ETS TOEIC®

Know English. Know Success.

COMPENDIUM STUDY

Comparison of Content, Item Statistics and Test-Taker Performance for the Redesigned and Classic TOEIC® Listening and Reading Tests

Chi-wen Liao, Natalie Hatrak and Feng Yu

January 2010

The Redesigned TOEIC® Listening and Reading test reflects the advances in English–language development observed in current and prospective employees in international businesses. New models of language proficiency have evolved, and the value of greater authenticity in language assessment has been recognized. The objective of the redesign of the TOEIC Listening and Reading test was to align the test with the developments in the field of English-language learning, while maintaining score comparability across redesigned and classic test versions. It is important to keep the scores on the Redesigned and Classic TOEIC Listening and Reading test on the same scale so that score users' day-to-day decision-making processes aren't interrupted. After years of research and evaluation, the Redesigned TOEIC Listening and Reading test was released in the May 2006 Secure Program (SP) administrations in Japan and Korea. This research report provides interim analyses results from the comparison of the Classic and the Redesigned TOEIC Listening and Reading test in three areas: (a) content, (b) item statistics and (c) test-taker performance.

Data

Analyses were performed in early 2007 on the 14 redesigned TOEIC SP forms administered from May 2006 through December 2006 and on the 22 classic TOEIC SP forms selected at random from January 2003 through December 2005. The sample sizes of the groups taking the Redesigned and the Classic TOEIC Listening and Reading test are shown in Table 1. The large sample sizes ensure stable comparison results.

As was the case with the Classic TOEIC Listening and Reading test, the redesigned test comprises two sections, listening and reading, and each section contains 100 items, for a total of 200. The score scale ranges for both the classic and the redesigned test are 5 to 495 points for each section, in increments of 5 points.

Table 1

Sample Sizes for Redesigned and Classic TOEIC Groups

	Classic TOEIC	Redesigned TOEIC
Japan	1,716,876	367,410
Korea	4,940,865	906,405
Total	6,657,741	1,273,815

Test Content

In the redesign process, some item types were retained, some were deleted or updated, and a small number of new item types were introduced. As shown in Table 2, more than 60% of the item types are common to both versions of the test.

The classic TOEIC listening section contained a total of 20 photo items used to assess language proficiency. Due to the low level of difficulty of the photo items, developers of the Redesigned TOEIC Listening and Reading test reduced the number of these items to 10. In the classic TOEIC listening section, the short conversations and short talks subsections were made up of discrete items that were based on short stimuli. To create a more authentic language text, the stimuli were lengthened and items were grouped into sets of three. In the reading section of the Redesigned TOEIC Listening and

Reading test, error-recognition items were removed in favor of text completion items, which allow for a more comprehensive evaluation of grammar and vocabulary skills. In the reading comprehension subsection, some of the longer passages were replaced by two related passages. See Chapter 2 for more detailed information regarding the content development process.

Table 2

Content Comparison Between the Redesigned and the Classic TOEIC Listening and Reading Test

	Classic TOEIC	Redesigned TOEIC
Listening		
Part 1	Photos: 20 discrete items	Photos: 10 discrete items
Part 2	Question-response: 30 discrete items	Question-response: 30 discrete items
Part 3	Short conversations: 30 discrete items	Conversations: 30 items 10 conversations, 3 questions each
Part 4	Short talks: 20 items 2-4 questions per talk	Talks: 30 items 10 talks, 3 questions each
Reading		
Part 5	Incomplete sentences: 40 discrete items	Incomplete sentences: 40 discrete items
Part 6	Error recognition: 20 discrete items	Text completion: 12 items 4 passages, 3 questions each
Part 7	Reading comprehension: 40 items 2-4 questions per passage	Reading comprehension: 48 items Single passages: 28 items 2-5 questions per passage Double passages: 20 items 5 questions per pair of passages

Reliability

Reliability refers to the extent to which assessment scores remain consistent over repeated administrations of the same form or alternate forms. Reliability also refers to the extent to which the assessment results are free from the effects of random variation caused by factors that may or may not be directly related to the purpose of the test (e.g., the kind of test given, the time of administration, test-takers' conditions or the scorers). There are different methods for calculating the reliability estimate for test scores. The reliability estimation for the TOEIC Listening and Reading test is based on an internal consistency method, and the reliability coefficient is called alpha. This calculated coefficient provides an indication of the consistency of test-taker responses to all of the items in each section. As seen in Table 3, the alphas estimated for the Redesigned TOEIC Listening and Reading test are as high as those for the Classic TOEIC Listening and Reading test. Overall, the reliability estimates for the Redesigned TOEIC Listening and Reading test are slightly higher than those for the Classic TOEIC Listening and Reading test.

