Have the Changes Made to the TOEIC Caused Any Differences in the Ways It Assesses Test Takers' English Language Abilities?

Brian D. Bresnihan

Abstract

This paper attempts to determine if there are any differences between the ways the new version and the original, old version of the TOEIC measure test takers' English language abilities using the scores achieved by about 1,200 students attending a university in Japan. It begins with a brief history of the TOEIC, a description of the test and its scoring and the changes made to it, and a discussion of issues related to the reliability of the scores. Following this, the data is presented and analyzed from two perspectives related to the inquiry. It ends noting that, regarding the question of whether or not the changes to the TOEIC altered the way it assesses test takers' English language abilities, these analyses suggest the answer is tentatively yes, but that neither version seems to have a clear advantage over the other.

1 . Background Information about the TOEIC

The Test of English for International Communication (TOEIC) was first administered in December of 1979, in Japan. It was created by the same organization, the Educational Testing Service (ETS), that produces the TOEFL, another norm-referenced test of English language proficiency, which was first administered in 1964 in the United States, and is based on the same design. This newer test of English language proficiency by ETS was created following two requests from Japan. The first request was made in early 1977 by an individual, Yasuo Kitaoka. He felt that a new test of English, which focused on its use in business contexts, rather than academia, and which aimed at lower ability users than the TOEFL did, was needed in order to urge Japanese people to learn how to use English for communicative purposes in work-related situations so they would be better able to compete in the world of international business. Although people at ETS were interested, the request was turned down because Kitaoka was the owner of a for-profit company. As ETS was a non-profit organization, it was felt it would be inappropriate for the organization to produce the test for a for-profit

company.

Next, Kitaoka tried to get Japan's Ministry of Education (now Ministry of Education, Culture, Sports, Science and Technology) interested in his idea. As it was already supporting the Japan-based Society for Testing English Proficiency, Inc. (STEP) tests, it had no interest in the development of another English language test. When this endeavor came to nothing, a friend on the board of directors for the non-profit public interest corporation World Economic Information Services, which was under the direction of the Ministry of Trade and Industry (now the Ministry of Economy, Trade, and Industry), assisted him in gaining that ministry's support for the test. So, in the latter half of 1977, representatives from the Ministry of Trade and Industry and the non-governmental comprehensive economic organization Japan Federation of Economic Organizations (now Federal Business Federation) met with people at ETS about Kitaoka's idea. Soon afterward, ETS began studying the possibilities for such a test and doing initial research. In Japan, Kitaoka's same friend became head of a newly formed non-profit public interest corporation within the World Economic Information Services, the TOEIC Steering Committee (now Institute for International Business Communication), which would be, and still is, responsible for advertizing, making arrangements for, and administering the TOEIC.[1]

The TOEIC is a norm-referenced standardized test of English proficiency. Therefore, it cannot be studied for in the way classroom tests and final exams, called criterion-referenced tests, can be studied for. Criterion-referenced tests assess the test takers' knowledge of a clearly defined, limited body of material which has been specified and made known to the test takers beforehand. Norm-referenced tests cover a much wider range of material than criterion-referenced tests, and the details of that material are not specified beforehand to the test takers. One can practice taking a norm-referenced test in order to become familiar with its format, but one cannot really study for the content of the test itself. If one's score on a norm-referenced test could be improved by studying specific material, then it would not be norm-referenced and the test would not be measuring the test takers' proficiency in the field being examined, only on the specific material that was studied. One's score on a norm-referenced test should not significantly increase unless one becomes more proficient in the field it is

[1] For more details about the beginnings of the TOEIC, see pages 14 to 16 of Bresnihan (2010), page 8 of Chapman (2004), McCrostie (2009), page 2 of McCrostie (2010), pages 6 and 18 of *TOEIC Newsletter, No. 105*, and page 2 of *TOEIC User Guide: Listening & Reading*.

assessing nor significantly decrease without a noticeable loss in proficiency. There is some overlap between these two types of tests, of course, but these two types of tests should be used for quite different purposes. In order for the TOEIC to be what ETS claims it to be, a test of English language proficiency, ETS does its best to be sure that it cannot be studied for it the way a criterion-referenced test can be.[2]

As just mentioned, the first TOEIC was administered near the end of 1979 in Japan. This was the TOEIC Secure Program (SP) Test. In 1981, a new type of TOEIC administration began, the TOEIC Institutional Program (IP) Test. The TOEIC IP Test has the same format and design as TOEIC SP Test. However, whereas the TOEIC SP Test can only be taken at certain locations and according to the schedule set by ETS, the TOEIC IP Test can be taken wherever and whenever an organization would like it to be held, as long as it has made arrangements far enough in advance. It is also somewhat less expensive than the TOEIC SP Test, and test takers can receive their scores much more quickly than they would if they had sat for the TOEIC SP Test. The disadvantages of the TOEIC IP Test are that ETS cannot guarantee the security of its administration, since it does not supervise the test taking procedure nor analyze the scores for oddities that might indicate cheating, and cannot guarantee the reliability of the scores obtained because it is constructed entirely from sets of questions from already administered tests and is administered to a restricted, not a general, population. Though neither test is inherently easier or more difficult than the other, these are important distinctions for TOEIC test takers and TOEIC score users to be aware of. Yet, it seems that many either ignore these differences or do not know of them. Even in much of the published research, it is unclear which type of test is being reported on.[3]

In January of 1982, the first administration of the TOEIC in South Korea was held. This was the first time the test was sat for outside of Japan. Initially, most test takers in South Korea took the TOEIC IP Test. Within one year of its debut, the TOEIC IP Test was also taken by more examinees in Japan than was the TOEIC SP Test. This has continued to be the case in Japan. Yet, the situation has not continued this way in South Korea. Nowadays, nearly all examinees in South Korea take the TOEIC SP

[2] Because of the confusion over this matter in Japan related to English language testing, Brown wrote an article specifically about this (Brown) in a book he edited and was published in Japan (Brown & Yamashita). Also, see Wood for comments by another language testing expert on related matters.

[3] For these details about the TOEIC IP Test and the differences between it and the TOEIC SP Test, see pages 2, 3, and 7 of *TOEIC Newsletter, No. 105*, page 1 of *TOEIC Test Data & Analysis 2009*, page 8 of Chapman (2004), "Differences between SP group application and IP," and "Group Application."

Test. Also, more people sit for the TOEIC in South Korea than in Japan. Considered worldwide, about 80% of the approximately five million yearly TOEIC test takers are in Japan or South Korea, though the TOEIC is now available in about 90 countries.[4]

The first changes ever made to the design of the TOEIC appeared in the TOEIC SP Test in May of 2006. These changes were implemented in the TOEIC IP Test in April of 2007. According to researchers for ETS, the test was revised "in order to better align test questions with everyday workplace language scenarios and to provide test-takers with more information about their listening and reading proficiency levels. . . . (T)hese changes (were) intended to align the test more closely with theories of communicative competence. . . . The revision is thought to be a valid measure of international communication today." [5] The overall changes made were that some of both the listening and the reading texts used for the questions were made longer than before and that the listening texts and questions were recorded by native English speakers not only from North America, as they were previously, but also from Britain, Australia, and New Zealand.[6]

The TOEIC was and is still divided into seven parts. The first four parts continue to assess listening ability. It takes examinees about 45 minutes to answer these 100 multiple-choice questions. The last three parts continue to assess reading ability. The examinees have 75 minutes to complete these 100 multiple-choice questions.

