ETS TOEIC®

Know English. Know Success.

COMPENDIUM STUDY

Field Study Results for the Redesigned TOEIC® Listening and Reading Test

Chi-wen Liao

January 2010

In January 2003, a team of content and statistical analysis specialists was formed to consider a redesign of the TOEIC® Listening and Reading test as it then existed, as noted in Compendium Study 2.1 of this compendium. There were four stages to the project: construct identification, prototyping, pilot testing, and field testing. The purpose of the field study, carried out in November 2004, was to evaluate the psychometric quality of the proposed TOEIC Listening and Reading redesign. This report documents the statistical results of the field study. Background information of the study is provided in this paper, including a description of the field test specifications, the data collection plan, and the approach to establish population and field study group equivalence. This is followed by a discussion of various statistical properties of the Redesigned TOEIC Listening and Reading test compared to the Classic TOEIC Listening and Reading test, such as item difficulty and discrimination, test difficulty and reliability, as well as the challenge to maintain the classic TOEIC scale.

Background

Specifications of the Field Test Forms

Two test forms were authored and assembled by the TOEIC redesign team. For field study testing purposes, the two forms were known as English Research Study Form C and English Research Study Form D. As in the Classic TOEIC Listening and Reading test, the Redesigned TOEIC Listening and Reading test was made up of two subtests or sections, listening and reading, with 100 multiple-choice items in each section. However, the numbers of items for some item types were changed, some new item types were added, and the number of questions that were presented as part of a set was increased. See Compendium Study 2.1 in this compendium for details on test content. Table 1 reiterates the test specifications of the Redesigned TOEIC Listening and Reading test along with the Classic TOEIC Listening and Reading test specifications for comparison.

TABLE 1

Test Specifications of The Classic and The Redesigned TOEIC Test

	Classic TOEIC Listening and Reading test	Redesigned TOEIC Listening and Reading tests
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

As mentioned, the Redesigned TOEIC Listening and Reading test consists of 100 items in each section, the same as the Classic TOEIC Listening and Reading test. The test is scored according to the number of items answered correctly. There is no penalty on guessing. The TOEIC scale for each section is 5 to 495 in increments of 5.

Data Collection Plan

A total of 1,958 candidates from Japan (N = 1,356) and Korea (N = 602) participated in this study in November 2004. In order to make various statistical comparisons of the Redesigned test TOEIC Listening and Reading with the Classic TOEIC Listening and Reading test, each candidate was required to take the operational TOEIC test (form A9) and one of the field test forms.

In order to compare the difficulty of forms C and D, equivalent groups were expected to take the two test forms. Instructions were thus given to the field to randomly assign candidates to take either form C or D. In Korea, candidates took form A9 as an operational administration in May 2004. Some of those candidates were then invited back in November to take either form C (N = 312) or D (N = 290). In Japan, form A9 and either form C or D were administered together in November 2004. A counter-balanced design was used when administering form A9 along with forms C (N = 696) or D (N = 660). A total of 1,008 and 950 candidates from the combined Japan and Korea group took forms C and D, respectively.

Samples and the TOEIC Population

In order to evaluate whether the difficulty of the Redesigned TOEIC Listening and Reading test was appropriate for the TOEIC population, it was important to check whether the ability of the field study sample was comparable to that of the TOEIC population. Also, the equivalence of the two sample groups was checked on for forms C and D, as the two forms could only be compared on various test qualities/properties if equivalent groups took them.

The abilities of the two groups were compared based on their performance on form A9. The raw score distributions of form A9 for the C and D groups are presented in Figures 1 and 2 for listening and reading, respectively. By examining the cumulative frequency curves in Figures 1 and 2, it can be seen that while the ability of C and D groups were quite close to each other, the D group was slightly more able than the C group on both listening and reading. This is also evident in the form A9 means for each group. The descriptive statistics for listening and reading form A9 by country and group are presented in Tables 2 and 3, respectively. The listening means (standard deviations) were 69.0 (13.5) and 70.1 (13.1) for groups C and D, respectively. The reading means (standard deviations) were 57.6 (15.3) and 58.4 (15.2) for groups C and D, respectively.