TABLE 3

Reliability Estimates of Scale Scores for the Redesigned and Classic TOEIC Listening and Reading Test

Test section	Number of items	Classic TOEIC Listening and Reading test	Redesigned TOEIC Listening and Reading test
Listening	100	0.90–0.93	0.92–0.93
Reading	100	0.90–0.94	0.92–0.93
Total	200	0.93–0.95	0.95–0.96

Standard Error of Measurement

The standard error of measurement (SEM) is a measure of the tendency of test takers' scores to vary because of random factors, such as the particular selection of items on the form the test taker happened to take, or the particular scorers who happened to score a test taker's responses. The smaller the SEM is, the smaller the influence of these factors. The SEM is the average of the differences between the test takers' scores on one testing and the average of the scores they would get if they took the test many times without actually improving their ability. In a large group of test takers, about 2/3 of them will have scores within one SEM of those long-term average scores. The SEM and the test's reliability are inversely related. A high reliability coefficient would indicate a low standard error of measurement. Conversely, a low reliability coefficient would result in less consistent test scores and more room for variation in scores. The **estimated SEM for the Redesigned TOEIC Listening and Reading test remains around 25 scaled score points for each section** as in the Classic TOEIC Listening and Reading test.

Item Statistics

Item Difficulty

The item difficulty for the TOEIC Listening and Reading test is characterized by a statistical index called delta (Δ). Delta is defined as 13 + 4z, where z is the normal deviate corresponding to proportion correct. Delta values ordinarily range from 6.0 for a very easy item (i.e., approximately 95% of test takers select the correct answer) to 20 for a very hard item (i.e., approximately 5% of test takers select the correct answer) with a mean of 13.0 (50% correct). Table 4 shows means and standard deviations of delta values for the Redesigned and Classic TOEIC Listening and Reading tests. Figure 1 shows the delta distributions for listening items on both the Redesigned and Classic Listening and Reading tests, and Figure 2 shows the reading items. Although the redesigned forms are slightly more difficult than the classic forms, the overall shape of the item difficulty distributions for the Redesigned and Classic TOEIC Listening and Reading tests closely resemble each other. As was the case with the Classic Listening and Reading test, the spread of item difficulty in the Redesigned Listening and Reading test contains an appropriate proportion of easy, medium and difficult items.

TABLE 4

Average Item Difficulties for the Redesigned and Classic TOEIC Listening and Reading Test

		Classic TOEIC Listening and Reading test	Redesigned TOEIC Listening and Reading test
Mean	Listening	12.2	12.6
SD		1.6	1.4
Mean	Reading	11.7	12.4
SD		2.1	1.9

FIGURE 1

Listening Item Difficulty for the Redesigned and Classic TOEIC Listening and Reading Test

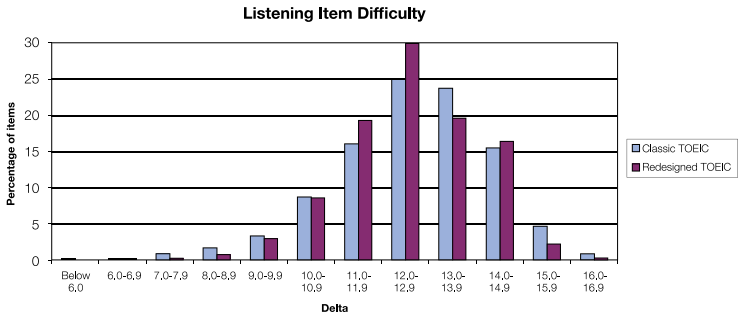
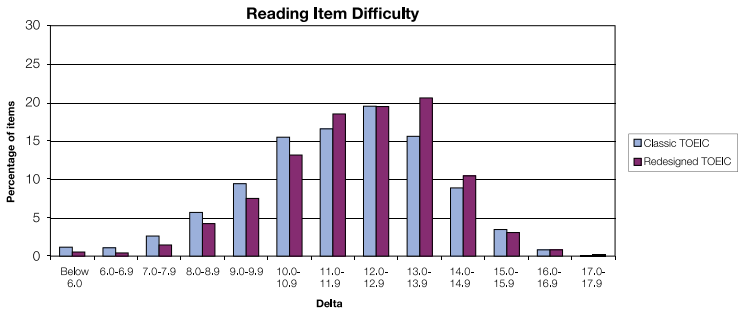