Part 1 used to contain twenty items, each with one photograph printed in the text booklet and four one-sentence options recorded on a tape from which to choose the one option that best described the photograph. There are now only ten items, using the same format, in this part. Part 2 was not changed. It contained and still contains thirty items, each with one statement or question along with three one-sentence options recorded on a tape from which to choose the one option that is the best response to the initial statement or question. Part 3 used to contain thirty items, each with one short

[4] For these details about the TOEIC in South Korea and TOEIC test takers in Japan, South Korea, and worldwide, see pages 3 and 7 of *TOEIC Newsletter, No. 105* and page 1 of *TOEIC Test Data & Analysis 2009*. For information about possible changes in situation concerning the TOEIC in South Korea, see Kang, Lee, and Oh & Kang.

[5] This quotation is on page 4 of Powers, Kim, & Weng.

[6] For a complaint that the changes made to the TOEIC were not extensive enough, see Chapman & Newfields. There are now other TOEIC tests in addition to the standard TOEIC, which is a listening and reading test. There is the TOEIC Speaking and the TOEIC Writing, which are administered together. See "About the TOEIC Speaking and Writing Tests," *Examinee Handbook: Speaking & Writing,* and "Speaking and Writing: Sample Tests" for details about these two tests. Also, there is a TOEIC test for English language learners of lower ability, the TOEIC Bridge. See "About the TOEIC Bridge," *Examinee Handbook*, and "Sample (TOEIC Bridge)" for details about this test.

conversation recorded on a tape and one question with four options printed in the text booklet from which to choose the one option that best answered the question based on the conversation. Now, there are still thirty items in Part 3, but only ten longer conversations with three questions per conversation. Otherwise, the format is the same. Part 4 used to contain twenty items based on between six and nine short talks recorded on a tape, each with between two and four questions along with four options per question printed in the test booklet. The one best answer to each question based on the short talk was to be chosen. The new test contains thirty items based on ten short talks, each of which has three questions. Otherwise, the format is the same.

As for the reading section, everything is printed in the test booklet. Part 5 has remained unchanged. It contained and continues to contain forty items, each with a single sentence with a blank in it followed by four options from which to choose the one best option to complete the sentence. Part 6 used to contain twenty items, each with a single sentence in which four words and/or groups of words were underlined. One of the underlined words and/or groups of words was an error in the sentence, which was to be chosen. Now, Part 6 contains twelve items based on four texts, each with three blanks and four options per blank from which to choose the one best option to complete each of the sentences and texts. Part 7 used to contain forty items based on a number of texts, each with between two and four questions along with four options per question from which to choose the one best answer to each based on the text. Part 7 now contains forty-eight items of two types. Twenty-eight items are of the same format as existed previously. Twenty items are based on four pairs of texts, each of which has five questions with four options per question from which to choose the one best answer to each based on the pair of texts.[7]

Examinees still receive three reported scores, not raw scores but scaled scores, after taking the TOEIC. One is a listening score based on their answers to the questions in parts 1 to 4, which will be between 5 and 495, inclusive. There is also a reading score based on their answers to the questions in parts 5 to 7, which will also be between 5 and 495, inclusive. They also receive a total score based on all of their answers to the questions on the test, which is calculated by simply adding the listening score and the reading score, and so will be between 10 and 990, inclusive. At around the same time

[7] These details about the TOEIC test items and the changes made to them come from pages 32 to 34 of Chapman & Newfields, *Examinee Handbook: Listening & Reading*, page 4 of Powers, Kim, & Weng, "Sample (TOEIC)," "Test Content (TOEIC)," and pages 3 and 4 of *TOEIC User Guide: Listening & Reading*.

that the changes to the TOEIC were implemented, ETS released a statement saying that "(a)ll TOEIC versions are equally valid and reliable," indicating that it considers scores from before and after the changes to the TOEIC to indicate equivalent measures.[8]

There are three other questions of interest concerning the reported scores. One is: How are the ranges of the reported scores determined? These ranges were created by ETS and used to determine the reported scores of the first test takers of the TOEIC. Describing the results of that first administration, ETS states:

> "The TOEIC scale has a range from 5 to 495 for each section. For the Listening Comprehension section the observed range--the scores actually obtained by examinees--went from a low of 40 to a high of 495. . . . The observed range of scaled scores for the reading section was from a low of 5 to a high of 455. . . . The total score for TOEIC is the sum of the two section scores . . . is quite gratifying to note that the scale functions as intended. Almost all points on the scale are utilized for both sections as well as for the total score." [9]

Another question concerning the reported scores is: Why is the total score determined by simply adding the listening score and the reading score? This is what ETS has to say about this question from its initial published study on the TOEIC:

> "It should be noted that because an examinee's listening comprehension and reading comprehension scores could be compared to each other, the section scores were scaled in such a way that the means and the standard deviations for the two sections are equal. An important result of this procedure is that the two sections have equal weight or importance in the total score." [10]

---

8 This quotation is from "New TOEIC test premieres in Japan and Korea; all TOEIC versions are equally valid worldwide."
9 This quotation is on page 9 of Woodford.
10 This quotation is on page 6 of Woodford.

There is nothing prior to this statement explaining why or how the listening scores and reading scores were or were able to be compared. However, it is later explained that "(t)he mean scaled (listening) score was 290" and "(t)he mean scaled (reading) score was 288. (No real score of 288 exists since all scores are reported in multiple of 5. A 288 score would be reported as 290.)" [11] Also appearing later is the following:

> "The reliability of the listening comprehension section was 0.916 . . . For the reading section, the reliability was 0.930 . . . Total test reliability was estimated at 0.956 . . . These reliabilities are well within the generally accepted limits for measurement of individual achievement. . . . The correlation between the two sections was 0.769 for the analysis sample. This would indicate that each score provides somewhat different information about the examinee and justifies reporting separate scores." [12]

The last question about the reported scores is: How can the scores from different forms of the test be considered equivalent? Of course, the TOEIC cannot be exactly the same test, cannot contain exactly the same questions, each time it is administered. Yet, the scores on different administrations of the test have to be considered to be equally correct or accurate for the scores to be used to measure ability or progress or to compare the abilities or amounts of progress of individuals or groups of individuals. ETS explains it this way:

> "The raw scores on every form of TOEIC will be converted to the common scale established at the first administration. . . . A statistical procedure called 'score equating' will be used to determine the appropriate conversion formula for each new form so that a given converted score . . . will represent the same level of ability regardless of the form taken or the ability level of the group with whom it was taken." [13]

[11] This quotation is on page 9 of Woodford.
[12] This quotation is on page 8 of Woodford.
[13] This quotation is on pages 5 to 6 of Woodford.