Figure 1. Form A9 – Listening Cumulative Frequency Curve.

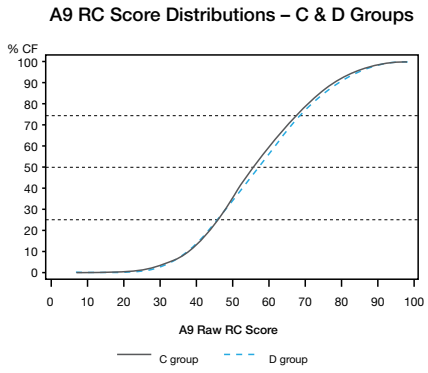


Figure 2. Form A9 – Reading Cumulative Frequency Curves.

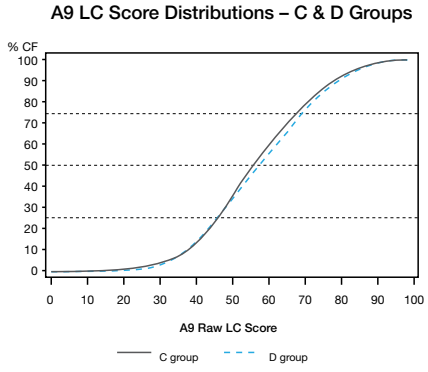


TABLE 2

Listening – Form A9 Descriptive Statistics by Country and Form

Country	Korea			Japan			Japan & Korea	Japan & Korea
Form	C & D	C	D	C & D	C	D	C	D
N	602	312	290	1,356	696	660	1,008	950
Mean	69.4	68.2	70.7	69.6	69.4	69.9	69.0	70.1
SD	12.2	12.1	12.2	13.8	14.1	13.5	13.5	13.1
Min	35	35	41	20	20	25	20	25
Max	97	97	96	99	99	98	99	98

TABLE 3

Reading – Form A9 Descriptive Statistics by Country and Form

Country	Korea			Japan			Japan & Korea	Japan & Korea
Form	C & D	C	D	C & D		C & D	C	D
N	602	312	290	1,356	696	660	1,008	950
Mean	61.4	60.5	62.5	56.5	56.4	56.6	57.6	58.4
SD	13.8	14.1	13.5	15.7	15.7	15.6	15.3	15.2
Min	29	29	34	7	7	16	7	16
Max	94	93	94	99	97	99	97	99

An analysis was then done to evaluate whether the field study samples who took form A9 were representative of the TOEIC population. There were more than 6 million examinees who took the TOEIC Listening and Reading test from 2001 to 2004. Their scaled scores were grouped into five ranges that represented various percentages of the examinees in the population. The percentages of the examinees and their scaled score means within each range for the population and the field study sample (on form A9) are presented in Tables 4 and 5 for listening, and Tables 6 and 7 for reading.

For listening, it was found that the field study sample was more able than the population. The listening scaled score means of groups C and D on form A9 were 338 and 345, respectively, compared to the population mean of 314. In particular, the lower end of the score distribution had a lower percentage of candidates than the population, and the higher end had a higher percentage of candidates than the population.

For reading, it was found that the field study sample was about equally able as the population. The percentages of examinees across the scaled score ranges were, in general, consistent between the field study sample and the population. The reading scaled score means of groups C and D on form A9 were 268 and 273, respectively, compared to the population mean of 267.

TABLE 4

Listening – Means and Standard Deviations of the TOEIC Population vs. Field Study Sample (Group C)

	Population (N = 6,121,672)			Form A9 (group C)			
Score range	**Pop %**	**M**	**SD**	**N**	**%**	**M**	**SD**
5-180	5	155	24	36	4	158	23
185-250	20	222	20	124	12	226	20
255-310	25	283	17	241	24	287	16
315-375	25	344	18	271	27	347	18
380-455	20	412	22	267	26	415	22
460-495	5	479	13	69	7	478	13
5-495	100	314	84	1008	100	338	81