FIGURE 2

Reading Item Difficulty for the Redesigned and Classic TOEIC Listening and Reading Test



Item Discrimination

The item discrimination for the Classic and Redesigned TOEIC Listening and Reading test is indicated by a statistical index known as the R-biserial correlation coefficient. Like the item difficulty index delta, the R-biserial is an item-level statistic that indicates the correlation between test-takers' scores on a particular item (e.g., 0s or 1s on an item) and the corresponding total scores (e.g., total scores for a

section). The R-biserial correlation shows how well an item differentiates between high and low ability test takers. Table 5 shows the means and standard deviations of R-biserial correlation coefficients across the Classic and Redesigned TOEIC Listening and Reading test. As seen in Table 5, the average R-biserials for both listening and reading items were slightly higher for the redesigned test. These average values, ranging from 0.44 to 0.46, are considered to be very high when compared to other tests of similar length and content. Figure 4 shows the R-biserial distributions for listening items on both the Redesigned and Classic Listening and Reading tests and Figure 5 shows the reading items. The two corresponding distributions reasonably resemble each other.

TABLE 5

Average Item Discrimination for the Redesigned and Classic TOEIC Listening and Reading Test

		Classic TOEIC	Redesigned TOEIC
Mean	**Listening**	0.44	0.45
SD		0.13	0.13
Mean	**Reading**	0.44	0.46
SD		.13	.13

FIGURE 3

Listening item discrimination for the Redesigned and Classic TOEIC Listening and Reading Test

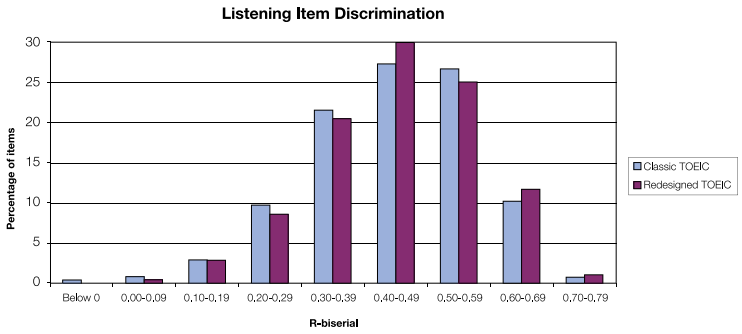
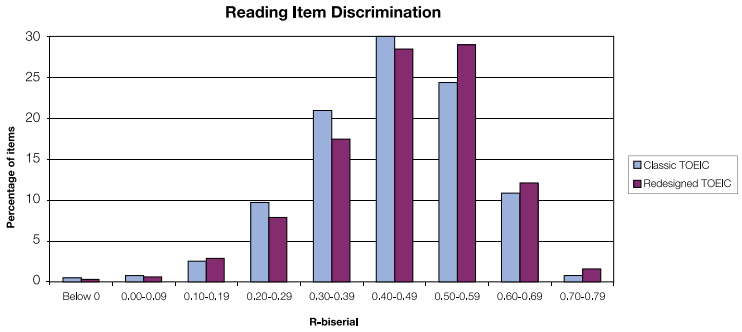


FIGURE 4

Reading Item Discrimination for the Redesigned and Classic TOEIC Listening and Reading Test



Test-Taker Performance

Table 6 shows the scaled score means and standard deviations for listening scores on the Classic and Redesigned TOEIC Listening and Reading tests for test takers in Japan and Korea. Very little variation in the means was observed in the scores for the redesigned and classic listening sections on both tests in each country. Test-taker performance on the listening sections of both tests aligned very closely, with an effect size of only 0.01. Figure 5 shows the cumulative percentile ranks of the scaled scores for test takers in Japan and Korea. The close proximity of the two curves indicates that the two test-taker groups performed very similarly.