> "Each TOEIC test form is equated back to two older TOEIC test forms by incorporating a block of items from each old form in the new test form. The equaters for each TOEIC test form are chosen by test developers based upon item reliability (*r*-biserials and delta values) and upon test specifications. A series of computations are used to equate the test forms. The equating computations are applied to a hypothetical sample, known as the 'equalized group.' . . . TOEIC test scores are reported on a scale which was instituted on the first TOEIC test administration . . ." [14]

> "Statistical analysis is conducted after each . . . Secure Program administration . . . and a unique raw-score-to-scaled-score conversion table is created for each test form based on statistical equating results. . . . As each test form will be reused multiple times in different areas of the world, the TOEIC program has a policy not to release test forms, for security reasons. Since test forms are not released to the public, the conversion table that is unique to each test form is not released either. A conversion table, used for reporting scaled scores, is of no practical use to the public when the test form to which it applies is not disclosed. In fact, the conversion table may fall subject to misuse if it is applied to the wrong test form and results in incorrect scaled scores being calculated." [15]

Other researchers have not been able to duplicate the strong reliabilities that ETS has reported. For example, although explaining why he could not be as accurate as ETS can be, Childs found that the reliability of the TOEIC total scores used in his study of company workers was 0.57,[16] far weaker than the 0.956 reported by ETS.[17] The present author found correlations between TOEIC listening section scores and the reading

[14] This quotation is on page II-5 of *TOEIC Technical Manual*. It is preceded from the page before by details similar to what was quoted from pages 5 to 6 of Woodford immediately above this quote. The entire passage is repeated word for word in Chapman's interview of a manager from the Institute for Business Communication (Chapman, 2004).
[15] This quotation is from "Frequently Asked Questions About the TOEIC Listening and Reading Test."
[16] This figure is from page 100 of Childs.
[17] This figure is from page 8 of Woodford.

section scores of university students in an earlier study to be between 0.35 and 0.49 for first-time test takers and to be between 0.47 and 0.53 for second -time test takers. [18] ETS reported this correlation to be 0.769. [19] "(N)ote that, as of the secure administration of . . . April 1998 . . . statistics are run on first-time test takers only, rather than on the total group of testers, as was previously the practice. First-time testers provide more unbiased results, as they have never previously been exposed to the TOEIC test or any of the test items." [20] This is another issue to consider when using TOEIC scores to judge people's English language abilities as most people will take the test more than once in hopes of obtaining a higher score. It is also a fact that a person's TOEIC scores can be quite variable across a number of administrations. [21] Childs even wrote, "Students may be counseled that if they take the test several times, they can expect that by chance alone they will achieve a score that is higher than their true score." [22]

The present author also attempted to find out if any differences could be detected in the TOEIC scores used in the same previously mentioned study between tests administered before the changes to the TOEIC and those administered after the changes were implemented. The basic statistics from tests taken before and after the changes and from the first test administration and the second yielded means and medians that were well within the associated standard errors of difference of each other for all three scores, listening, reading, and total, which are stated by ETS to be about 35 points for the Listening scores and for the Reading scores and are assumed by this author to be about 70 points for the Total scores. However, the total scores' and listening scores' means and medians of the group that took the tests after the changes increased more than those of the group that sat for the tests before the changes, though not in amounts greater than the standard errors of difference. This was not the case for the reading scores' means and medians, which increased almost exactly the same amounts for both groups. [23]

For the scores achieved before the changes to the TOEIC were implemented, the correlation between the listening scores and the reading scores on the first test was 0.45

[18] These figures are from pages 96, 121, and 122 of Bresnihan (2010).
[19] This figure is from page 8 of Woodford.
[20] This figure is from page VIII-1 of *TOEIC Technical Manual*.
[21] For examples of such variations in scores, see pages 75 to 86, 114 to 120, 158 to 164, 184 to 189, 195 to 199, and 308 to 328 of Bresnihan (2010), pages 69 to 70 of Childs, and pages 172, 174, and 179 of Saegusa.
[22] This quotation is on page 74 of Childs.
[23] For the details of these differences, see pages 126 to 132 of Bresnihan (2010). For details concerning the standard errors of difference, see pages IV-4 to IV-7 of *TOEIC Technical Manual*.

and on the second test was 0.49. For scores achieved after the changes to the TOEIC were implemented, these correlations were 0.38 for between the first test listening and reading scores and 0.50 for between the second test listening and reading scores. From these results, it did not seem that the changes to the TOEIC improved or worsened its capabilities to measure those students' English language abilities, compared with before the changes were implemented, and therefore indicated that the test still did not assess those students English language abilities well.[24]

2 . Materials, Procedures, and Purposes

The TOEIC IP Test scores used in this study are from tests which were taken by first-year non-English majors in one department at a university in Japan.[25] These students were required to take the TOEIC IP Test administered at the school at approximately the end of the eleventh week of each semester as a part of their mandatory English courses. Without a TOEIC score, the students could not pass any of their three compulsory English courses, as each student's TOEIC Total score was used in determining her/his final grades. (See Appendix A for details.)

All of the first-year students were enrolled in three independent required English courses. Each course met once a week for 90 minutes throughout the two 15-week semesters. One of the courses emphasized listening, with some speaking. Another emphasized reading, with some discussion. The third course emphasized grammar, with some writing. The department taught eight sections or classes of each course. Students were placed in classes based on their student identification number, mostly in numerical order. Each class had about 25 students. No student had the same teacher for two different courses.

In Japan, the school year begins in April, and there is a summer vacation between the two fifteen-week semesters of the university school year. Therefore, these students had approximately 11-weeks of their three English classes before taking the TOEIC IP Test the first time. Before taking it the second time, they had approximately 4 more weeks of English classes, 11 weeks of no classes, and then 11 more weeks of English classes. The attendance requirement for each of these three English courses in order to be eligible to pass was 67%.

[24] These figures are from pages 132 to 136 of Bresnihan (2010).
[25] A large portion of these scores was also used by the present author in an earlier study, Bresnihan (2010), and all were used in Bresnihan (2012).

Approximately 2,400 TOEIC IP Test scores were used in this study. About half of them were taken before changes were implemented in the TOEIC in 2007, and the other half of them were taken after this. Only the scores for students who sat for two tests in one year were included. This totaled about 200 pairs of scores per year over the six-year period from 2004 to 2009.

Microsoft Excel 2004 for Macintosh was used to generate basic statistics and correlations. JMP 5.0 for Macintosh was used to carry out one-way analyses of variance.[26] On line programs were used to calculate statistics related to effect size and associated correlations.[27] Any slight discrepancies in certain figures that may appear within and/or among the tables is due to rounding.

The main research question was: Did the changes to the TOEIC make it a better test of English language proficiency than it was before the changes were made? Evidence indicating that this might be the case would be increased stability (less variation) in individual's scores across test administrations (because changes of any kind would not be expected for most students due to there having been less classroom study than required between tests for noticeable increases to be expected to take place and due to there having been less time than necessary between tests for noticeable decreases to be expected to take place[28]) and increased similarity between individual's listening scores and reading scores on the same administrations (because similarity would be expected since listening and reading are both receptive abilities and are both important aspects of general or overall language proficiency, and vice versa). If, as ETS

[26] I am grateful to Michael Redfield for running the one-way analyses of variance for me and for helping me interpret the results.

[27] Cohen's d and related correlation coefficient figures were calculated using "Calculators" and "Calculators (2)."

[28] Concerning the former, the only large-scale study that has attempted to estimate the amount of classroom English language study time needed for learners to increase their English language abilities as measure by a standardized test, which was the TOEIC, certain amounts was done by Saegusa. On page 174 of Saegusa, he states that "less than 80 hours of instruction is not very effective. In such cases, 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." On page 42 of Wood, the language testing expert being interviewed says that "(n)either TOEIC nor Bridge are designed for re-testing with less than 90-120 hours of instruction time in between each attempt. On pages 213 to 214 of Bresnihan (2010), the present author points out that Saegusa's estimates are actually too low because of the use of a wrong constant in his calculations. With 100% attendance, the maximum amount of classroom English language study time these students engaged in between the two test administrations in their three required English courses combined was about 67.5 hours, with a long summer break dividing this time in two, making it unlikely to be a single unit of study. This is explained in more detail on page 90 of Bresnihan (2010). Concerning the latter, ETS suggests on page 22 of *Examinee Handbook: Listening & Reading*, in "Frequently Asked Questions About the *TOEIC Listening and Reading* Test," and on page 10 of *TOEIC User Guide: Listening & Reading* that a TOEIC score is acceptable for up to two years.

claims, both the old version and the new version of the TOEIC yield equivalent scores,[29] then the expectation is that neither version will prove to be better at assessing test takers' English language abilities than the other. Therefore, no differences should be found in the assessments of these students' scores on the two versions of the test.

