consistent"[35]

Chapman interprets these finding differently and takes a different stance stating, "This report by Wilson (1989) seems to indicate that a separate speaking test and the TOEIC will provide different information about examinees. . . . (T)o test the ability . . . to speak English, employing the TOEIC test in isolation is unlikely to be the most accurate method available."[36] This is the point of view it seems one would take based on the explanation in the *TOEIC Technical Manual* noted earlier concerning the reporting of TOEIC scores.[37] Here is a slightly more detailed explanation by a professor of educational research about the same issue:

---

35 In order, these five quotations are on pages 46, 47, 47, 48, and 48, respectively, of Wilson (1989).

36 This quotation is on page 76 of Chapman (2003).

37 This refers to the statement that "scores with reliabilities over 0.90 are considered to be adequate for reporting and usage" and that those with correlations lower than this are not, which is on page IV-2 of *TOEIC Technical Manual*.

> "The *Standards for Educational and Psychological Testing* provides direction for test score reporting and usage in the credentialing of persons in many occupations and professions (1999). . . . Important testing information including reliability coefficients are useful in comparing scores from these different tests, but interpretation allowances must be made for the variability of scores from different samples of examinees, administration techniques from which the reliability coefficients were obtained, the source(s) of error indicated by the reliability coefficient, the number of items on the test, and the length of time allowed for testing. Nunnally and Bernstein (1994) provided guidance in the interpretation of the reliability coefficient by stating that a value of .70 is sufficient for early stages of research, but that basic research should require test scores to have a reliability coefficient of .80 or higher. When important decisions are to be made with test scores, a reliability coefficient of .90 is the minimum with .95 or higher a desirable standard."[38]

Childs questioned the internal reliability of TOEIC scores. In his study of company workers who were studying English, he estimated raw scores for the reported scores he had based on information given by ETS about one form or version of the test.[39] Although this does not lead to the most accurate findings, the results are likely very close to the best that could be determined. Using the same reliability formula as ETS, his data produced a Total score correlation coefficient of 0.57,[40] far lower than that reported by ETS, which is 0.95 to 0.96.[41] He also reported his findings concerning score gains over three TOEIC administrations: About one third of the subjects' scores increased twice, about two thirds increased once and decreased once, and a few subjects' scores decreased twice.[42] In addition, he showed that there were great differences between many examinees' estimated scores, which were based on previous scores, mean scores, and mean changes in scores, and their actual scores.[43] Based on all of this, Childs concluded that "jumping around is the nature of TOEIC scores. . . . The fact is simply that TOEIC . . . is not the best gauge of

[38] This quotation is from Schumacker.
[39] This is explained on page 69 of Childs.
[40] This figure is on page 69 of Childs.
[41] These figures are on page IV-2 of *TOEIC Technical Manual*.
[42] These figures are on page 71 of Childs.
[43] This is detailed on page 70 of Childs.

individual learning. . . . The use of TOEIC for gauging individual learning is, in general, ineffective or wrong. . . . (T)est-to-test differences will display very great variability."[44]

Hirai carried out research into the relationships between TOEIC Listening scores and speaking test scores, between TOEIC Reading scores and writing test scores, and between TOEIC Total scores and both speaking test scores and writing test scores. His data generated a correlation between TOEIC Listening scores and speaking test scores of 0.74.[45] ETS reports such correlations to be 0.83, 0.74, and 0.75.[46] The correlation Hirai found between TOEIC Reading scores and writing test scores was 0.59.[47] Such correlations are reported by ETS to be 0.83.[48] Correlations between TOEIC Total scores and speaking test scores were found by Hirai to be 0.78 and 0.66.[49] ETS reports correlations of 0.74, 0.76, and 0.73 for such pairings.[50] Hirai also reports correlations in his data between TOEIC Total scores and writing test scores of 0.66 and 0.69.[51] ETS gives no such correlations in the reports referred to in the present study. So, Hirai's data produced correlations which are generally just a little lower than those reported by ETS, except for those of 0.83 from the initial study.

However, Hirai doubted the claims made by ETS that their correlations between TOEIC Listening scores and scores on tests of speaking and between TOEIC Reading scores and scores on tests of writing were high enough to make accurate predictions about the test takers' capabilities in the abilities not tested. "While the correlation coefficient is a general indicator of how closely two quantities relate to each other, one should be cautious about the potential pitfall of predicting the value of one quantity . . . from that of the other . . . on the basis of the correlation coefficient, unless it is extremely close to +/-1."[52] He also noted that the highest of these correlations were from early studies which "may tend to have an inherent bias toward collecting higher scores. . . . As a result, the data collected in such studies tends to be skewed toward the high end, which effectively increases the correlation coefficient."[53]

---

44 The first two of these quoted phrases and sentences are on page 73 of Childs, and the second two are on page 74 of Childs.

45 This figure is from page 2 of Hirai (2002).

46 These figures are from page 14 of Woodford, page 40 of Wilson (1989) and page 9 of Wilson (1993) and pages I-2 and III-2 of *TOEIC Technical Manual,* and page 6 of Wilson (1993), respectively.

47 This figure is on page 5 of Hirai (2002).

48 This figure is on page 15 of Woodford and pages I-2 and III-4 of *TOEIC Technical Manual.*

49 These figures are from page 2 of Hirai (2002) and page 17 of Hirai (2009), respectively.

50 These figures are on page 40 of Wilson (1989) and page 9 of Wilson (1993) and pages I-2 and III-2 of *TOEIC Technical Manual,* page 6 of Wilson (1993), and page 9 of Wilson (1993), respectively.

51 These figures are on page 5 of Hirai (2002) and page 38 of Hirai (2008), respectively.

52 This quotation is on page 7 of Hirai (2002).

53 This quotation is from pages 13 and 14 of Hirai (2009).

Furthermore, Hirai followed the procedure described by Woodford of finding correlations between only specific TOEIC scores and scores on other measures,[54] which was explained earlier. He did this using both the TOEIC scores chosen by Woodford and other scores. In all cases, his correlation coefficients increased.[55] He also tried finding correlations between segments of his data sets as determined by TOEIC scores and correlations on other measures. In these cases, all of his correlations decreased.[56] The latter procedure, however, somewhat replicates the real-world situation in which companies and schools have groups of their employees or students, respectively, take the TOEIC. In such cases, it is probably never the case that there is anywhere near as much variation in scores on a given TOEIC IP Test administration as on the administrations used in the studies by ETS or on a public TOEIC SP Test administration. Therefore, correlations from any TOEIC IP Test administration would be lower than those reported by ETS, which means the reliability of the TOEIC scores would be weaker, too.

In a prior study, the present author investigated TOEIC IP Test scores taken six months apart by first-year university students, who were not English majors and were taking three 90-minute English classes per week during each semester, in one department at a university in Japan. What was revealed was that evidence of overall language proficiency, which would be demonstrated by consistency in scores, could not be demonstrated through correlations between those students' TOEIC Listening scores and TOEIC Reading scores grouped in many different ways. The correlations were lower, often much lower, than would be expected to support the idea that something in common was the basis of the scores. Between Listening scores and Reading scores on the same test administrations, these correlations were all between 0.35 and 0.53, inclusive.[57] In the initial study by ETS, the correlation between the Listening and Reading scores was reported to be 0.769.[58]

The present author also did not find as strong relationships between the Listening scores on the first and second test administrations and between the Reading scores on the first and second administrations as would be expected considering the relationships ETS studies have found between TOEIC Listening scores and other measures of Listening and between TOEIC Reading scores and other measures of Reading. They all fell in or near the lower end of what ETS reports. All of the correlations between the two Listening scores and between the two Readings scores in this author's

54 This procedure is explained on page 12 of Woodford.
55 This procedure is detailed and the results are reported on page 18 of Hirai (2009).
56 This procedure is detailed and the results are reported on pages 3 and 4 of Hirai (2002), pages 38 and 39 of Hirai (2008), and pages 17 and 18 of Hirai (2009).
57 These figures are on pages 92, 93, 96, 121, and 123 of Bresnihan.
58 This figure is from page 8 of Woodford.

earlier study were between 0.60 and 0.72, inclusive.[59] The *TOEIC Technical Manual*, which includes the results of the initial study, presents correlations between TOEIC Listening scores and other listening measures of 0.67 to 0.92 and correlations between TOEIC Reading scores and other reading measures of 0.73 to 0.87.[60]

The Listening scores and the Reading scores in each of the ten test administrations of the present author's earlier study were also normally distributed with the means and medians never deviating by more than 5.8 points, but usually less than half that amount.[61] And, the numbers of Listening scores and the numbers of Reading scores increasing and decreasing, on the second testing from the first testing for all of the scores from the ten test administrations combined, were fairly similar. Of the Listening scores, about 54% increased, 5% remained the same, and 42% decreased. Of the Reading scores, about 53% increased, 5% were unchanged, and 43% decreased.[62] This author interpreted these two analyses along with the just mentioned correlations between the Listening scores and Reading scores on the first test sitting and the Listening scores and Reading scores on the second test sitting, respectively, to be evidence of regressions to the mean, which is commonly a result of guessing at answers to questions on a test. This was taken to be an indication that the test was likely too difficult for these students and, therefore, called into question the ability of these scores to have been reliable evaluations of these students' English language abilities.[63]

4. Materials, Procedures, and Purposes

The TOEIC IP Test scores used in this study were from tests taken by first-year students in one department at a university in Japan.[64] The students were not English majors. They were all enrolled in three distinct mandatory English courses, each of which met for 90 minutes once a week throughout both 15-week semesters. One course emphasized reading, with some discussion; one course emphasized listening, with some speaking; and one course emphasized grammar, with some writing. There were eight sections of each course with about 25 students in each section. Placement in sections was done primarily in student identification number order.