TABLE 5

Listening – Means and Standard Deviations of the TOEIC Population vs. Field Study Sample (Group D)

	Population (N = 6,121,672)			Form A9 (group D)			
Score range	Pop%	M	SD	N	%	M	SD
5-180	5	155	24	16	2	161	25
185-250	20	222	20	122	13	224	21
255-310	25	283	17	200	21	285	16
315-375	25	344	18	259	27	346	18
380-455	20	412	22	282	30	416	22
460-495	5	479	13	71	7	476	13
5-495	100	314	84	950	100	345	80

TABLE 6

Reading – Means and Standard Deviations of the TOEIC Population vs. Field Study Sample (Group C)

	Population (N = 6,121,672)			Form A9 (group C)			
Score range	Pop %	M	SD	N	%	M	SD
5-115	5	92	21	44	4	94	20
120-195	20	162	22	177	18	162	24
200-265	25	233	20	291	29	231	20
270-335	25	302	20	263	26	300	20
340-415	20	373	22	174	17	375	21
420-495	5	411	17	59	6	439	16
5-495	100	267	92	1008	100	268	90

TABLE 7

Reading – Means and Standard Deviations of the TOEIC Population vs. Field Study Sample (Group D)

	Population (N = 6,121,672)			Form A9 (group D)			
Score range	Pop %	M	SD	N	%	M	SD
5-115	5	92	21	42	4	98	15
120-195	20	162	22	169	18	164	21
200-265	25	233	20	227	24	231	19
270-335	25	302	20	277	29	300	19
340-415	20	373	22	181	19	376	22
420-495	5	411	17	54	6	437	17
5-495	100	267	92	950	100	273	90

In summary, groups C and D were approximately equivalent in ability. However, the field study sample was more able than the TOEIC population in listening, and about equally able as the TOEIC population in reading.

Results

Item Difficulty

Item difficulty is characterized by two types of statistical indices, one is p-value and the other is delta (Δ). The p-value ranges from 0 to 1 and is defined as the proportion of people who answer the item correctly. The closer the number is to 1, the easier the item, and vice versa. Tables 8-9 show the mean and standard deviation (SD) of p-values for listening and reading items, respectively. For listening, it was found that the mean p-value for form A9 was .70, and the mean p-values for forms C and D were only .57 and .63, respectively (see Table 8). These indicated that the items on the field test forms, on average, were more difficult than the items on the operational TOEIC Listening and Reading test form, form A9. For reading, it was found that the mean p-value for form A9 was .58, and the mean p-values for forms C and D were .58 and .60, respectively (see Tables 10). These indicated that, on average, the difficulty of the three test forms were equivalent.

TABLE 8

Listening Mean and Standard Deviation of P-Values

	Form A9	Form C	Form D
Mean	0.70	0.57	0.63
SD	0.17	0.20	0.20

TABLE 9

Reading Mean and Standard Deviation of P-Values

	Form A9	Form C	Form D
Mean	0.58	0.58	0.60
SD	0.18	0.18	0.19

Similar evidence was observed from delta. 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). Tables 10-11 show the mean and standard deviation (SD) of delta values for listening and reading, respectively. For listening, it was found that the mean delta value for form A9 was 10.6, and the mean p-values for forms C and D were 12.2 and 11.4, respectively (see Table 10). This indicates that items on form C were, on average, the most difficult, and items on form A9 were the easiest. For reading, it was found that the mean p-value for form A9 was .58, and the mean p-values for forms C and D were .58 and .60, respectively (see Tables 10). Again, the closeness of these values indicates that the items on the listening sections of the three forms, on average, were of equal difficulty.

TABLE 10

Listening Mean and Standard Deviation of Delta Values

	Form A9	Form C	Form D
Mean	10.6	12.2	11.4
SD	2.30	2.40	2.4

TABLE 11

Reading Mean and Standard Deviation of Delta Values

	Form A9	Form C	Form D
Mean	12.0	12.1	11.9
SD	2.10	2.0	2.2

Item Discrimination

Item discrimination 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. Tables 12 and 13 show the mean and standard deviation (SD) of R-biserial values for listening and reading, respectively. For both listening and reading, it was found that the mean R-biserial values for the three forms were similar to one another, and the mean values ranged from .42 to .43. The closeness of these values indicates that the items on the three forms were, on average, equally discriminating.