3 . Examining the Stability of Scores across Test Administrations

Table 1 presents the basic statistics of the TOEIC IP Test Total scores, Listening scores, and Reading scores on the first and second test administrations for the two groups that will be considered in this paper. The first group includes the scores of 597 students who took the TOEIC IP Test twice in one year from 2004 to 2006, before the changes were made it. The second group includes the scores of 610 students who sat for the test twice in one year after the changes were implemented, from 2007 to 2009.

Table 1
Maximums, Minimums, Variations, Means, Medians, Standard Deviations, and Numbers of Scores Greater Than 3 Standard Deviations from the Mean

	2004-2006: n=597					
	Total1	Total2	Listen1	Listen2	Read1	Read2
Maximum	725	770	395	425	335	360
Minimum	160	230	100	115	55	75
Variation	565	540	295	310	280	285
Mean	457	462	254	253	203	209
Median	455	460	255	250	205	205
Stdv	83.6	90.5	47.1	50.1	51.2	54.8
No.>3 Stdv	1	2	0	3	0	0
No.<3 Stdv	1	0	1	0	0	0

	2007-2009: n=610					
	Total1	Total2	Listen1	Listen2	Read1	Read2
Maximum	820	835	435	465	385	385
Minimum	230	255	110	115	95	85
Variation	590	580	325	350	290	300
Mean	468	483	248	263	220	221
Median	465	480	245	260	220	220
Stdv	78.5	87.0	44.2	46.9	49.9	53.5
No.>3 Stdv	3	2	4	3	3	2
No.<3 Stdv	1	0	1	2	0	0

[29] The claim of equivalency for the original and new TOEIC scores can be found in "New TOEIC test premieres in Japan and Korea; all TOEIC versions are equally valid worldwide."

Comparing the first two rows of each group, we see that all of the corresponding maximum and minimum scores in the second group are higher than those in the first group, except for the minimum Listening score on the second administration, which is the same (115 points). Comparing the two third rows shows us that the variations in scores is always greater for the second group than the first, with the differences in the variations of the Reading scores being the smallest (10 points on the first test administration and 15 points on the second test administration). When comparing the means and medians of the two groups, we see that all of them are larger for the second group, except for those of the Listening scores on the first test administration, which are larger for the first group instead (254 points and 255 points versus 248 points and 245 points, respectively). Looking at the last two rows for each group, there more scores which are greater than three standard deviations from the mean for all of the scores on the second test administration than the first, except for the Total scores, which have the same number (two are greater than three standard deviations above the mean). From this initial survey of the data, it appears possible that either the students who sat for the tests after the changes were implemented were better at English than those who sat for the test before the changes were implemented or the changes made to the test resulted in forms that included questions that these students found easier to answer, although all but one of the differences in the these figures are less than one standard error of difference, which is about 35 points for Listening scores and Reading scores and is assumed by this author to be about 70 points for Total scores, as mentioned earlier. The exception is for the minimum Reading score on the first test administration, which is 40 points higher in the second group than in the first group.

It is also of interest to compare where most of the scores fall within the possible range of scores for each group. This is shown in Table 2. All of the ranges of the variations in scores for about two thirds of the scores surrounding the means and medians for the corresponding test administrations are higher for the second group, except for the first test administration Listening scores, which are almost the same, with the only difference being that the upper end of the range is five points higher for the first group than the second (295 points versus 290 points, respectively) while the lower end is the same for the two groups (210 points). Otherwise, the differences between the lower ends of the ranges of variations in scores is always greater than the differences between the upper ends, with those of the second group always being

higher for both. Recalling the means and medians just mentioned above, these ways the ranges of variations in scores vary from each other is to be expected. Also, these differences all fall within one standard error of difference of each other. However, we see that the spread of the variations of these scores is always a little wider for the first group than for the second and that the differences in the spreads is always greater on the second test administration than the first. Unlike a test given by a class teacher to assess students' knowledge and abilities related to the materials that had been taught and studied, where the teacher would hope for and expect most students' scores to gather between 75% or 80% and 100%, a standardized test is generally considered better the more it spreads out the scores of the test takers. From this point of view, it appears that the old form of the test was slightly better than the new form. In addition, all of the spreads are greater on the second test administration than the first, except for the spread of the Listening scores of the second group, which are the same (80 points). This tells us that the there is greater variability in two thirds of the scores surrounding the means from the second test administration than from the first.

Table 2
Ranges of Variations in Scores for Two Thirds of the Scores Surrounding the Means/Medians

Total Scores

Group	Test1	%	Spread	Test2	%	Spread
2004-	380-530	66.2	150 pts	375-545	67.2	170 pts
2007-	395-540	67.5	145 pts	400-560	67.5	160 pts

Listening Scores

Group	Test1	%	Spread	Test2	%	Spread
2004-	210-295	65.0	85 pts	205-300	66.5	95 pts
2007-	210-290	68.7	80 pts	225-305	67.0	80 pts

Reading Scores

Group	Test1	%	Spread	Test2	%	Spread
2004-	155-250	67.5	95 pts	155-260	68.0	105 pts
2007-	175-265	65.9	90 pts	170-265	66.7	95 pts

Table 3 presents basic statistics concerning the differences in each student's scores on the two test administrations when the first scores obtained are subtracted from the second scores obtained. The maximum increase in Total score for the first group is 200 points, while for the second group it is 255 points. The maximum decrease

in Total score for the first group is -195 points and for the second group is -235 points. The first group's maximum increase in Listening score is 150 points and is 185 points for the second group, while the maximum decrease is -130 points for the first group and -115 points for the second group. For the Reading scores, the maximum increase is 135 points for the first group and 140 points for the second group, with the maximum decreases being -130 points and -150 points, respectively. The variations in these maximum and minimum changes in scores are always greater for the second group, with the spread being particularly wide for the changes in Total scores. The variation in the changes of Total scores is 395 points for the first group, while it is 490 points for the second group. All of each groups's corresponding means and medians are nearly the same. Although none of these are very different between the two groups, the first group's changes in Listening scores and changes in Reading scores means are -1 points and 6 points, respectively, while they are 15 points and 0 points, respectively, for the second group. The differences it these means are reflected in the changes in Total scores means being 5 points for the first group and 15 points for the second group. Almost all of these figures indicate that there is greater variability between the scores on these tests for the second group's students than for the first group's students.