At the end of approximately the eleventh week of each semester, the students were required to take the TOEIC IP Test administered at the school. Without doing so, a student could not pass

[59] These figures are on pages 92, 93, 98, 121, and 122 of Bresnihan.
[60] These figures are on pages III-2 to III-4 of *TOEIC Technical Manual*.
[61] These figures are from pages 56, 60, 105, 106, and 109 of Bresnihan.
[62] These figures are on page 171 of Bresnihan.
[63] This point is explained and elaborated on pages 86, 88, 98 to 102, 120, 124, and 216 of Bresnihan.
[64] A large portion of these scores was also used by the present author in an earlier study, Bresnihan.

any of the three compulsory English courses, and each student's TOEIC Total score was used in determining her/his final grade. (See Appendix A for details.) Between the two semesters, there was the two and a half months' long summer vacation. So basically, between the two tests, there were 4 weeks of classes followed by 11 weeks without classes, and then another 11 weeks of classes. Students needed to attend at least two thirds of a course's classes in order to be eligible to pass it.

The approximately 2,400 TOEIC IP Test scores used in this study were achieved by about 1,200 students, who took the test twice a year, about 200 students per year, over a six-year period, from 2004 to 2009. Approximately half of the tests were taken before the first changes to the TOEIC were made in 2007, and about half were taken after. There were few, if any, changes in the teachers who taught the three mandatory English courses. Each teacher taught the same group(s) of students, the same section(s), for both semesters. None of the students were taught by the same teacher for two different courses.

Basic statistics were generated using Microsoft Excel 2004 for Macintosh. One-way analyses of variance were carried out using JMP 5.0 for Macintosh.[65] Other statistics related to effect size were calculated on line.[66] Any slight discrepancies among figures within or among tables are due to rounding.

A number of issues are reported on concerning this data. First an initial viewing of various ranges of the scores are presented and are compared with TOEIC scores from the general population, with what is possible for TOEIC test scores, and with the scores reported by ETS in its initial study. Then, comparisons are made between the average TOEIC scores achieved in this study on the first administration and the second administration, followed by similar comparisons for each individual student's scores. Lastly, comments are made concerning the use of TOEIC scores in courses of English language study.

5. Comparisons of These Scores with the Scores of Other University Students in Japan and the Scores Used in the Initial ETS Study

Table 1 shows the average TOEIC IP (not SP) Test Total, Listening, and Reading scores achieved by all undergraduate students who sat for the test in Japan from 2004 to 2009, as well as the number of examinees. The lowest average Total score is 425 in 2004 and the highest is 439 in 2009, a spread of 15 points. The lowest average Listening score is 242 in 2004 and 2008 and the

[65] The one-way analyses of variance were run for me by Michael Redfield. I am grateful for this and for his help in understanding the results of these analyses.

[66] Cohen's d and correlation coefficient figures were calculated using "Effect Size Calculators" and "Effect Size Calculators (2)."

highest is 251 in 2005, the range being 10 points. The lowest average Reading score is 181 in 2006 and the highest is 195 in 2009, a variation of 15 points.

Table 1
Average TOEIC IP Test Scores of All Undergraduate Students in Japan: 2004 to 2009[67]

	2004 n=214,741	2005 n=243,286	2006 n=271,857	2007 n=300,511	2008 n=304,906	2009 n=309,311
Total	425	435	428	431	430	439
Listening	242	251	247	245	242	244
Reading	183	184	181	186	188	195

Table 2 gives the average TOEIC IP (not SP) Test Total, Listening, and Reading scores achieved by only first-year undergraduate students who took the test in Japan from 2004 to 2009, and the numbers of test takers. The spread in average Total scores is from 387 in 2004 to 412 in 2009, 26 points. The lowest average Listening score is 222 in 2004 and the highest is 233 in 2006, a range of 12 points. The lowest average Reading score is 165 in 2004 and the highest is 184 in 2009, a variation of 20 points.

Table 2
Average TOEIC IP Test Scores of All First-Year Undergraduate Students in Japan: 2004 to 2009[68]

	2004 n=91,853	2005 n=108,636	2006 n=132,470	2007 n=146,901	2008 n=148,772	2009 n=152,937
Total	387	401	401	406	405	412
Listening	222	232	233	232	227	228
Reading	165	169	168	174	178	184

Table 3 shows the average TOEIC IP (not SP) Test Total, Listening, and Reading scores achieved by only students whose major was similar to that of the students whose scores are being used in this study and who took the test in Japan from 2004 to 2009, and the numbers of students these were. The lowest average Total score is 408 in 2007 and the highest is 433 in 2005, a

[67] These figures are from page 8 of *TOEIC Test Data & Analysis 2004*, page 8 of *TOEIC Test Data & Analysis 2005*, page 8 of *TOEIC Test Data & Analysis 2006*, page 9 of *TOEIC Test Data & Analysis 2007*, page 9 of *TOEIC Test Data & Analysis 2008*, and page 9 of *TOEIC Test Data & Analysis 2009*.

[68] These figures are from page 8 of *TOEIC Test Data & Analysis 2004*, page 8 of *TOEIC Test Data & Analysis 2005*, page 8 of *TOEIC Test Data & Analysis 2006*, page 9 of *TOEIC Test Data & Analysis 2007*, page 9 of *TOEIC Test Data & Analysis 2008*, and page 9 of *TOEIC Test Data & Analysis 2009*.

variation of 26 points. The lowest average Listening score is 230 in 2008 and the highest is 246 in 2005, a spread of 17 points. The lowest average Reading score is 177 in 2007 and the highest is 191 in 2009, a range of 15 points.

Table 3
Average TOEIC IP Test Scores of All Students in Japan Whose Major Was Similar to That of the Students Whose Scores Are Being Used in This Study: 2004 to 2009[69]

	2004 n=52,891	2005 n=56,240	2006 n=63,138	2007 n=60,046	2008 n=60,953	2009 n=61,779
Total	418	433	425	408	412	424
Listening	235	246	242	231	230	233
Reading	183	187	183	177	182	191

Table 4 displays the average TOEIC IP Test Total, Listening, and Reading scores on each administration of the test achieved by the students whose scores are being used in this study, and the numbers of test takers. The lowest average Total score is on the second test in 2005, 435, and the highest is on the second test in 2009, 491. They vary by 57 points. Only in 2005 is the average Total score lower on the second test than on the first, and both of the average Total scores in 2005 are noticeably lower than all others. The next lowest average Total score is 461 on the first test in 2007, varying by only 31 points from the highest. The average Listening scores are spread from 242 on the second test in 2005 to 269 on the second test in both 2006 and 2008, a 28-point

Table 4
Average TOEIC IP Test Scores on Each Adminstration of All Students Whose Scores Are Being Used in This Study

	2004 n1=213 n2=204	2005 n1=207 n2=200	2006 n1=207 n2=205	2007 n1=224 n2=218	2008 n1=210 n2=206	2009 n1=199 n2=194
Total 1	462	446	469	461	466	479
Total 2	472	435	479	476	484	491
Listening1	249	252	264	255	244	243
Listening2	247	242	269	257	269	263
Reading1	212	194	205	206	222	236
Reading2	225	194	210	219	216	228

[69] These figures are from page 8 of *TOEIC Test Data & Analysis 2004*, page 8 of *TOEIC Test Data & Analysis 2005*, page 8 of *TOEIC Test Data & Analysis 2006*, page 9 of *TOEIC Test Data & Analysis 2007*, page 9 of *TOEIC Test Data & Analysis 2008*, and page 9 of *TOEIC Test Data & Analysis 2009*.

difference. The lowest average Reading score is 194 on both tests in 2005 and the highest is 236 on the first test in 2009, a range of 43 points.

Whether comparing the averages of the TOEIC IP Test scores being used in this study to those for the same years of all undergraduate students in Japan, of all first-year undergraduate students in Japan, or of all students in Japan whose major was similar to that of the students whose scores are being used in this study, the averages of the TOEIC IP Test scores being used in this study are somewhat higher than those for students throughout the whole of Japan. The two lowest average Total scores of the scores being used in this study are 435 and 446 on the second and first administrations in 2005, respectively. The next lowest average Total score is 461 on the first test in 2007 and the highest is 491 on the second test in 2009. The highest average Total score of all undergraduate students in Japan is 439 in 2009, of all first-year students in Japan is 412 in 2009, and of all students in Japan whose major is similar to that of the students whose scores are being used in this study is 433 in 2005. These are all clearly lower than the highest of the average Total scores in this study, 491. The lowest average Total score of all undergraduate students in Japan is 425 in 2004, of all first-year students in Japan is 387 in 2004, and of all students in Japan whose major is similar to that of the students whose scores are being used in this study is 408 in 2007. These are all lower than the lowest of the average Total scores in this study, 435, and much lower than lowest achieved outside of 2005, 461.

The lowest average Listening score of the scores being used in this study is 242 on the second test in 2005 and the highest is 269 on the second test in 2006 and in 2008. The lowest average Listening score of all undergraduate students in Japan is also 242 in 2004 and in 2008, of all first-year students in Japan is 222 in 2004, and of all students in Japan whose major was similar to that of the students whose scores are being used in this study is 230 in 2008. Although the first of these scores is the same as lowest of the average Listening scores in this study, 242, the latter two are noticeably lower. The highest average Listening score of all undergraduate students in Japan is 251 in 2005, of all first-year students in Japan is 244 in 2009, and of all students in Japan whose major was similar to that of the students whose scores are being used in this study is 246 in 2005. These are all lower than the highest of the average Listening scores in this study, 269.

The lowest average Reading score of the scores being used in this study is 194 on the first and second tests in 2005 and the highest is 236 on the first test in 2009. The lowest average Reading score of all undergraduate students in Japan is 181 in 2006, of all first-year students in Japan is 165 in 2006, and of all students in Japan whose major was similar to that of the students

whose scores are being used in this study is 177 in 2007. These are all lower than the lowest average Reading score in this study, 194. The highest average Reading score of all undergraduate students in Japan is 195 in 2009, of all first-year students in Japan is 184 in 2009, and of all students in Japan whose major was similar to that of the students whose scores are being used in this study is 191 in 2009. These are all far below the highest average Reading score in this study, 236.