TABLE 6

Listening Scaled Score Means and Standard Deviations for the Redesigned and Classic TOEIC Listening and Reading Test

	Classic TOEIC		Redesigned TOEIC	
	Means	SD	Means	SD
Japan	313	85	311	86
Korea	320	83	322	82
Japan & Korea	318	83	319	84

FIGURE 5

Listening Scaled Scores for Test-takers in Japan and Korea on the Redesigned and Classic TOEIC Listening and Reading Test

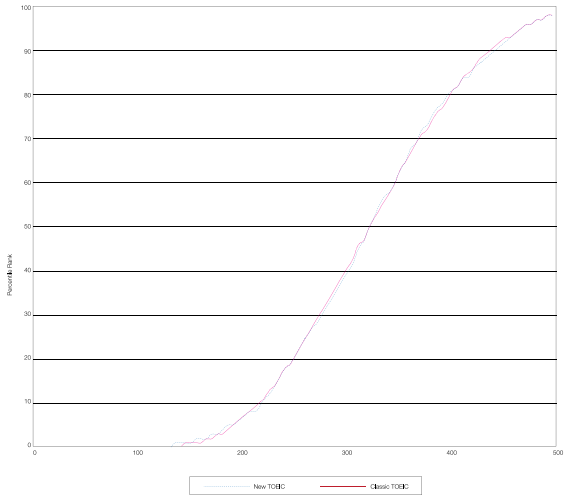


Table 7 shows the scaled score means and standard deviations for reading scores on the Classic and Redesigned TOEIC Listening and Reading tests for test takers in Japan and Korea. In general, the redesigned TOEIC test-taker group performed slightly better than the classic TOEIC test-taker group on the reading section. The effect size is approximately 0.06, which is considered to be an insignificant difference. Figure 6 displays the cumulative percentile ranks of the reading scaled scores for the two studied groups. The redesigned TOEIC group was 4-scaled score points higher on average than the classic group in Japan, and 7-scaled score points higher on average in Korea. The Redesigned and Classic tests use the same scoring scale, and the scores from the different forms were equated. Both Table 6 and Figure 6 show that the groups taking the redesigned TOEIC tests are slightly more able in reading, and the pattern is consistent in Japan and Korea.

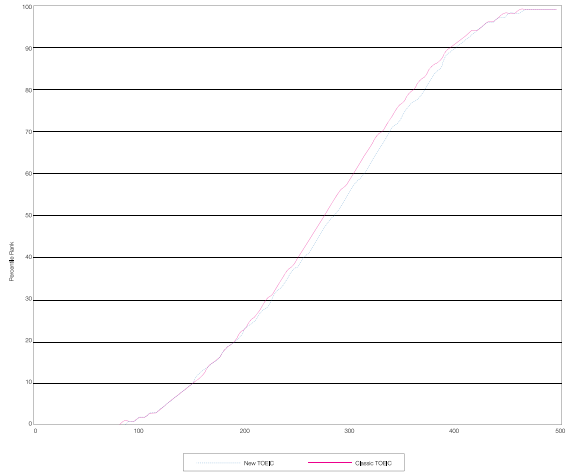
TABLE 6

Reading Scaled Score Means and Standard Deviations for the Redesigned and Classic TOEIC Listening and Reading Test

	Classic TOEIC		Redesigned TOEIC	
	Means	**SD**	**Means**	**SD**
Japan	255	91	259	94
Korea	275	91	282	93
Japan & Korea	270	92	276	94

FIGURE 6

Reading Scaled Scores for Test-takers in Japan and Korea on the Redesigned and Classic TOEIC Listening and Reading Test



Summary

In summary, the preliminary results at both the item and test level, along with the results for test-taker performance, reveal the close similarity in statistical characteristics between the Redesigned TOEIC Listening and Reading test and the Classic TOEIC Listening and Reading test. This conclusion is supported by how closely the scaled score distributions of the redesigned and classic tests correspond to each other across separate groups in the study. On the basis of this study, one could claim that test scores on the Redesigned and Classic TOEIC tests are comparable, even though the two versions contain slightly different item type structures.