Table 3
Maximums, Minimums, Variations, Means, Medians, Standard Deviations, and Numbers of Scores Greater Than 3 Standard Deviations from the Mean for Changes in Scores

	2004-2006: n=597			2007-2009: n=610		
	T2-T1	L2-L1	R2-R1	T2-T1	L2-L1	R2-R1
Maximum	200	150	135	255	185	140
Minimum	-195	-130	-130	-235	-115	-150
Variation	395	280	265	490	300	290
Mean	5	-1	6	15	15	0
Median	5	0	5	15	15	0
Stdv	60.2	41.2	40.7	63.9	41.4	43.1
No.>3 Stdv	2	2	2	1	2	1
No.<3 Stdv	1	3	1	1	1	1

Table 4 also concerns the numbers of students whose scores are the same and are different on the two test administrations for both groups. There are very small numbers of students, between 3% and 6%, whose scores remained the same on both test

administrations for both groups, as shown in the third column. Comparing the figures in the second column to those in the fourth, the numbers of students whose scores increased are always more than the number whose scores decreased. This could be due to increased abilities, but a more likely cause is greater familiarity with the actually taking of the test, as it was the second time the students sat for it. As mentioned earlier, for this reason, ETS now only uses scores from people who sat for the test for the first time in its statistical studies. However, these differences between the numbers of students with increases and decreases in scores are not very great in most cases. Two of the numbers of increases and decreases in scores are almost the same (for the first group's Listening scores, 49% and 47%, respectively, and for the second group's Reading scores, also 49% and 47%, respectively) while one varies quite a bit (for the second group's Listening scores, 62% and 32%, respectively) and the others vary by amounts in between (for the first group's Reading scores, 54% and 42%, respectively, for the first group's Total scores, 53% and 43%, respectively, and for the second group's Total scores, 57% and 40%, respectively). Because only the number of students in the second group whose Listening scores are higher on the second test administration is noticeably larger than the number whose Listening scores are lower, the scores achieved by these students after the test was changed indicate either greater increases in English language abilities, which would not be predicted, or greater ability to find the correct answers, due simply to greater test taking abilities, for those students who took the test after the changes to the TOEIC were made than before. To a lesser degree, we see the same differences for the first group's changes in Reading scores when compared to the second group's changes in Reading scores.

Table 4
Numbers of Students Whose Scores Are Different
on Test 2 Than on Test 1 and Ranges and Means of the Changes

Total Scores

Group	Number Increased		Number No Change		Number Decreased		Maximum Increase	Maximum Decrease	Mean Change
2004-	316	53%	22	4%	259	43%	200	-195	5
2007-	349	57%	17	3%	244	40%	255	-235	15

Listening Scores

Group	Number Increased	Number No Change	Number Decreased	Maximum Increase	Maximum Decrease	Mean Change
2004-	292 49%	23 4%	282 47%	150	-130	-1
2007-	381 62%	34 6%	195 32%	185	-115	15

Reading Scores

Group	Number Increased	Number No Change	Number Decreased	Maximum Increase	Maximum Decrease	Mean Change
2004-	321 54%	26 4%	250 42%	135	-130	6
2007-	297 49%	27 4%	286 47%	140	-150	0

The last column in Table 4 again shows us the quite small average changes in scores, which were mentioned while discussing Table 3. These average changes are all well within the standard errors of difference. However, looking at the fifth and sixth columns, at the maximum increases and decreases that also appeared in Table 3, we see, in all cases, that they are far outside two standard errors of difference, which would be 138 points for Total scores and 69 points for Listening scores and for Reading scores. As is demonstrated in this discussion of Table 3 and Table 4, averages often hide specifics. On the other hand, sometimes changes are, in fact, meaningless as they are too small to have any consequences or indicate any real differences. Therefore, it is very useful to also find out which and what percentages of the differences in scores indicate real changes in the scores and the test takers' English language abilities, as far as the assumptions associated with the test's statistics are concerned.

Table 5 displays the numbers of students whose changes in scores indicate statistically real differences in the scores with 68% and 95% certainty, respectively. Comparing column two to column four and column three to column five, we see that more scores increased than decreased in all cases except for the Listening scores for the first group and the Reading scores of the second group (but only by more than 69 points), in which slightly more scores decreased than increased. We also see that the differences between the numbers of increases and the numbers of decreases is small for the first group's Total scores (13% and 2% versus 10% and 1%, respectively) and Listening scores (17% and 5% versus 18% and 5%, respectively) and the second group's Reading scores (20% and 5% versus 19% and 7%, respectively) while it is somewhat larger for the second group's Total scores (17% and 3% versus 8% and 1%, respectively) and Listening scores (30% and 9% versus 11% and 2%, respectively) and the first group's

Reading scores (21% and 6% versus 13% and 4%, respectively). This seems to suggest that these students found taking the listening section of the test the second time easier after the changes were made than before and that these students found taking the reading section of the test the second time easier before the changes were made than after.

Table 5
Numbers of Students Whose Scores Are More Than 70 and 138 Points or More Than 35 and 69 Points Different on Test 2 Than on Test 1

Total Scores

Group	No. of Scores Increased by More Than 70		No. of Scores Increased by More Than 138		No. of Scores Decreased by More Than 70		No. of Scores Decreased by More Than 138	
2004-	77	13%	11	2%	61	10%	7	1%
2007-	104	17%	18	3%	50	8%	6	1%

Listening Scores

Group	No. of Scores Increased by More Than 35		No. of Scores Increased by More Than 69		No. of Scores Decreased by More Than 35		No. of Scores Decreased by More Than 69	
2004-	100	17%	28	5%	105	18%	31	5%
2007-	183	30%	52	9%	69	11%	15	2%

Reading Scores

Group	No. of Scores Increased by More Than 35		No. of Scores Increased by More Than 69		No. of Scores Decreased by More Than 35		No. of Scores Decreased by More Than 69	
2004-	123	21%	37	6%	80	13%	22	4%
2007-	124	20%	32	5%	118	19%	42	7%

When combining the changes in scores in Table 5, we find with 68% and 95% confidence, respectively, that for Total scores, 23% and 3% of the students in the first group and 25% and 4% of the students in the second group scored truly differently on the two test administrations. For Listening scores with the same confidences, respectively, we find that 35% and 10% of the students in the first group and 41% and 11% in the second group performed truly differently on the two sittings. For Reading scores, we find 34% and 10% in the first group and 39% and 12% in the second group obtained scores that were truly different on the two administrations with 68% and 95% confidence, respectively. At the 68% confidence level, it appears that quite a large

number of these scores are different on the two test administrations for both groups, if one did not expect there to be very many noticeable differences. At the 95% confidence level, these amounts of differences are much smaller. They are very small for the Total scores, but still around 10% for the Listening scores and the Reading scores of both groups.

Table 6 presents the correlation coefficients and the covariability coefficients between the same scores, the Total scores, Listening scores, and Reading scores, on the two test administrations for both groups. Each of these coefficients for the same scores are a little higher for the first group. This indicates that the first group's scores on the two test administrations have a slightly stronger relationship with each other than do the same scores in the second group. All of the correlation coefficients, falling between 0.764 and 0.587, inclusively, suggest a medium, neither strong nor weak, relationship between the associated scores. The covariability coefficients suggest that between about one third (0.345, 0.412, 0.429) and one half (0.584, 0.498, 0.500) of the scores on one test administration can account for or predict the scores on the other test administration.

Table 6
Correlations and Covariabilities Across Test Administrations

	2004-2006: n=597		2007-2009: n=610	
	Correlation	Covariability	Correlation	Covariability
T1 & T2	0.764	0.584	0.706	0.498
L1 & L2	0.642	0.412	0.587	0.345
R1 & R2	0.707	0.500	0.655	0.429

Correlation equals the summation of each listening score minus the average of the listening scores times each reading score minus the average of the reading scores all divided by the square root of the summation of the square of each listening score minus the average of the listening scores times the summation of the square of each reading score minus the average of the reading scores.
Covariability equals the correlation squared.