Based on these first four tables and comparisons, it appears that the scores achieved by the students in this study are higher than those of similar students throughout Japan, on average. The lowest average Total score in this study is 435 while the highest of those in the other three categories of students from throughout Japan is 439. The highest in this study is 491. The lowest average Listening score in this study is 242 and the highest is 269. The highest of those in the other three categories is 246. The lowest average Reading score in this study is 194 while the highest of those in the other three categories is 195. The highest in this study is 236. In the initial study reported by ETS, upon which the TOEIC's scoring system, explained earlier, is based, the average Total score was 578, the average Listening score was 290, and the average Reading scores was 288.[70] These average scores are quite a bit higher than those in this study and those of undergraduate and first-year students in Japan and of students in Japan whose major is similar to those of the students whose scores are being used in this study.

The initial study also reports, for the scores being used there, that the Listening scores varied from 40 to 495, with about 68% falling between 200 and 370 and with a mean, or average, of 290, and that the Reading scores varied from 5 to 455, with about 70% between 210 and 385 and with a mean of 288. About 68% of the Total scores fell between 400 and 745, with the mean being 578, but no overall range of variation was given for the Total scores.[71] It was not mentioned, but the present author assumes that these ranges of about 68% to 70% of the scores are around the respective means or medians. Tables 5 and 6 include similar information concerning the scores being used in the present study.

In Table 5, we find that the lower ends of the ranges of variation for about two thirds of the Total scores around the mean/median in this study are usually a bit lower than that from the initial study published by ETS, being between 380 and 390 as compared to 400. Two are much lower, 355 and 365 in 2005, and two are a little higher, 410 and 415 in 2009. The higher ends of the ranges of variation for about two thirds of the Total scores around the mean/median in this study are

[70] These figures are on page 9 of Woodford.
[71] These figures are on page 9 of Woodford.

a great deal lower than that reported in the initial study. The former are between 525 and 585 while the latter is 745.

Table 5
Ranges of Variations in Scores Being Used in This Study for Two Thirds of the Scores Surrounding the Means/Medians

Year	Total1	%	Total2	%	Listen1	%	Listen2	%	Read1	%	Read2	%
2004	380-555	68.1	385-565	69.1	200-290	70.9	210-295	68.6	165-265	69.0	170-280	70.1
2005	365-525	68.6	355-525	70.5	205-295	69.6	195-285	70.5	140-240	68.6	140-245	70.0
2006	390-540	68.1	385-565	69.8	215-305	68.1	215-320	67.8	140-245	69.1	160-255	68.8
2007	385-550	70.1	385-545	68.8	210-300	69.6	215-300	70.6	150-250	69.6	170-260	68.8
2008	390-530	69.0	395-575	69.4	210-290	68.6	230-315	70.3	170-270	68.6	165-275	68.4
2009	415-550	69.3	410-585	69.6	210-275	69.8	225-305	68.6	190-290	71.3	175-270	68.6

The lower ends of the ranges of variation for about two thirds of the Listening scores around the mean/median in this study are usually about the same as that in the initial study. From the present study, most of the lower ends are between 195 and 215 and the ones in 2008 and 2009 are 230 and 225, respectively, while it is 200 in the initial study. The higher ends of the ranges of variation for about two thirds of the Listening scores around the mean/median in this study are between 275 on the first test in 2009 and 320 on the second test in 2006. In the initial study, it is 370, much higher than in the present study.

The lower ends of the ranges of variation for about two thirds of the Reading scores around the mean/median in this study are usually rather lower than that in the initial study. In this study, the lower ends are between 140 and 175, except on the first test in 2009 when it is 190. It is 210 in the initial study. The higher ends of the ranges of variation for about two thirds of the Reading scores around the mean/median in this study are very much lower than that in the initial study, being between 240 on the first test in 2005 and 290 on the first test in 2009 while it is 385 in the initial study.

This data indicates that a great many of the scores being used in this study are noticeably lower than most of those in the initial study, as determined especially by the upper limits of the variations in scores for about the middle 68% to 71% of the students. Table 5 also shows that the majority of the scores occur in much more restricted ranges than the scores in the initial study. In the initial study, the lower and upper scores of about the middle two thirds of the Total scores vary by 345 points, while the variation is between 135 and 180 points for the scores being used in this study. For the Listening scores in the initial study, this variation is 170, but it is only 65 to 105

points for the scores being used in this study. About the middle two thirds of the initial study's Reading scores vary by 175 points from lowest to highest, yet they vary by only 95 to 110 points for the scores being used in this study. Table 6 continues this inquiry, giving the basic statistics from the TOEIC scores of the 12 administrations being used in this study.[72]

Table 6
Maximums, Minimums, Means, Medians, Standard Deviations, and Number of Scores Greater Than 3 Standard Deviations from the Mean

	2004: n1=213, n2=204					
	Total1	Total2	Listen1	Listen2	Read1	Read2
Maximum	890	695	465	375	425	335
Minimum	235	265	150	140	60	85
Mean	462	472	249	247	212	225
Median	450	470	250	245	210	225
Stdv	86.1	85.1	47.9	45.5	52.8	53.5
No.>3 Stdv	1	0	1	0	1	0
No.<3 Stdv	0	0	0	0	0	0

	2005: n1=207, n2=200					
	Total1	Total2	Listen1	Listen2	Read1	Read2
Maximum	725	705	395	425	330	310
Minimum	210	230	145	115	55	75
Mean	446	435	252	242	194	194
Median	445	435	255	245	190	193
Stdv	84.7	85.3	46.4	46.8	51.4	51.3
No.>3 Stdv	1	1	1	2	0	0
No.<3 Stdv	0	0	0	0	0	0

	2006: n1=207, n2=205					
	Total1	Total2	Listen1	Listen2	Read1	Read2
Maximum	675	770	380	415	335	360
Minimum	160	230	100	125	60	80
Mean	469	479	264	269	205	210
Median	480	480	265	270	210	210
Stdv	83.6	94.6	48.4	53.1	49.4	55.2
No.>3 Stdv	0	1	0	0	0	0
No.<3 Stdv	1	0	1	0	0	0

[72] Visual displays in the form of bar graphs of much of this data sorted in many ways can be seen in parts II and III and appendices A to M of Bresnihan.

Table 6 cont.
Maximums, Minimums, Means, Medians, Standard Deviations, and Number of Scores Greater Than 3 Standard Deviations from the Mean

	2007: n1=224, n2=218					
	Total1	Total2	Listen1	Listen2	Read1	Read2
Maximum	820	835	435	465	385	370
Minimum	230	280	125	115	90	95
Mean	461	476	255	257	206	219
Median	455	473	255	255	200	215
Stdv	82.0	88.1	46.8	49.5	48.9	51.2
No.>3 Stdv	2	2	2	2	1	0
No.<3 Stdv	0	0	0	0	0	0

	2008: n1=210, n2=206					
	Total1	Total2	Listen1	Listen2	Read1	Read2
Maximum	710	735	390	390	385	385
Minimum	275	275	110	120	115	85
Mean	466	484	244	269	222	216
Median	470	483	240	270	220	215
Stdv	73.1	86.2	41.7	44.6	47.3	55.1
No.>3 Stdv	1	0	0	0	2	1
No.<3 Stdv	0	0	1	0	0	0

	2009: n1=199, n2=194					
	Total1	Total2	Listen1	Listen2	Read1	Read2
Maximum	685	730	385	410	360	410
Minimum	280	255	125	145	95	100
Mean	479	491	243	263	236	228
Median	480	490	240	260	235	230
Stdv	79.5	88.3	41.4	45.8	51.3	55.3
No.>3 Stdv	0	0	2	1	0	1
No.<3 Stdv	0	0	1	0	0	0

Of the maximum Total scores, three are above 800, on the first test in 2004 and the two tests in 2007. Six other maximum Total scores are between 700 and 800, and three are in the upper 600s, on the second test in 2004 and first tests in 2006 and 2009. One minimum Total score is below 200, on the first test in 2006. All of the other minimum Total scores are between 200 and 300. All of the means for Total scores are between 435 and 491, inclusive. The Total scores' mean reported in the initial study by ETS is 578. The range of those Total scores is not given.

Half of the maximum Listening scores are above 400, and half are in the upper 300s. All of

the minimum Listening scores are between 100 and 150, inclusive. The Listening scores' means are between 242 and 269, inclusive. In the initial study, the Listening scores vary from 40 to 495, and the Listening scores' mean is reported to be 290.

Ten of the maximum Reading scores are between 300 and 400, and two, on the first test in 2004 and the second test in 2009, are in the lower 400s. Ten of the minimum Reading scores are below 100, and two are in the lower 100s, on the first test in 2008 and the second test in 2009. The means for the Reading scores are between 194 and 236. The variation in the Reading scores reported in the initial study is from 5 to 455, and the mean is 288.

In the initial study, Woodford states, "It is quite gratifying that the (grading) scale functions as intended. Almost all of the points on the scale are utilized for both sections of the test as well as for the total score."[73] The entire range is from 5 to 495 for each of the two sections, Listening and Reading, and 10 to 990 for the Total score.[74] In Table 6, it is clear that the scores being used in this study do not vary as much as those in the initial study. Scores are not found near either end of the three scales for most of this data. The only ones that are close are the highest Listening scores and the lowest Reading scores. Therefore, these scores are more restricted in their variation than are the scores in the initial study. They are from a population more homogeneous in English language abilities than the population tested in the initial study. At the end of Section 2 above, referring to information provided by ETS, it is explained that the greater the homogeneity of the group in the abilities being tested, the less reliable will be the test scores of those abilities. Also, the averages of the Listening scores are somewhat lower and the averages of the Reading scores are much lower than those in the initial study. So, the language abilities of these students as measured by the TOEIC are quite a bit lower than those of a great of the subjects as measured by the TOEIC for the initial study.

Table 6 also shows that the means and medians, or numerical mid points, of the scores are very similar in all cases for the Total, Listening, and Reading scores. Almost all are within a few points of each other. There are only two differences greater than 10 points. They are of 12 and 11 points for Total scores on the first test in 2004 and the first test in 2006, respectively. As the scales for the TOEIC scores are so wide, these small differences between the means and medians indicate that these sets of scores are distributed very close to normally, close enough to be assumed they are normally distributed for further types of statistical analysis. This means that about 68% of

[73] This quotation is from page 9 of Woodford.
[74] These figures are from page 5 of Woodford.

the scores are expected to be within plus or minus one standard deviation of the mean, about 95% of the scores are expected to be within plus or minus two standard deviations of the mean, and 99.7% of the scores are expected to be within plus or minus three standard deviations of the mean.