TABLE 12

Listening Mean and Standard Deviation of R-Biserial

	Form A9	Form C	Form D
Mean	0.43	0.42	0.44
SD	0.13	0.15	0.13

TABLE 13

Reading Mean and Standard Deviation of R-Biserial

	Form A9	Form C	Form D
Mean	0.43	0.43	0.42
SD	0.12	0.13	0.14

Test Difficulty

The total test difficulty of form A9 compared with the total test difficulties of field test forms C and D were analyzed, and the means and standard deviations (SD) of the three forms for listening and reading are shown in Table 14. Consistent with the finding on mean item difficulty, it was found that the listening form C was the most difficult with a mean (M= 56.5) much lower than form A9 (M = 69.0). The reading mean of the three forms are much more comparable to one another.

TABLE 14

Mean and Standard Deviation of Forms A9, C, and D

	Group C (N = 1,008)		Group D (N = 950)	
	Form A9	Form C	Form A9	Form D
Listening				
Mean	69.0	56.5	70.1	62.9
SD	13.5	14.0	13.1	14.1
Reading				
Mean	57.6	56.6	58.4	59.0
SD	15.3	15.5	15.2	14.5

The total score distributions for the three forms were compared. The cumulative frequency curves comparing form A9 with form C and with form D are presented in Figures 3 and 4 for listening, and 5 and 6 for reading. Consistent with the information about mean item difficulty, it was found that forms C and D were both more difficult than form A9 in listening. For example, half of the sample (at about 50th percentile) correctly answered 57% of form C and 62% of form D, but correctly answered about 70% of form A9.

FIGURE 3

Form A9 and Form C – Listening Cumulative Frequency Curves

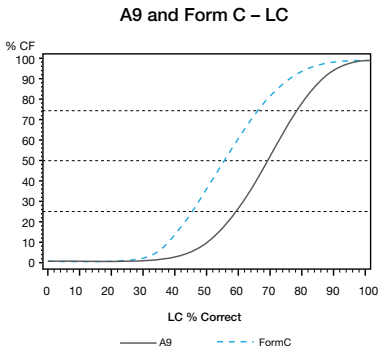
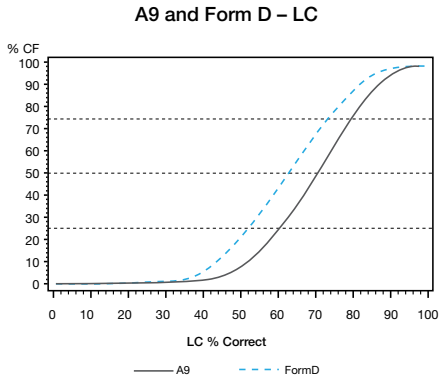


FIGURE 4

Form A9 and Form D – Listening Cumulative Frequency Curves



For reading, the cumulative frequency curves comparing form A9 with forms C and D are presented in Figures 5 and 6, respectively. It was found that forms C and D and form A9 were of similar difficulty. The two curves for form A9 and form C were almost identical. The form D curve only slightly departed from the form A9 curve.

FIGURE 5

Form A9 and Form C – Reading Cumulative Frequency Curves

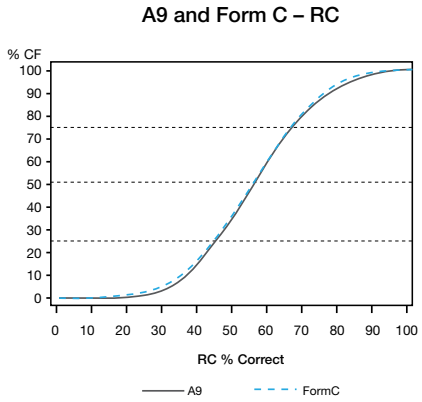
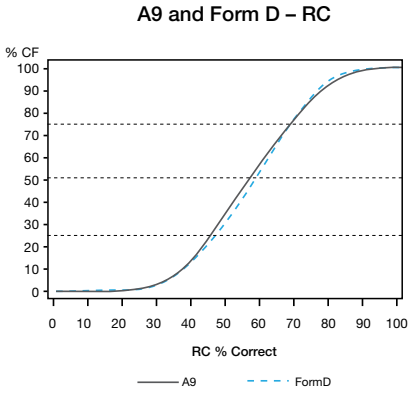