Table 7 displays the results of the one-way analyses of variance for the Total scores' means, Listening scores' means, and Reading scores' means, respectively, for both test administrations of the pre-changes and post-changes groups. The results of each of the three analyses indicated significant differences at the alpha=0.05 level in each of the data sets ((F(3, 2410)=10.7997, $p<0.0001$) for the Total scores' means, (F(3, 2410)=10.1163, $p<0.0001$) for the Listening scores' means, and (F(3, 2410)=16.0282,

p<0.0001) for the Reading scores' means). When Tukey-Kramer's Honestly Significant Difference procedures were carried out with alpha set at 0.05, one significant difference relevant to this study was found for the Total scores' means and one for the Listening scores' means, both for the second group's data. There were no relevant significant differences found for the Reading scores' means. This seems to indicate that there is more variability (less stability) in the Listening scores and, probably as a result, the Total scores of the test administrations from after the changes were made to the TOEIC IP Test than in the Listening scores and the Total scores from before the changes were implemented.

Table 7
One-way Analyses of Variance for Scores' Means
of Test 1 and Test 2 Administrations for Each Group, 2004-2006 and 2007-2009

Total Scores' Means

Source	Degrees of Freedom	Sum of Squares	Mean Square	F Ratio	Probability > F Ratio
Between Groups	3	234546	78181.9	10.7997	<0.0001
Within Groups	2410	17446659	7239.3		
Corrected Total	2413	17681204			

Alpha equals 0.05.

Listening Scores' Means

Source	Degrees of Freedom	Sum of Squares	Mean Square	F Ratio	Probability > F Ratio
Between Groups	3	67425.6	22475.2	10.1163	<0.0001
Within Groups	2410	5354272.1	2221.7		
Corrected Total	2413	5421697.7			

Alpha equals 0.05.

Reading Scores' Means

Source	Degrees of Freedom	Sum of Squares	Mean Square	F Ratio	Probability > F Ratio
Between Groups	3	132200.4	44066.8	16.0282	<0.0001
Within Groups	2410	6625896.6	2749.3		
Corrected Total	2413	6758097			

Alpha equals 0.05.

Table 8 presents information concerning the averages of the Total scores, the Listening scores, and the Reading scores for the two groups. In the third row, and as mentioned in the discussion of Table 3, we see that the differences in the means from

the first test administration to the second is greater for the Total scores and Listening scores of the second group than the first group (15 points and 15 points compared to 5 points and -1 points, respectively) and is greater for the Reading scores of the first group than the second group (6 points compared to 1 point, respectively). However, all three changes in means are quite small when considering that the range of possible scores is from 10 to 990 for the Total score and from 5 to 495 for the Listening score and the Reading score. The Cohen's d effect size figures in the eighth row indicate very weak relationships between all of these pairs of means, except for the Listening scores' means of the second group (0.329), which indicates a slightly better than weak relationship. However, when we look further at the correlation coefficients (all with absolute values between 0.010 and 0.162, inclusive) and the covariability coefficients (all between 0.000 and 0.026, inclusive) in the ninth and tenth rows, we find that almost no relationships at all are indicated between any of these pairs of means because the figures are so small.

Table 8
Means, Changes in Means, Standard Deviations, Standard Errors, Effect Sizes, Correlation Coefficients, and Covariability Coefficients for Each Pairing of Scores, 2004-2006 and 2007-2009

	Total Scores		Listening Scores		Reading Scores	
	2004-06 n=597	2007-09 n=610	2004-06 n=597	2007-09 n=610	2004-06 n=597	2007-09 n=610
Mean1	457	468	254	248	203	220
Mean2	462	483	253	263	209	221
Change	5	15	-1	15	6	1
Stdv1	83.6	78.5	47.1	44.2	51.2	49.9
Stdv2	90.5	87.0	50.1	46.9	54.8	53.5
Std Err1	3.4	3.1	1.9	1.8	2.1	2
Std Err2	3.7	3.5	2.1	1.9	2.2	2.2
d	0.057	0.181	-0.021	0.329	0.113	0.019
r	0.029	0.090	-0.010	0.162	0.056	0.010
r^2	0.001	0.008	0.000	0.026	0.003	0.000

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 effect size.
r equals d divided by the square root of the summation of d squared plus 4, the correlation coefficient.
r^2 equals r times r, the covariability coefficient.

Almost all of the means and medians, the maximum and minimum scores, and the

ranges of the variations in scores around two thirds of the means and medians suggested either that the new version of the TOEIC was a little easier for these students than the old version was or that those students who entered this school's department after the changes were made to the TOEIC were a little better at English than those who entered before the changes were implemented. This was confirmed by the ranges of two thirds of the scores surrounding the means and medians. These variations of two thirds of the scores surrounding the means and medians were a little higher for the scores obtained after the changes to the TOEIC than before. More students' scores were higher the second time they took the test than were lower in all cases, with the biggest difference between the two being for the Listening scores obtained after the changes were made to the TOEIC. The maximum gains and losses in scores and, also, the ranges of the differences in scores on the two test administrations were larger for the students who sat for the test after the changes were implemented than before except for the maximum loss of the Listening scores, which was larger for the students who took the test before the changes than after. Of these increases and decreases, the percentages of those that were found to be true changes with 68% and 95% confidence were slightly greater for the scores achieved after the changes were made than for those that were achieved before the changes. The correlation and covariability coefficients between the same scores on the two test administrations suggested that the scores obtained before the changes to the TOEIC were implemented had a slightly stronger relationship with each other than did the scores obtained after the changes were made. One-way analyses of variance and the subsequent Tukey-Kramer's Honestly Significant Difference procedures revealed that there was less stability in the Listening scores' means and in the Total scores' means between the two tests after the changes to the TOEIC were implemented than before, although the Cohen's d, correlation coefficient, and covariability coefficient figures indicated that there were from very weak to no relationships between any of the paired means in either group. Generally, it seems that these students' TOEIC IP Test scores were slightly more stable for the two tests that were administered six months apart before the changes to the test were implemented than subsequently.

4 . Examining the Similarities between Listening Scores and Reading Scores on the Same Test Administrations

In Table 1, where the basic statistics concerning the Total scores, Listening scores, and Readings scores are presented separately, we saw that in general the Listening scores tended to be higher than the Reading scores. Table 2 also confirmed this. For this reason, when comparing the Listening scores with the Reading scores, it was decided to subtract the latter from the former. Table 9 show basic statistics generated from these differences. There are very large variations in the maximum and minimum Listening score minus Reading score differences shown in the third row, in the first group of 390 points on the first test and of 335 points on the second test and in the second group of 310 points on the first test and 300 points on the second test. The average differences on the first administration are 51 points for the first group and 28 points for the second group. As the Listening score and Reading score are simply added to form the Total score, the latter average is closer to what would be a desired average difference in these scores. The average differences on the second administration are 44 points for the first group and 42 points for the second group, neither difference looking very desirable from the aforementioned point of view.