The last two rows of each chart in Table 6 give the number of scores that were more than three standard deviations greater than and less than the means. In one case there were three scores that fell more than three standard deviations from the mean, for the Listening scores on the first test in 2009. All of the other administrations have between zero and two scores that are greater than and/or less than three standard deviations from the mean combined. The "68-95-99.7 rule" for data that is normally distributed specifies that it is not unusual for three scores in every one thousand to be more than three standard deviations from the mean. If many more than this were to occur, then it would suggest that the data was unusual in some way. Our data sets do not violate this rule. Therefore, it is further support for these data sets being normally distributed, and no scores need to be deleted from them before continuing our analysis.

When a standardized test is administered to large numbers of people, a normal distribution is considered to indicate that the test performed well in separating the test takers based on their abilities. In a classroom testing situation, however, a normal distribution of scores on a test would nearly always be considered not good, despite the desires for such a distribution by some administrators and teachers. Teachers want all of their students to have learned close to all of the material covered. Therefore, teachers should want most of the test scores to be at least something like 80% or better and none to be lower than about 70%. So, a normal distribution on such a test might indicate that something was wrong with the test. The most likely reason would be that many of the questions on the test were too difficult for many of the students, who were uncertain of the correct answers and so responded by guessing. In such cases, the normal distribution would indicate a regression to the mean, something one does not want in test results as it indicates that the test takers' abilities were not measured accurately.

6. Comparisons of First and Second Test Administration Scores' Averages

Using only the scores by students who sat for both test administrations, analyses of variance were carried out to discover if there were any significant differences between any two means for Total scores, Listening scores, and Reading scores. Table 7 displays the results of the one-way analysis of variance for Total scores' means of all twelve test administrations. It shows that significant differences do exist in the data at the alpha = 0.05 level ($F(11, 2402) = 7.1269$, $p <$

0.0001). Therefore, Tukey-Kramer's Honestly Significant Difference procedure was carried out with alpha set at 0.05 to discover which pairs of means are significantly different. Of the 66 pairings, 17 differences are significant, with values between 27.633 and 0.165, inclusive. However, of the six pairings that would be relevant to consider here, those between the means on the first test and the second test of the same year, none are significantly different. Therefore, according to these analyses, there are no significant differences between these average Total scores on the first test and the second test in each of these six years.

Table 7
One-way Analysis of Variance for Total Scores' Means
of All Test 1 and Test 2 Administrations

Source	Degrees of Freedom	Sum of Squares	Mean Square	F Ratio	Probability > F Ratio
Between Groups	11	558836	50803.2	7.1269	<0.0001
Within Groups	2402	17122369	7128.4		
Corrected Total	2413	17681204			

Alpha equals 0.05.

Table 8 shows the one-way analysis of variance for Listening scores' means' results for all of the test administrations. Significant differences in the data are shown to exist at the alpha = 0.05 level ($F(11, 2402) = 9.1120$, $p < 0.0001$). Tukey-Kramer's Honestly Significant Difference procedure revealed that of the 66 pairings, 20 means are significantly different, with values between 12.936 and 0.217, inclusive. Two of these are relevant to this study, between the means of the first and second tests in 2008 and 2009, with values of 9.065 and 4.294, respectively. This indicates that there are significant differences between the Listening scores' means on the first test and the second test in 2008 and in 2009, but not in 2004, 2005, 2006, and 2007.

Table 8
One-way Analysis of Variance for Listening Scores' Means
of All Test 1 and Test 2 Administrations

Source	Degrees of Freedom	Sum of Squares	Mean Square	F Ratio	Probability > F Ratio
Between Groups	11	217176.3	19743.3	9.112	<0.0001
Within Groups	2402	5204521.5	2166.6		
Corrected Total	2413	5421697.7			

Alpha equals 0.05.

Table 9 displays the results of the one-way analysis of variance for Reading scores' means for all test administrations. Again, significant differences in the data are indicated at the alpha = 0.05 level ($F(11, 2402) = 11.9221$, $p < 0.0001$). According to Tukey-Kramer's Honestly Significant Difference procedure, however, of the 25 significant differences in the 66 pairings, with values between 23.419 and 0.099, inclusive, none are between the six pairs of interest in this study, between pairs for the same year. Therefore, there are no significant differences found between any of the first test and second test Reading scores' means for the same year using these procedures.

Table 9
One-way Analysis of Variance for Reading Scores' Means
of All Test 1 and Test 2 Administrations

Source	Degrees of Freedom	Sum of Squares	Mean Square	F Ratio	Probability > F Ratio
Between Groups	11	349872.6	31806.6	11.9221	<0.0001
Within Groups	2402	6408224.4	2667.9		
Corrected Total	2413	6758097			

Alpha equals 0.05.

Based on these one-way analyses of variance, there are very few instances of meaningful differences between the means of the scores on the first and second test administrations. Only for the 2008 and 2009 Listening scores' data are there significant differences. All of the other pairings indicate no significant differences in the means.

More detailed information concerning comparisons of the Total scores' means for the same year are presented in Table 10. Recalling that the possible range of the Total scores is from 10 to 990, the differences in the Total scores' means on the two test administrations for each year, shown in the third row, varying from -9 to 17 points when subtracting the first mean from the second, with standard errors of between 5.1 and 6.3, shown in the sixth and seventh rows, seem quite small. The absolute absolute values of the figures in the eighth row, of between 0.105 and 0.213, indicate the effect sizes of the pairs of means. (Whether positive or negative is of no importance when considering these figures.) Though subject to different interpretation depending on the exact details of the data being analyzed, the general understanding of Cohen's d figures is that a value of 0.5 would be considered medium in size while a value of 0.2 would be considered small.[75] The covariability coefficients, which are in the bottom row, indicate that small would be an optimistic

[75] This interpretation of and more information about Cohen's d figures can be found in Becker.

description. The covariabilities vary from 0.003 to 0.011, meaning that the largest amount of the average Total score for one year that can be explained by the other average Total score of the same year, is just 1.1%, as far as this analysis can determine.

Table 10
Total Scores' Mean, Change in Mean, Standard Deviation, Standard Error, Cohen's d, Correlation Coefficient, and Covariability Coefficient for Each Pairing of Scores: 2004 to 2009

	2004 n=198	2005 n=197	2006 n=202	2007 n=218	2008 n=204	2009 N=188
Mean1	458	444	468	462	467	477
Mean2	472	435	479	476	484	491
Change	14	-9	11	14	17	14
Stdv1	80.8	85.3	83.7	81.8	72.9	80.2
Stdv2	85.7	85.9	94.0	88.3	86.3	86.3
Std Err1	5.7	6.1	5.9	5.5	5.1	5.9
Std Err2	6.1	6.1	6.6	6.0	6.0	6.3
d	0.168	-0.105	0.124	0.164	0.213	0.168
r	0.084	-0.052	0.062	0.082	0.106	0.084
r2	0.007	0.003	0.004	0.007	0.011	0.007

d equals the subtraction of Mean2 minus Mean1 divided by the square root of the division of the summation of Stdv2 squared plus Stdv1 squared divided by 2, the Cohen's d figure.
r equals d divided by the square root of the summation of d squared plus 4, the correlation coefficient.
r2 equals r times r, the covariability coefficient.

Table 11 displays the same information for the Listening scores' means as is displayed for the Total scores' means in Table 10. Listening scores can vary from 5 to 495, and the differences in the Listening scores' means when subtracting the first mean from the second vary from -10 to 24, as is shown in the third row. The standard errors, displayed in the sixth and seventh rows, are all between 2.9 and 3.7, inclusive. So, in some cases, in 2008 and 2009, there is more variation in the Listening scores' means than the Total scores' means, though the range of possible scores is half as large. These two differences suggest real differences. The figures shown in the eighth row give varied effect sizes, with one value, 0.000, indicating no relationship, three values indicating very small to small relationships, 0.021, 0.138, and -0.213, and others indicating medium relationships, 0.456 and 0.553. The covariability coefficients, however, indicate that the effects are still very slight, indeed. The covariabilities, in the bottom row, vary from 0.000 to 0.071, which means that, according to this analysis, the most of any average Listening score that can be explained by the other average Listening score of the same year is 7.1%.

Table 11
Listening Scores' Mean, Change in Mean, Standard Deviation, Standard Error, Cohen's d, Correlation Coefficient, and Covariability Coefficient for Each Pairing of Scores: 2004 to 2009

	2004 n=198	2005 n=197	2006 n=202	2007 n=218	2008 n=204	2009 n=188
Mean1	248	251	263	256	245	243
Mean2	248	241	270	257	269	263
Change	0	-10	7	1	24	20
Stdv1	45.3	46.8	48.0	46.8	42.0	42.3
Stdv2	45.7	47.2	53.1	49.6	44.8	45.3
Std Err1	3.2	3.3	3.4	3.2	2.9	3.1
Std Err2	3.2	3.4	3.7	3.4	3.1	3.3
d	0.000	-0.213	0.138	0.021	0.553	0.456
r	0.000	-0.106	0.069	0.010	0.266	0.222
r2	0.000	0.011	0.005	0.000	0.071	0.049

d equals the subtraction of Mean2 minus Mean1 divided by the square root of the division of the summation of Stdv2 squared plus Stdv1 squared divided by 2, the Cohen's d figure.
r equals d divided by the square root of the summation of d squared plus 4, the correlation coefficient.
r2 equals r times r, the covariability coefficient.

The same information as is presented for Total scores' means and Listening scores' means in Table 10 and Table 11, respectively, is shown in Table 12 for Reading scores' means. Reading scores have the same possible variation as Listening scores, from 5 to 495. When the first Reading scores' mean is subtracted from the second, the differences vary from -6 to 14, inclusively, as is shown in the third row. There is less variation here than between either Total scores' means or Listening scores' means. Presented in the sixth and seventh rows, the standard errors vary from 3.3 to 3.9, which is quite similar to the standard errors of the Listening scores' means. The absolute values of the figures displayed in the eighth row are between 0.000 and 0.267, the maximum value being slightly larger than the largest for the Total scores' means but much smaller than the largest for the Listening scores' means and indicating from no relationship to little relationship. These are confirmed by the covariability coefficients, which vary from 0.000 to 0.017, inclusively, suggesting that none of these Reading scores' averages can explain more than 1.7% of its corresponding Reading scores' average.