FIGURE 6

Form A9 and Form D – Reading Cumulative Frequency Curves



Since the field study samples were more able than the TOEIC population in listening, the test design team wanted to know how the TOEIC population would score if they had taken the field tests. The following question was asked: How would the TOEIC population have scored if they had taken forms C and D?

To answer this question, the performance of the TOEIC population on the redesigned test needed to be estimated, and it was done through the following steps. First, forms C and D were equated to form A9 and placed on the TOEIC scale. This was done through equating form C to form A9, and form D to form A9, through a single group design. The scaled score distributions of forms C and D were then obtained and compared to the scaled score distribution of the TOEIC population from 2001 to 2004. For example, a person at the 50th percentile in the TOEIC population had a listening scaled score of about 310. By checking the raw-to-scaled conversion for form C, it was found that a raw score of 51 was equivalent to a scaled score of 310. It could be concluded that a person in the population with a listening scaled score of 310 could score around 51 items right on the field test form C.

Tables 15 and 16 present the estimated listening and reading score if the TOEIC population had taken Forms C, D, and A9. The population was separated into six subgroups with the cuts around the 5th, 25th, 50th, 75th, and 95th percentiles. For listening, it was found that the 50th percentile of the TOEIC population would answer about 51, 57, and 65 items correctly in forms C, D, and A9. It was also observed that the 95th percentile of the TOEIC population would answer about 78, 83, and 88 items correct for forms C, D, and A9. The information was provided to the content redesign team to help the team decide how they would adjust the difficulty level of the operational redesigned listening test forms.

TABLE 15

Listening – Estimated Raw Scores for Forms C, D, and A9

Pop %	Population	Form C	Form D	Form A9
5%	460 – 495	79 – 96	84 – 99	89 – 100
20%	380 – 455	63 – 78	69 – 83	76 – 88
25%	315 – 375	52 – 62	58 – 68	66 – 75
25%	255 – 310	42 – 51	48 – 57	56 – 65
20%	185 – 250	33 – 42	36 – 47	43 – 55
5%	5 – 180	0 – 32	0 – 35	0 – 42

For reading, it was found that the 50th percentile of the TOEIC population would answer about 55, 58 and 56 items correctly on forms C, D and A9. It was also observed that the 95th percentile of the TOEIC population would answer about 81, 81 and 83 of the forms C, D and A9 items correctly. These results suggest that forms C and D reading are comparable in difficulty to form A9. Again, the information was provided to the content redesign team to help the team decide if they wanted to adjust the difficulty level of the operational redesigned reading test forms.

TABLE 16

Reading – Estimated Raw Scores for Forms C, D and A9

Pop %	Population	Form C	Form D	Form A9
5%	420 – 495	82 – 100	82 – 98	84 – 100
20%	340 – 415	69 – 81	70 – 81	70 – 83
25%	270 – 335	56 – 68	59 – 69	57 – 69
25%	200 – 265	45 – 55	47 – 58	46 – 56
20%	120 – 195	32 – 44	34 – 46	34 – 45
5%	5 – 115	0 – 31	0 – 33	0 – 33

In summary, it was found that listening forms C and D were more difficult than form A9 for the field study sample and that the field study sample was more able than the TOEIC population in listening, which suggests that the TOEIC population would find the Redesigned TOEIC Listening test more difficult than the Classic TOEIC Listening test. For reading, forms C and D were about as difficult as form A9 for the field study sample and the sample was about equally able as the TOEIC population,

which suggests that the TOEIC population would find the Redesigned TOEIC Reading test as difficult as the Classic TOEIC Reading test.