Table 9
Maximums, Minimums, Variations, Means, Medians,
Standard Deviations, and Numbers of Scores
Greater Than 3 Standard Deviations from the Mean
for Differences between Listening and Reading Scores

	2004-2006: n=597		2007-2009: n=610	
	L1-R1	L2-R2	L1-R1	L2-R2
Maximum	245	215	165	160
Minimum	-145	-120	-145	-140
Variation	390	335	310	300
Mean	51	44	28	42
Median	50	45	30	45
Stdv	51.9	53.3	52.2	50.4
No.>3 Stdv	3	1	0	0
No.<3 Stdv	1	1	1	2

Table 10 displays the ranges of the variations in the differences between the Listening scores and the Reading scores for about two thirds of the scores surrounding the means and medians of the differences in scores for both administrations for both groups. These variations are all 95 points or 100 points. The ranges are a bit lower for

the second group than the first, on the first test being from 0 points to 95 points around the mean/median of 51 points/50 points for the first group and from -20 points to 80 points around the mean/median of 28 points/30 points for the second group and on the second test being from -5 points to 95 points around the mean/median of 44 points/45 points for the first group and from -10 points to 85 points around the mean/median of 42 points/45 points for the second group. Neither group seems to have any advantages over the other concerning these ranges of variations.

Table 10
Ranges of Variations in Differences between
Listening and Reading Scores for Two Thirds of the Scores
Surrounding the Means/Medians

Group	L1-R1	%	Spread	L2-R2	%	Spread
2004-	0-95	67.0	95 pts	-5-95	68.2	100 pts
2007-	-20-80	65.9	100 pts	-10-85	68.5	95 pts

Table 11 concerns the numbers of students whose Listening scores and Reading scores are different from each other on the two test administrations for each group. As expected from what we have seen prior to this, there are many more students whose Listening scores are greater than their Reading scores on both test administration in both groups. Looking at the second and fourth columns, on the first administration, there are about six times as many students with Listening scores greater than their Reading scores than vice versa in the first group (being 496 students and 86 students, respectively) and more than twice as many in the second group (being 412 students and 175 students, respectively). On the second test administration, there are more than four times as many students with Listening scores greater than their Reading scores than the reverse in both groups (being 476 students and 107 students in the first group and 494 students and 102 students in the second group). The numbers of students with Listening and Reading scores that are the same, shown in the third column, is very small in all cases, being 2%, 4%, 2%, and 2%. The maximum higher Listening scores are quite a bit larger from the first group than the second, being 245 points and 215 points higher and 165 points and 160 points higher, respectively. The maximum higher Reading scores are exactly the same on the first administration for both groups, 145 points. On the second administration, the first group's maximum higher Reading score is 120 points higher, while the second group's is 140 points higher. From these

details, we can see that the Listening scores and Reading scores vary from each other in nearly equivalent amounts in both groups on both administrations, and on the second administration, in almost exactly the same ways.

Table 11
Numbers of Students Whose Listening and Reading Scores Are Different from Each Other on Test 1 and on Test 2 and Ranges and Means of the Differences

First Test Administration

Group	No. Listening Scores Are Higher		No. L & R Scores Are Equal		No. Reading Scores Are Higher		Maximum L Score Is Higher	Maximum R Score Is Higher	Mean Difference in Scores
2004-	496	83%	14	2%	86	14%	245	145	51
2007-	412	68%	23	4%	175	29%	165	145	28

Second Test Administration

Group	No. Listening Scores Are Higher		No. L & R Scores Are Equal		No. Reading Scores Are Higher		Maximum L Score Is Higher	Maximum R Score Is Higher	Mean Difference in Scores
2004-	476	80%	14	2%	107	18%	215	120	44
2007-	494	81%	14	2%	102	17%	160	140	42

Since the standard error of difference is about 35 points for both the Listening section and the Reading section of the TOEIC and since the two scores are simply added to calculate the Total score, as explained earlier, it seems possible to use the standard error of difference as a means to decide which differences in the two scores are likely to be real differences and which are not. Table 12 displays the numbers of students whose Listening scores and Reading scores differ by more than one and two standard errors of difference, by more than 35 and 69 points, on each of the two test administrations for both groups. Based on the differences in the two scores that we have seen earlier, it is not surprising to find that there are many more students whose Listening scores that are truly higher than their Reading scores than there are the opposite. In the first group on the first test administration, 62% of the students achieved Listening scores that indicate they truly scored higher on this section of the test than on the Reading section, that indicate they had stronger listening abilities than reading abilities, with 68% confidence, and 37% with 95% confidence, while with 68% confidence it seems 5% of the students scored truly better on the Reading section than the Listening section, had stronger reading abilities than listening abilities, and 1% with 95% confidence. In

the second group on the first test administration, 44% of the students seem to have scored truly higher on Listening section than the Reading section with 68% confidence and 23% with 95% confidence, while 10% of the students seem to have scored truly better on the Reading section with 68% confidence and 3% with 95% confidence. Adding these together, 67% and 38% of the students in the first group appear to have scored truly differently on the two sections with 68% and 95% confidence, respectively, on the first test administration. In the second group, 54% and 26% of the students appear to have scores that are truly different on the two sections with 68% and 95% confidence, respectively, on the first test. If one were expecting the students' English language abilities being measured by the two sections to be quite similar, then the new TOEIC seems to have measured their abilities a little better. However, there are large numbers of students with scores that are statistically considered truly different on the two sections in both groups. The percentages of students scoring differently on the two sections of the test on the second test administration are almost the same for both groups. With 68% and 95% confidence that the Listening score is higher, in the first group the figures are 55% and 33%, respectively, and in the second group they are 55% and 31%, respectively. With 68% and 95% confidence that the Reading score is higher, in the first group the figures are 7% and 3%, respectively, and in the second group they are 7% and 2%. Added together, 62% of the students in both groups appear to have scored differently on the two sections with 68% confidence, while with 95% confidence, it appears that 36% of the students scored differently on the two sections in the first group and 33% of the students in the second group.

Table 12
Numbers of Students Whose Listening and Reading Scores Are More Than 35 and 69 Points Different from Each Other on Test 1 and on Test 2

First Test Administration

	No. Listening Scores More Than		No. Listening Scores More Than		No. Reading Scores More Than		No. Reading Scores More Than	
Group	35 Pts.	Higher	69 Pts.	Higher	35 Pts.	Higher	69 Pts.	Higher
2004-	372	62%	220	37%	31	5%	7	1%
2007-	270	44%	142	23%	64	10%	17	3%

Second Test Administration

Group	No. Listening Scores More Than 35 Pts.	Higher	No. Listening Scores More Than 69 Pts.	Higher	No. Reading Scores More Than 35 Pts.	Higher	No. Reading Scores More Than 69 Pts.	Higher
2004-	327	55%	199	33%	40	7%	16	3%
2007-	338	55%	192	31%	44	7%	15	2%

Lastly, we will examine the reliabilities of the two sections' scores in comparison with each other, which are presented in Table 13. The first two rows concern the comparisons we have been considering above. On the first test administration, the correlation and covariability are nearly the same for the two groups, being 0.445 and 0.198, respectively, for the first group and 0.390 and 0.152, respectively, for the second group. These are in the very weak range. They are also very similar for the two groups on the second test administration, being 0.488 and 0.238, respectively, for the first group and 0.502 and 0.252, respectively, for the second group. These are slightly stronger than those for the first test administration, and indicate weak relationships. The third and fourth rows, of correlations and covariabilities for the same pair of abilities across administrations are included for comparison. We see that they are very similar to those in the first and second rows, suggesting that the assessments on the two test administrations are very similar to each other.