These more detailed analyses of the differences between the means on the first and second test administrations using basic and effect size statistics indicate that very little change occurred in

most of the scores' means. There is more change in two of the Listening scores' means than in the Total scores' means or the Reading scores' means. These are the changes in Listening scores' means for the 2008 and 2009 data. However, even these changes are small.

Table 12
Reading Scores' Mean, Change in Mean, Standard Deviation, Standard Error, Cohen's d, Correlation Coefficient, and Covariability Coefficient for Each Pairing of Scores: 2004 to 2009

	2004 n=198	2005 n=197	2006 n=202	2007 n=218	2008 n=204	2009 n=188
Mean1	211	194	205	206	222	234
Mean2	225	194	209	219	216	228
Change	14	0	4	13	-6	-6
Stdv1	50.9	51.9	49.9	48.9	46.7	50.7
Stdv2	54.0	51.6	54.7	51.3	54.9	54.0
Std Err1	3.6	3.7	3.5	3.3	3.3	3.7
Std Err2	3.8	3.7	3.8	3.5	3.8	3.9
d	0.267	0.000	0.076	0.259	-0.118	-0.115
r	0.132	0.000	0.038	0.129	-0.059	-0.057
r2	0.017	0.000	0.001	0.017	0.003	0.003

d equals the subtraction of Mean2 minus Mean1 divided by the square root of the division of the summation of Stdv2 squared plus Stdv1 squared divided by 2, the Cohen's d figure.
r equals d divided by the square root of the summation of d squared plus 4, the correlation coefficient.
r2 equals r times r, the covariability coefficient.

7. Comparisons of Each Student's First and Second Test Administration Scores

Table 13 shows the numbers of students whose Total scores increased, decreased, or remained the same the second time they took the TOEIC IP Test as compared to the first for each year. The first row indicates that the amounts of students whose Total scores increased are usually a bit less than 60%, except for 2005 where it is just 43%. So, usually, a little more than 40% of the students' Total scores did not increase on the second test. The average increases in Total scores, given in the last row, are between 11 and 18 points, inclusive, with one average decrease in Total scores of -9 in 2005. The variations in changes in scores, seen in the fourth and fifth rows, are between 225 and -235, both in 2009. The smallest maximum increase is 135 in 2005 and the smallest maximum decrease is -140 in 2007.

Table 13
Numbers of Students Whose Total Scores Are Different on Test 2 Than on Test 1 and Ranges and Means of the Changes

	2004	2005	2006	2007	2008	2009
No. Increased	113 57%	85 43%	118 58%	122 56%	117 57%	110 59%
No. No Change	7 4%	10 5%	5 2%	5 2%	7 3%	5 3%
No. Decreased	78 39%	102 52%	79 39%	91 42%	80 39%	73 39%
Max. Increase	190	135	200	195	195	225
Max. Decrease	-165	-175	-195	-140	-175	-235
Mean Change	14	-9	11	13	18	14

The numbers of students whose Listening scores increased, decreased, or remained the same when they took the TOEIC IP Test the second time when compared with the first are presented in Table 14. The first row discloses what seem to be large differences in the performances of the students in 2008 and 2009 compared to the four previous years. From 2004 to 2007, the amounts of students whose Listening scores increased are 48%, 41%, 57%, and 48%, respectively, while they are 75% in 2008 and 65% in 2009. As would be suspected, the last row reflects similar changes between the 2004 to 2007 average changes in Listening scores, being 0, -9, 7, and 1, respectively, and those in 2008 and 2009, being 24 and 20, respectively. In the fourth and fifth rows, the largest increase in Listening score is 185 in 2009 and the largest decrease is -130 in 2005. The smallest maximum increase is 100 points in 2005 and the smallest maximum decrease is -85 in 2009.

Table 14
Numbers of Students Whose Listening Scores Are Different on Test 2 Than on Test 1 and Ranges and Means of the Changes

	2004	2005	2006	2007	2008	2009
No. Increased	95 48%	81 41%	116 57%	104 48%	154 75%	123 65%
No. No Change	10 5%	8 4%	5 2%	14 6%	9 4%	11 6%
No. Decreased	93 47%	108 55%	81 40%	100 46%	41 20%	54 29%
Max. Increase	110	100	150	140	130	185
Max. Decrease	-125	-130	-125	-105	-115	-85
Mean Change	0	-9	7	1	24	20

In Table 15, the numbers of students whose Reading scores changed, or not, the second time they took the TOEIC IP Test when compared with their first attempt and in what way are displayed. The first row indicates the reverse of what is seen in Table 14 for the Listening scores. From 2004 to 2007, the amounts of students whose Reading scores increased are 60%, 51%, 51%, and 61%,

respectively. In 2008 and 2009, the amounts whose Reading scores increased are 41% and 43%, respectively. The last row shows the average changes in Reading scores. Again, these reflect what is seen in the first row. The average changes from 2004 to 2007 are 13, 0, 4, and 12 points, respectively, and in 2008 and 2009, they each -6 points. The largest increase in Reading score on the second test as compared with the first is 140 points in 2008, and the largest decrease is -150 in 2009. The smallest maximum increase is 95 in 2005, and the smallest maximum decrease is -80 in 2004.

Table 15
Numbers of Students Whose Reading Scores Are Different on Test 2 Than on Test 1 and Ranges and Means of the Changes

	2004		2005		2006		2007		2008		2009	
No. Increased	118	60%	100	51%	103	51%	133	61%	83	41%	81	43%
No. No Change	6	3%	8	4%	12	6%	10	5%	11	5%	6	3%
No. Decreased	74	37%	89	45%	87	43%	75	34%	110	54%	101	54%
Max. Increase	135		95		120		125		140		115	
Max. Decrease	-80		-115		-130		-115		-105		-150	
Mean Change	13		0		4		12		-6		-6	

Tables 13, 14, and 15 show that usually about one third to one half of the students' scores are lower on the second test administration than the first and that usually about two fifths to three fifths of the students' scores are higher on the second test administration than the first. Only the Listening scores' data for 2008 and 2009 are somewhat different, both displaying larger percentages of increases, of 75% and 65%, respectively, in scores from the first test administration to the second, and correspondingly lower percentages of decreases. So generally, more students achieved higher scores on the second test than on the first, but many students achieved higher scores on the first test than on the second, and sometimes more students' scores were higher on the first test than the second (for Total score in 2005, for Listening score in 2005, and for Reading score in 2008 and in 2009). Also, sometimes the numbers of students who scores increased and decreased on the second test were almost the same (for Listening score in 2004 and in 2007).

More important than the numbers of students whose scores increase or decrease the second time they take a test is the numbers of students whose scores increase or decrease enough to indicate a real change in their scores, and presumably in their English language abilities, if these scores are accurate measures of these abilities for these test takers. Inherent to all tests are errors in the

testing method or instrument that need to be considered along with the score in order for the score to be correctly understood in its measurement of the ability being tested. It was mentioned earlier that, for the TOEIC test, the standard error of difference between two Listening scores or between two Reading scores is approximately +/-35 points with about 68% confidence. To be about 95% confident, the error or confidence band is approximately +/-69 points for each section's score.[76] It seems that it would be better to use the latter amount so as to be quite certain that changes in ability have taken place. However, both amounts will be included in the following tables for consideration. Also, although ETS does not provide a standard error of difference for the TOEIC Total score, it will be assumed to be the sum of the standard errors of difference for the Listening scores and the Reading scores, or +/-70 points.

Table 16
Numbers of Students Whose Total Scores Are More Than 70 and 138 Points Different on Test 2 Than on Test 1

	2004 n=198		2005 n=197		2006 n=202		2007 n=218		2008 n=204		2009 n=188	
No.>70	32	16%	25	13%	25	12%	35	16%	39	19%	30	16%
No.>138	6	3%	0	0%	5	2%	7	3%	7	3%	5	3%
No.<70	13	7%	35	18%	13	6%	20	9%	13	6%	17	9%
No.<138	1	1%	2	1%	4	2%	1	0%	3	1%	2	1%

Using the assumed confidence bands mentioned immediately above, which are based on information supplied by ETS, Table 16 presents the amounts of students whose second Total scores on the TOEIC IP Test are more than 70 points and more than 138 points different than their first Total scores. The first row indicates that between 12% and 19%, inclusive, of the students' Total scores are more than 70 points higher on the second administration, suggesting with 68% confidence that a real improvement in their Total score and English language ability took place. With 95% confidence, or considering an increase of more than 138 points, the amounts of students whose scores truly increase are from 0% to 3%, inclusive, as seen in the second row. The two bottom rows show the amounts of students whose Total scores decease by these same amounts. Between 6% and 18%, inclusive, decrease by more than 70 points and between 0% and 1%, inclusive, decrease by more than 138 points, suggesting with 68% and 95% confidence, respectively, a real decrease in Total score and English language ability. About 69% to 82%,

[76] These figures are from pages IV-6 to IV-7 of *TOEIC Technical Manual*.

inclusive, of the students' Total scores do not change by more than 70 points, and about 96% to 99%, inclusive, of the students' Total scores do not change by more than 138 points, indicating no changes in English language ability at those confidence levels.

Table 17 displays the amounts of students whose Listening scores change by more than 35 points and by more than 69 points. The first row shows that between 11% and 35%, inclusive, of the students' Listening scores are more than 35 points higher on the second test, indicating with 68% confident that their scores increase and their English listening abilities improved. The second row shows that between 4% and 12%, inclusive, of the students' Listening scores increase by more than 69 points on the second test, suggesting that their scores and their English listening abilities improved, with 95% confidence. The two bottom rows present decreases of more than 35 points for between 6% and 26%, inclusive, of the students' Listening scores and of more than 69 points for between 1% and 8%, inclusive, of the students' Listening scores, suggesting with 68% and 95% confidence, respectively, a real decrease in score and English listening ability. The amounts of students whose Listening scores indicate no changes in English listening ability are between 57% and 73%, inclusive, with 68% confidence and between 86% and 93%, inclusive, with 95% confidence.