Table 13
Correlations and Covariabilities Across Abilities

	2004-2006: n=597		2007-2009: n=610	
	Correlation	Covariability	Correlation	Covariability
L1 & R1	0.445	0.198	0.390	0.152
L2 & R2	0.488	0.238	0.502	0.252
L1 & R2	0.426	0.181	0.422	0.178
L2 & R1	0.460	0.212	0.369	0.136

Correlation equals the summation of each listening score minus the average of the listening scores times each reading score minus the average of the reading scores all divided by the square root of the summation of the square of each listening score minus the average of the listening scores times the summation of the square of each reading score minus the average of the reading scores.
Covariability equals the correlation squared.

The average difference of the Listening scores minus Reading scores on the

first administration was smaller for the students who took the test after the changes were made to it than before the changes were made. Because the Listening score and Reading score are simply added to determine the Total score, the former average is a more desirable average difference in scores. On the second administration, the average differences in the Listening scores and Reading scores were almost exactly the same. Since the ranges of the variations in the differences between the Listening scores and the Reading scores for about two thirds of the scores surrounding the means and medians of the differences in scores for both administrations for both groups are nearly the same, no advantages were found for either group. There also seemed to be no advantages for either group when considering the numbers of students whose Listening scores and Reading scores were different from each other on the two test administrations, as the numbers of students whose Listening scores and Reading scores were the same on a given test administration were nearly the same for both groups. Although large numbers of students achieved scores that statistical analyses indicated were truly different on the two sections in both groups, there were somewhat fewer students who did so after the changes were implemented on the first administration. This would be favorable when expecting the students' English language abilities to be measured quite similarly on the two sections of the test. This difference did not appear between the two groups on the second test administration. The reliabilities of the two sections' scores in comparison with each other on the same administration varied from very weak to weak relationships, showing no advantages for either group. From these considerations, it seems that these students' Listening scores and Reading scores were slightly more similar to each other for the group that sat for the test after the changes were made than before.

5. Conclusions

This study investigated whether or not the changes made to the TOEIC resulted in it being better able to assess these university students' English language abilities than it did before the changes were implemented. In order to do this, the TOEIC IP Test scores of 1,207 students, who took the test twice within six months, 597 before the changes were made and 610 after the changes were made, were analyzed. The initial survey of the data, of the maximum and minimum scores, the means and medians, and the ranges of the variations in scores around two thirds of the means and medians,

suggested, though not significantly in most cases, either that the majority of the students who sat for the test after the changes were made were a little better at English than the majority who sat for the test before the changes were made or that the questions on the new version of the test were a little easier for a majority of the students to find the answers to than was the case on the original version of the test. When a number of the teachers were asked informally whether or not the students generally seemed to be better at English from the 2007 school year on compared to before that, none replied that their students had been worse at English in the earlier years. Also, almost all of these same figures, in addition to actual counting, indicated that the students' scores on the second test for both groups tended to be higher on the second administration than on the first, though again not significantly. Since the amount of classroom study time between the two test administrations was insufficient to expect many students to have improved their English language abilities a measurable amount with this test, and since the amount of time between the two tests was too short to expect many students to have lost measurable amounts of English language abilities, it was supposed that familiarity with the test, increased test taking abilities, was probably the cause of these general small increases in the scores.

When comparing the scores on the two test administrations for both groups of students, for those who took the test before the changes were implemented and for those who sat for the test after the changes were made, in many different ways, there were almost always slightly greater similarities in the scores for the former group than the latter. Almost all comparisons of simple gains and losses and significant gains and losses, including correlation, covariability, and Tukey-Kramer's Honestly Significant Difference figures, indicated more stability in the students' scores in the first group and more variation in the students' scores in the second group. As similarity was predicted based on the amount of English language study time and on the amount of calendar time between the two test administrations, these results suggested that the former version of the test assessed these students' English language abilities a little more accurately than the new version of the test.

Comparisons of the Listening scores with the Reading scores for each test administration for both groups of students revealed very few advantages for one version of the test over the other. The only differences were that the average difference between the Listening scores and the Reading scores on the first administration was

smaller, though not significantly, for the students who sat for the test after the changes were implemented than before and that there were fewer students in the second group whose two section scores on the first test administration were significantly different from each other than there were in the first group. Otherwise, all other considerations, such as the total numbers of students who Listening scores and Reading scores were different from each other on the same test administration, the ranges of the variations in these differences surrounding the means and medians, and the correlations and covariabilities, did not reveal any advantages for either group or version of the test. However, concerning the idea that there should be a close similarity between these Listening scores and Reading scores because listening abilities and reading abilities are both important aspects of overall language proficiency and because these two scores are simply added to obtain the Total score, these results suggested that the new version of the test assessed these students' English language abilities with slightly better accuracy than the prior version of the test.

The results of these analyses related to the two sub-questions investigated in this study suggested that the scores obtained by these students taking the two different versions of the TOEIC IP Test were slightly different. The results of the question concerning the expected stability of the students' scores achieved on two test administrations that were six months apart seemed to favor the original version of the test. The results of the question concerning the expected similarity of the Listening and Reading scores achieved by individual students on the same test administration seemed to favor the new version of the test. Perhaps more intriguing, though, was the fact that the scores obtained after the changes to the test had been implemented had a tendency to be higher than those obtained before the changes were made, even if the differences were not significant in most cases. As no other uniform tests had been given to use for comparison, it is impossible to determine the meaning of this. These results tentatively suggest that the TOEIC scores test takers are now receiving may not be as equivalent to the TOEIC scores test takers received before the changes to the test were implemented as ETS claims they are, though which version of the test better assessed the students' English language abilities could not be determined definitively.

References

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/>.

Bresnihan, B. 2010. *Possible Reliability Problems Affecting Use of TOEIC IP Test Scores*. Kobe: Institute for Policy Analysis and Social Innovation, University of Hyogo.

Bresnihan, B. 2012. "Using TOEIC Scores to Evaluate Student Performance in English Language Courses." *The Jimbun Ronshu,* Vol. 43. pp. 1-47. 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.

Brown, J.D., & Yamashita, S.O. (Eds.) 1995. *Language Testing in Japan*. Tokyo: JALT.

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, & 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>.

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>.

Examinee Handbook. 2007. ETS.TOEIC Bridge. Dec. 11, 2011.
<http://www.ets.org/Media/Tests/TOEIC_Bridge/pdf/TOEIC_BridgeExam.pdf>.

Examinee Handbook: Listening & Reading. 2008. ETS.TOEIC. Dec. 11, 2011.
<http://www.ets.org/Media/Tests/TOEIC/pdf/TOEIC_LR_examinee_handbook.pdf>.

Examinee Handbook: Speaking & Writing. 2009. ETS.TOEIC. Dec. 11, 2011. <http://www.ets.org/Media/Tests/TOEIC/pdf/TOEIC_Speaking_and_Writing_Examinee_Handbook.pdf>.

Frequently Asked Questions About the TOEIC Listening and Reading Test. 2011. ETS. TOEIC. Nov. 1, 2011. <http://www.ets.org/toeic/listening_reading/faq>.

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

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." April 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>.

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>.

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). ETS.TOEIC. 2006. 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>.

Speaking and Writing: Sample Tests. 2008. ETS.TOEIC. Dec. 11, 2011.
<http://www.ets.org/Media/Tests/TOEIC/pdf/TOEIC_sw_sample_tests.pdf>.

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

TOEIC Newsletter, No. 105. Digest Version. Special Feature. 30 Years of TOEIC. Nov. 2009. Tokyo: IIBC & ETS. Feb. 15, 2010.
<http://www.toeic.or.jp/toeic_en/pdf/newsletter/newsletterdigest105.pdf>.

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 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>.

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