Table 17
Numbers of Students Whose Listening Scores Are More Than 35 and 69 Points Different on Test 2 Than on Test 1

	2004 n=198		2005 n=197		2006 n=202		2007 n=218		2008 n=204		2009 n=188	
No.>35	38	19%	21	11%	41	20%	40	18%	71	35%	62	33%
No.>69	10	5%	7	4%	11	5%	12	6%	18	9%	22	12%
No.<35	30	15%	51	26%	24	12%	39	18%	12	6%	18	10%
No.<69	11	6%	16	8%	4	2%	9	4%	2	1%	4	2%

The amounts of students whose Reading scores increase and decrease by more than 35 points and by more than 69 points on the second test administration are presented in Table 18. The first row shows that between 14% and 26%, inclusive, of students' Reading scores are more than 35 points higher on the second test than the first, indicating a real increase in score and in English reading ability with 68% confidence. The second row suggests, with 95% confidence, a real increase in Reading score and in English reading ability for between 4% and 10%, inclusive, of the students. The two bottom rows indicate that, with 68% confidence, between 9% and 29%,

inclusive, of the students' Reading scores are truly lower on the second test administration than the first because of a decease in score of more than 35 points, and their English reading abilities were really weaker, and, with 95% confidence, that between 2% and 9%, inclusive, of the students' Reading scores are truly lower on the second test administration than the first because of a decrease in score of more than 69 points, and their English reading abilities were really weaker. The amounts of students whose Reading scores suggest no changes in their real scores or in their reading abilities are between 29% and 46%, inclusive, with 68% confidence, and between 86% and 91%, inclusive, with 95% confidence.

Table 18
Numbers of Students Whose Reading Scores Are More Than 35 and 69 Points Different on Test 2 Than on Test 1

	2004 n=198		2005 n=197		2006 n=202		2007 n=218		2008 n=204		2009 n=188	
No.>35	51	26%	39	20%	33	16%	57	26%	35	17%	26	14%
No.>69	18	9%	7	4%	12	6%	21	10%	10	5%	11	6%
No.<35	18	9%	35	18%	27	13%	21	10%	59	29%	44	23%
No.<69	4	2%	11	6%	7	3%	4	2%	18	9%	10	5%

Tables 16, 17, and 18 show that the numbers of students whose scores indicate real changes in scores on the second test when compared with the first test is quite low. With 95% confidence, one would expect to feel pretty sure that an increase in scores is real. This is the case for between 0% and 3% of the students' Total scores, between 4% and 12% of the students' Listening scores, and between 4% and 10% of the students' Reading scores. This tells us that most of the students whose scores increased did not achieve both a higher Listening score and a higher Reading score but a higher score for one or the other. Also, with 95% confidence that the changes in scores on the second test when compared with the first are accurate, between 0% and 2% of the Total scores, between 1% and 8% of the Listening scores, and between 2% and 9% of the Reading scores decreased. It would be very unusual and disturbing if real decreases in some of the students' abilities being measured took place while the students were studying and practicing them. However, this is what these analyses suggest.

8. Discussion and Concluding Remarks

It seems to be implicit, given the extremely widespread and pervasive use of the TOEIC and

TOEIC scores in schools throughout Japan, that many people connected with English teaching in Japan believe that TOEIC scores are able to precisely measure students' English language abilities and to accurately measure students' learning in their English courses. If you were in charge of the program from which the scores being used in this study were collected, how would you evaluate its results, based on its students' performances on the TOEIC, which have been presented and analyzed above? If you believed that TOEIC scores were a good tool for measuring students' learning in their English courses, probably you would not evaluate the results of the teaching and learning in this program very highly, and perhaps you would wonder if it were sensible to continue offering these courses. However, numbers of experts in the field of language testing disagree with the basic assumption that TOEIC scores are accurate instruments for monitoring and evaluating students' progress in English language courses in schools.

Language testing expert Brown, for instance, clearly disagrees. He wrote an article addressing the misuse of standardized test scores particularly addressing people who need to evaluate students in English language courses and make decisions about English language programs in schools in Japan. At that time, the TOEIC was not yet a very widely known or used test, so he specifically mentions TOEFL scores, but not TOEIC scores. However, TOEIC scores and TOEFL scores are of the same type. In the article, he explains the differences between criterion-referenced tests and norm-referenced tests. The TOEFL and the TOEIC are norm-referenced tests. He states that norm-referenced tests "are not typically designed to test material that is specifically and directly related to a single course or program."[77] "(T)he characteristics of norm-referenced tests make them inappropriate for assessing what percent of material covered in class each student has learned for diagnostic, progress, or achievement decisions."[78] "(T)he content of the TOEFL (and the TOEIC, as well) is entirely too broadly defined to be useful in tracking the progress of students, or measuring their achievement in semester-length, or even year-long English courses."[79] This seems obvious as the people who make standardized tests know nothing about the great many different courses the students who take the tests are studying in. It seems that many schools and English language programs ignore this fact.

On the other hand, Brown explains that "(c)riterion-referenced tests are usually based on the very specific objectives of a course or program."[80] They are created by teachers and/or individual

[77] This quotation is from page 13 of Brown.
[78] This quotation is on page 18 of Brown.
[79] This quotation is on page 18 of Brown. The parenthetic phase is added.
[80] This quotation is from page 14 of Brown.

programs in order to measure the students' increased understanding, retention, and level of performance related to the specific materials, content, language, tasks, and abilities covered in a course of study. So, why are so many English courses and programs in Japanese schools using TOEIC scores to determine grades and to award credits and certification? Are students being evaluated in any other courses based on standardized tests rather than on tests created for use in those specific courses and/or for the materials used in those courses? Likely, it is because nowadays many people have come to believe that all things can be considered as uniform, as are the natural laws of the physical world (although even the uniformity of those is now being demonstrated to be less constant than has been imagined) and that everything should be and can be measured in quantitative ways similar to the mechanical ways many things of concern in the physical sciences can be and are measured. TOEIC scores appear to measure students' English language abilities in just such a mechanical and precise way because specific scores are reported. However, this precision is a fallacy. TOEIC scores presented by Bresnihan, Childs, and Saegusa, for example, clearly demonstrate that this is the case. [81]

Also, TOEIC scores are seductively easy to use and understand, even if the truthfulness to support their appearance is lacking, as it may be when considering an individual person's scores. They are easy to use because, once a scale has been created to translate specific TOEIC scores into school grades, teachers do not need to make end-of-term grades for students; they only need to record them. In fact, teachers might not even feel the need to grade tests or papers or to do any kind of evaluating at all if TOEIC scores are the sole means for evaluating their students. They are easy to understand for people with little or no knowledge about English language learning because they have become well known and accepted due to their widespread usage during recent years and to the many detailed descriptions published by its makers and promoters, even if many people's understanding of them is incorrect because certain aspects about them, like the wide ranges of the standard errors of measurement and difference and the possible fluctuation in scores, are often unknown and almost never considered. They have subconsciously entered people's thinking, their way of picturing, considering, and understanding the world around them and their experiences. However, this does not logically imply that their usage is necessarily correct or beneficial, just that it has become fashionable.

That TOEIC scores have now become established, have become part of the accepted norm, also gives them prestige and authority, even in ways that they cannot truthfully support, like their

[81] See parts VI and VII of Bresnihan, pages 69 and 70 of Childs, and pages 177 to 180 of Saegusa.

ability to measure individual's English language abilities precisely or to measure people's English language learning gains in courses or programs that do not include sufficient amounts of study time. However, people in positions of authority, such as program directors and administrators, can use them to put pressure on and to assure and maintain their status over those below them, who are likely to be teachers or those needing to study English, and to force standardization and uniformity, which are wrongly believed to achieve better results than creativity and variety.

Yet from another point of view, Brown states, "Administrators and teachers alike should also realize that using (norm-referenced tests) for (criterion-referenced test) purposes minimizes the possibilities that their program will look good."[82] This is because they "simply are not sensitive enough"[83] to measure gains in individual courses and from short amounts of study time. Concerning the latter, professional test writer Woodhead, in a recent interview by Wood, states that "(n)either TOEIC nor Bridge are designed for re-testing with less than 90-120 hours of instruction time in between each attempt."[84] Childs, as mentioned earlier, also concurs with this. "TOEIC seems a poor instrument for gauging the short-term learning gains of individuals . . . in programs with relatively few teaching hours."[85] Referring to his study, he states, "We have seen that, in a teaching program that totaled 53 hours, the variability of TOEIC results defeated their usefulness in measuring learning gains because the (standard error of measurement) of TOEIC was in the range of expected individual gains."[86] Therefore, real gains in English language abilities could not be demonstrated through the TOEIC scores obtained.

Perhaps the only large-scale study that attempts to determine the amounts of classroom English language study time needed to make certain gains in TOEIC scores, as a measure of English language proficiency, is an article by Saegusa published some 25 years ago. Using "thousands of TOEIC scores,"[87] mostly of young adult workers in company-arranged English classes of usually 3 to 6 months duration with about 10 workers per class taught by native speakers of English,[88] supplied by those in charge of the TOEIC in Japan, Saegusa employed multiple correlations, regression equations, and standard errors of measurement to determine how much time studying in English language classes is needed in order to ensure that a high percentage of the students improve

82 This quotation is on page 18 of Brown.
83 This quotation is from page 18 of Brown.
84 This quotation is from page 42 of Wood.
85 This quotation is on page 73 of Childs.
86 This quotation is on page 74 of Childs.
87 This quotation is from page 165 of Saegusa.
88 This information is on page 167 of Saegusa.

their TOEIC scores by certain amounts. In summarizing his findings, he states,

> "(L)ess 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."[89]

> "It usually takes more time to improve English proficiency than is generally believed. Our studies show that it will take an average of 400 hours of instruction to raise the proficiency of TOEIC 450 . . . to that of TOEIC 600 . . . . The general definition of (TOEIC 450) is the elementary proficiency of (survival English); and that of (TOEIC 600) is the minimum working proficiency. This distinction is very important, because (TOEIC 600) can be a criterion upon which to distinguish between working and non-working proficiency. To successfully carry out business in English, however, a higher level . . . roughly equivalent to TOEIC 730 . . . will be required, and to reach that level it is estimated that another 400 hours of instruction will be needed."[90]

Based on the explanations provided by ETS on how to use standard errors of measurement and standard errors of difference,[91] the present author pointed out in an earlier study that Saegusa should have used standard errors of difference instead of standard errors of measurement to make his calculations. The latter are 29% smaller than the former. Therefore, the amounts of classroom study time needed to expect a good percentage of the students to increase their TOEIC scores by the amounts indicated by Saegusa are actually shorter than they should be.[92]

Considering college and university English language courses in Japan, usually each course meets for 90 minutes once a week for 15 weeks per semester, and 20 students or less per class is

---

89 This quotation is on page 174 of Saegusa.

90 This quotation is on page 181 of Saegusa. The TOEIC scores in parentheses are substituted for another test's scores as per comparisons made by the author in the article.

91 These explanations are on pages IV-4 to IV-7 of *TOEIC Technical Manual*.

92 An explanation of this mistake is given on pages 213 to 214 of Bresnihan.

quite rare. Twice that number of students per class is not unusual. Many schools require two different English language courses to be taken during each semester of the first year. There is a more than two months' long summer break in between semesters. As pointed out earlier, the students whose scores are being used in this study were registered for three such courses per semester, and so had the potential to attend English language classes for 67.5 hours per semester. Also, the potential amount of English language study time between the two TOEIC IP Test administrations was about 18 hours, then about 11 weeks of no classes, and then another 49.5 potential hours of study. This number of hours does not even come close to the specific minimum number of hours recommended by Saegusa or Woodhead, cited just above. Therefore, to expect many of these students to noticeably improve their overall English language abilities or proficiency in ways that would be identifiable by significantly higher TOEIC scores on the second test than the first in such short periods of time, due to their limited studies in their university English courses, would be unreasonable. It would also be unreasonable to expect many students' overall English language abilities or proficiency or their TOEIC scores to increase much or at all during their first year at college or university almost anywhere in Japan, unless the students study and practice using English a great deal more than they need to just to pass their English courses. However, this does not mean that English language courses ought to be eliminated. Even though the amounts of time allotted for students to study the English language in courses is inadequate, these courses probably help many students to retain more of their English language abilities for longer than they would without these courses.

Students will be graded more fairly and accurately in courses based on things like their effort to participate in the work of the course, their improvement in abilities practiced and/or introduced in the course, and their increase in knowledge of the material covered in the course, as measured by their ability to reproduce those tasks and recall that information when required, than by TOEIC scores. TOEIC scores may be able to add some information about students' English language abilities, but they certainly cannot adequately replace what the students' teacher would know about them concerning the course they are studying in. Even a representative for the organization that promotes the TOEIC in Japan says, "We do not recommend corporations use TOEIC in isolation." [93] To so obviously grade students in English language courses essentially based on their English proficiency when they enter these courses, which is the case if a standardized test is used because it cannot measure specific gains from short amounts of study, lacks face validity. What would be the

[93] This quotation is on page 10 of Chapman (2004).

reason for students to take such courses if they are being graded in this way? To evaluate students in courses using a test that students are unlikely to improve their scores on during the time period they attend the courses and that is too difficult for the students to do well on, which seems to be the case with the TOEIC for most students in Japan based on the analyses presented in this study and which might possibly be partially due to the business-related contexts and contents of the test items, discourages students from studying in the courses for which their scores are used and causes them to have less interest in improving their English. Especially, many of the students whose scores will go down the second time they take the test, as will certainly be the case due to the variability of scores inherent in standardized testing, will likely feel discouraged and lose motivation for their English language studies. Emphasizing getting high TOEIC scores rather than improving English abilities also distracts students from setting appropriate goals and from engaging enthusiastically in appropriate and useful activities for their studies in their English courses and of English in general. As Woodhead said, "First the student needs to be motivated to learn English and NOT simply to pass the test." [94]

These are some of the reasons why the reliabilities found in this study are so weak. However, instead of getting rid of English language courses at colleges and universities because of poor results on tests like the TOEIC, a more logical response would be to get rid of the standardized tests, or to use them sparingly. In either case, English language teachers can then use tests of their own individual choice and making and can evaluate their students based on their own grading policies, which is the only way that truly makes sense.

If nothing is known about a person's English language abilities, for example, when a person is applying for a job and/or the person who needs to judge a person's language ability does not know English well, then something like a TOEIC score might be the only means of assessment available, and so its usage in isolation would be justifiable to a certain extent. Yet, those making decisions using those TOEIC scores, and those being judged by them, would be benefited by knowing that, for example, someone who has achieved a TOEIC score of 700 is very likely to be better at using English than someone who has achieved a TOEIC score of only 400. However, when the scores are closer to each other, say 550 and 500 or even 550 and 450, it much less certain that the person with the higher score is truly better at using English than the person with the lower score because, according to the statistics published by ETS for the TOEIC in general, these scores do not indicate actual differences in English proficiency. It would be very important to keep this in

[94] This quotation is on page 44 of Wood.

mind.

TOEIC scores can add an additional perspective on a student's English language abilities, but they cannot adequately replace what an English language teacher would know about the student's English language abilities and her/his performance and effort in the class. Also, the teacher can take into account whether or not the contexts, settings, and assumed experiences of the contents of the TOEIC match with the knowledge and experiences of her/his students, who in Japan have very likely not yet entered the business world. Therefore, they should not be the sole means, nor a significant means, of evaluating students in English language courses. Standardized testing of this sort does not have a long history. Its modern beginnings outside of psychology, along with the use of statistics in these matters, can be found in the 1900s in the USA. In recent years, quantitative testing has come to overshadow almost all other aspects of teaching, learning, and education in many places and in many fields. This attempt to measure quantitatively what is much more qualitative than quantitative is having very negative effects on people's learning and development in certain areas, such as languages. The sooner that administrators and teachers come to realize this and eliminate many of the inappropriate uses of standardized testing, the better it will be for students and for the prospects of beneficial personal and societal development.

References

About the TOEFL Test. 2010. ETS.TOEFL. Dec. 10, 2010. <http://www.ets.org/toefl/institutions/about/>.

About the TOEIC Bridge. 2010. ETS.TOEIC. Sept. 29, 2010.
<http://www.toeic.or.jp/toeic_en/bridge/about.html#a>.

About the TOEIC Speaking and Writing Tests. 2010. ETS.TOEIC. Dec. 10, 2010.
<http://www.ets.org/toeic/speaking_writing/about/>.

Becker, L. 2000. “Effect Sizes (ES).” Effect Sizes (ES). Nov. 9, 2010.
<http://www2.jura.uni-hamburg.de/instkrim/kriminologie/Mitarbeiter/Enzmann/Lehre/StatIIKrim/EffectSizeBecker.pdf>.

Bradford-Watts, K., Ikeguchi, C., & Swanson, M. (Eds.) 2005. *JALT 2004 Conference Proceedings*. Tokyo: JALT.

Bresnihan, B. 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 Brown, J.D. & Yamashita, S.O.(Eds.) *Language Testing in Japan*. pp. 12-19. Tokyo: JALT.

The Cambridge, IELTS, TOEFL and TOEIC compared for equivalencies. Courses-of-English-in-England.com. Dec. 20, 2008.
<http://www.courses-of-english-in-england.com/TOEFL,%20IELTS,%20TOEFL%20First,%20CAE.htm>.

Can-Do Levels Table. ETS.EUROPE. Dec. 20, 2008.
<http://www.uk.toeic.eu/fileadmin/free_resources/UK%20website/UK_Current/TOEIC_Can-do_table.pdf>.

Can-Do Levels Table (2). ETS. Dec. 15, 2009.
<http://www.uk.toeic.eu/fileadmin/free_resources/ETS_Global_master/TOEIC_L_R_can-do_table.pdf>.

Chapman, M. 2003. “TOEIC: Tried But Undertested.” *Shiken: JALT Testing & Evaluation SIG Newsletter*, Vol. 7, No. 3, Autumn. pp. 2-7. Tokyo: JALT. Feb. 12, 2010, <http://jalt.org/test/cha_1.htm>.

Chapman, M. 2004. “Insights in Language Testing: An Interview with Kazuhiko Saito.” *JALT Testing & Evaluation SIG Newsletter*, Vol. 8, No. 2, Aug. pp. 8-11. Tokyo: JALT. Apr. 3, 2010.
<http://www.jalt.org/test/sai_cha.htm>.

Chapman, M. 2005. “TOEIC & TOEFL: A Partnership of Equals?” In Bradford-Watts, K., Ikeguchi, C., & Swanson, M. (Eds.) *JALT 2004 Conference Proceedings*. pp. Tokyo: JALT. Dec. 20, 2008.
<http://jalt-publications.org/archive/proceedings/2004/E19.pdf>.

Chapman, M. 2006. “An Over-reliance on Discrete Item Testing in the Japanese Business Context.” *Achievement of 2006 International Conference on English Instruction and Assessment*. Apr. 22-23. Taiwan: National Chung Cheng University. Mar. 2, 2010.
<http://fllcccu.ccu.edu.tw/conference/2005conference_2/download/C07.pdf>.

Chapman, M, & Newfields, T. 2008. “The ‘New’ TOEIC.” *Shiken: JALT Testing & Evaluation SIG Newsletter*, Vol. 12, No. 2, Apr. pp. 32-37. Tokyo: JALT. Feb. 12, 2010.
<http://jalt.org/test/cha_new.htm>.

Childs, M. 1995. “Good and Bad Uses of TOEIC by Japanese Companies.” In Brown , J.D. & Yamashita, S.O. (Eds.) *Language Testing in Japan*. pp. 66-75. Tokyo: JALT.

Differences between SP group application and IP. ETS.TOEIC. Dec. 8, 2010. <http://www.toeic.or.jp/toeic_en/corpo/guide02.html>.

Educational Testing Service. FundingUniverse. Dec. 20, 2008. <http://www.fundinguniverse.com/company-histories/Educational-Testing-Service-Company-History.html>.

Effect Size Calculators. 1999. Becker, L. Nov. 9, 2010. <http://www.uccs.edu/~faculty/lbecker/>.

Effect Size Calculators (2). 2009. Ellis, P. Nov. 9, 2010. <http://myweb.polyu.edu.hk/mm/effectsizefaqs/calculator/calculator.html>.

ETS Europe UK Launches new TOEIC Speaking and Writing tests May 22, 2008. ETS.org. Dec. 20, 2008. <http://www.uk.toeic.eu/no_cache/toeic-sites/toeic-default/news-toeic/?news=412&view=detail>.

ETS Premieres World's First Internet-Based English-Proficiency Test. Prometric. Dec. 7, 2010. <http://www.prometric.com/News/Press/ETS+Premieres+Worlds+First+Internet-Based+English-Proficiency+Test.htm>.

Frequently Asked Questions About the TOEIC Listening and Reading Test. 2010. ETS.TOEIC. Sept. 29, 2010. <http://www.ets.org/toeic/listening_reading/about/faq>.

Group Application. ETS.TOEIC. Dec. 8, 2010. <http://www.toeic.or.jp/toeic_en/corpo/guide01.html>.

Hirai, M. 2002. "Correlations between Active Skill and Passive Skill Test Scores." *Shiken: JALT Testing & Evaluation SIG Newsletter*, Vol. 6, No. 3, Sept. 2002. pp. 2-8. Tokyo: JALT. Feb. 12, 2010. <http://jalt.org/test/hir_1.htm>.

Hirai, M. 2008. "Correlations between STEP BULATS Writing and TOEIC Scores." In Newfields, T., Wanner, P., & Kawate-Mierzejewska, M. (Eds.) *Divergence and Convergence. Educating with Integrity: Proceedings of the 7th Annual JALT Pan-SIG Conference.* May 10-11. pp. 36-46. Kyoto: Doshisha University Shinmachi Campus. Mar. 9, 2010. <http://jalt.org/pansig/2008/HTML/Hirai1.htm>.

Hirai, M. 2009. "Correlations between STEP BULATS Speaking and TOEIC Scores." In Skier, E. & Newfields, T. (Eds.) *Infinite Possibilities. Expanding Limited Opportunities in Language Education. Proceedings of the 8th Annual JALT Pan-SIG Conference.* May 23-24. pp. 12-25. Chiba: Toyo Gakuen University, Nagareyama Campus. Feb. 15, 2010. <http://jalt.org/pansig/2009/HTML/Hirai.htm>.

Kang, S.W. 2009. "Home-Grown English Test to Replace TOEFL, TOEIC." May 20. *The Korea Times*. Apr. 3, 2010. <http://www.koreatimes.co.kr/www/news/special/2010/01/181_45304.html>.

Lee, B.M. 2007. "TOEFL and TOEIC Need Replacements." Apr. 20. *Munhwa Ilbo*. Dec. 7, 2010. <http://www.koreafocus.or.kr/design1/layout/content_print.asp?group_id=101583>.

McCrostie, J. 2009. "TOEIC No Turkey at 30." Aug. 11. *The Japan Times: Online*. Feb. 15, 2010. <http://search.japantimes.co.jp/cgi-bin/fl20090811zg.html>.

McCrostie, J. 2010. "The TOEIC in Japan: A Scandal Made in Heaven." *Shiken: JALT Testing & Evaluation SIG Newsletter*, Vol. 14, No. 1, Feb. pp. 2-11. Tokyo: JALT. Mar. 2, 2010. <http://jalt.org/test/mcc_1.htm>.

New TOEIC test premieres in Japan and Korea; all TOEIC versions are equally valid worldwide. ETS.org. Apr. 3, 2010. <http://www.ea.toeic.eu/toeic/ea/news/?news=805&view=detail>.

Nunnally, J.C. & Bernstein, I.H. 1994. *Psychometric Theory, 3rd Edition*. McGraw-Hill Series in Psychology. pp. 264-265. New York: McGraw-Hill, Inc.

Oh, Y.J, & Kang, S.W. 2009. “Korea to Replace TOEFL with State Tests.” Nov. 1. *The Korea Times*. Apr. 3, 2010. <http://www.koreatimes.co.kr/www/news/nation/2009/12/117_54652.html>.

Powers, D.E., Kim, H.J., & Weng, V.Z. 2008. *TOEIC Can-Do Guide--Executive Summary. The Redesigned TOEIC Listening and Reading Test*. ETS.org. Dec. 20, 2008. <http://www.ets.org/Media/Tests/Test_of_English_for_International_Communication/TOEIC_Can_Do.pdf>.

Registration Opens in Japan for TOEIC Speaking and Writing Tests. CCUN. Dec. 20, 2008. <http://www.ccun.com.cn/gjpd/2007-05-10/content.1178777110000d8717.html>.

Research and Design. 2010. ETS.TOEFL. Dec. 10, 2010. <http://www.ets.org/toefl/institutions/about/research_design/>.

Saegusa, Y. 1985. “Prediction of English Proficiency Progress.” *Musashino English and American Literature*, Vol. 18. pp. 165-185. Tokyo: Musashino Women’s University.

Sample (TOEIC). 2006. ETS.TOEIC. Sept. 29, 2010. <http://www.ets.org/Media/Tests/TOEIC/pdf/TOEIC_LR_sample_tests.pdf>.

Sample (TOEIC Bridge). 2010. ETS.TOEIC Bridge. Sept. 29, 2010. <http://www.ets.org/Media/Tests/TOEIC_Bridge/pdf/Bridge_Sample_Test.pdf>.

Schumacker, R. 2005. “Standards for Interpreting Reliability Coefficients. Applied Measurement Associates.” Apr. 10, 2010. <http://www.appliedmeasurementassociates.com/White%20Papers/Standards%20for%20Interpreting%20Reliability%20Coefficients.pdf>.

South Koreans Take First TOEIC Speaking and Writing Tests. ETS. Dec. 20, 2008. <http://korea.etsasiapac.org/09122006>.

Taylor, K. “ESL Exams: A Teacher’s Guide.” Leo Network. Dec. 20, 2008. <http://www.learnenglish.de/Teachers/eslexams.htm>.

Test Content (TOEFL). 2010. ETS.TOEFL. Dec. 10, 2010. <http://www.ets.org/toefl/institutions/about/content/>.

Test Content (TOEIC). 2010. ETS.TOEIC. Dec. 10, 2010. <http://www.ets.org/toeic/speaking_writing/about/content/>.

Tests & Products. 2010. ETS. Dec. 10, 2010. <http://www.ets.org/tests_products>.

TOEFL Equivalency Table. Multiply, Inc. Dec. 20, 2008. <http://genkeis.multiply.com/journal/item/209>.

TOEFL Equivalency Table (2). Vancouver English Centre. Dec. 20, 2008. <http://secure.vec.bc.ca/toefl-equivalency-table.cfm>.

TOEIC Newsletter, No. 105. Digest Version. Special Feature. 30 Years of TOEIC. Nov. Tokyo: IIBC & ETS. Feb. 15, 2010. <http://www.toeic.or.jp/toeic_en/pdf/newsletter/newsletterdigest105.pdf>.

TOEIC Speaking and Writing Tests Launched in the UK. ETS. Dec. 20, 2008. <http://www.uk.toeic.eu/no_cache/toeic-sites/toeic-default/news-toeic/?news=694&view=detail>.

TOEIC Technical Manual. 1998. Princeton: The Chauncey Group International & ETS. Dec. 20, 2008. <http://www.toeic.cl/images/toeic_tech_man.pdf>.

TOEIC Test Data & Analysis 2004. 2005. Tokyo: IIBC & ETS.

TOEIC Test Data & Analysis 2005. 2006. Tokyo: IIBC & ETS.

TOEIC Test Data & Analysis 2006. 2007. Tokyo: IIBC & ETS.

TOEIC Test Data & Analysis 2007. 2008. Tokyo: IIBC & ETS.

TOEIC Test Data & Analysis 2008. 2009. Tokyo: IIBC & ETS. Feb. 15, 2010.
<http://www.toeic.or.jp/toeic_en/pdf/data/TOEIC_DAA2008.pdf>.

TOEIC Test Data & Analysis 2009. 2010. Tokyo: IIBC & ETS. Sept. 22, 2010.
<http://www.toeic.or.jp/toeic_en/pdf/data/TOEIC_DAA2009.pdf>.

TOEIC User Guide: Listening & Reading. 2007. ETS.TOEIC. Dec. 20, 2008.
<http://www.ets.org/Media/Tests/Test_of_English_for_International_Communication/TOEIC_User_Gd.pdf>.

Who We Are. 2010. ETS. Dec. 10, 2010. <http://www.ets.org/about/who/>.

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: ETS. Apr. 8, 2010.
<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: ETS. Dec. 20, 2008.
<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*, Vol. 34, No. 6, Nov./Dec. pp. 41-45. Tokyo: JALT.

Woodford, P. 1982. *TOEIC Research Summaries: An Introduction to TOEIC: The Initial Validity Study*. Princeton: ETS. Dec. 20, 2008.
<http://www1.ets.org/Media/Research/pdf/TOEIC-RS-00.pdf>.

Appendix A

There were two minimum criteria for students to be eligible to pass any of the three required first-year English courses.

1) Each student was required to attend at least two thirds of the classes for a course in order to pass that course. (However, if a student achieved a TOEIC score of 730 or higher, then this criterion was waived.)

2) Each student was required to achieve a TOEIC score of at least 220 in order to pass any of the three required English courses.

If the above criteria were met, then each teacher individually determined each of her/his students' final grades based on their classwork, attendance, quiz scores, homework, etc., within the parameters of the chart below.

TOEIC Score Range	Final Grade Range
220 - 339	40 - 75
340 - 469	50 - 80
470 - 599	60 - 85
600 - 729	70 - 90
730 - 859	80 - 95
860 – 990	90 - 100