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 not be directly related to the purpose of the test (e.g., the time of administration, test-taking conditions, or the scorers employed). 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 or each content type.

Table 17 shows the reliability estimates for the listening and reading section scores and content scores from forms A9, C and D. The reliability of listening and reading section scores of forms C and D was comparable to those of form A9. As to the content areas, there was a drop in the reliability of the photo section in the field test forms. This was because the number of items was reduced to 10 from 20. The talk items were increased from 20 to 30, and there was an increase of reliability in the talk section for the field test form. The traditional passage had only 28 items in the field test forms, but the reliability was quite high for both forms, .78.

TABLE 17

Reliability of Forms A9, C and D

	Classic test form		Field test forms		
	# of items	Form A9	# items	Form C	Form D
Listening	100	0.91	100	.90	.91
Photo	20	0.67	10	.52	.49
Question-response	30	0.78	30	.80	.78
Conversation	30	0.78	30	.68	.76
Talk	20	0.61	30	.75	.77
Reading	100	0.92	100	.92	.91
IS	40	0.84	40	.82	.82
ER	20	0.69	N/A	N/A	N/A
Cloze	NA	N/A	12	.62	.57
Traditional passage	40	.84	28	.78	.78

	Classic test form		Field test forms		
	# of items	Form A9	# items	Form C	Form D
Double passage	NA	N/A	20	.75	.69

Correlations

With all of the changes in the redesign, it was important to examine how the Redesigned TOEIC test Listening and Reading scores (forms C and D) were correlated with the Classic TOEIC Listening and Reading test scores (form A9). Tables 18 and 19 show the correlations between section scores for forms A9, C and D. High correlations (.87 and .88) were found between the redesigned and classic TOEIC test scores for listening and reading. These correlations are nearly as high as their reliability, suggesting that the redesigned test measures the same constructs as the classic test.

TABLE 18

Correlations Between Listening and Reading for Forms A9 and C

	Form A9 Listening	Form A9 Reading	Form C Listening	Form C Reading
Form A9 – Listening	1.00			
Form A9 – Reading	0.76	1.00		
Form C – Listening	0.88	0.74	1.00	
Form C – Reading	0.73	0.87	0.76	1.00

TABLE 19

Correlations Between Listening and Reading for Forms A9 and D

	Form A9 Listening	Form A9 Reading	Form C Listening	Form C Reading
Form A9 – Listening	1.00			
Form A9 – Reading	0.73	1.00		
Form D – Listening	0.87	0.74	1.00	
Form D- Reading	0.71	0.88	0.77	1.00

Speededness

The TOEIC test's listening is paced by a tape recording; therefore, the issue of speededness does not apply. In contrast, the TOEIC reading section is intended to be a power test so that score differences reflect individual differences in ability and not individual differences in working speed.

Criteria frequently used in judging the speededness of a test include: (a) percentage of examinees completing the whole section, (b) percentage of examinees completing 75% of the section, and (c) number of items reached by 80% of the examinees. Each of these is arbitrary and should not be strictly applied in isolation. As a rule of thumb at ETS, a test is usually regarded as essentially unspeeeded if at least 80% of the test takers reach the last question and if virtually everyone reaches 75% of the items.

It was found that at least 80% of the examinees reached the last reading question on both Forms C and D, and almost the whole sample (97% and 99%) reached 75% of the items in the reading sections of both forms. For form A9, 80% of the examinees reached the last reading question and 99.5% of the group reached 75% of the reading items. The redesigned and classic TOEIC reading sections have similar speededness indices, and all forms appear to be unspeeded.

Maintaining the TOEIC Scale

During the redesign process, it was hoped that the existing classic TOEIC scale could continue to be used to report performance on the proposed Redesigned TOEIC Listening and Reading test. This issue was examined by checking whether the redesigned test continued to measure the same constructs as the classic test (see the section on Correlations above) and whether the difficulty level of the redesigned test was kept at about the same level as the classic test (see the section on Test Difficulty above). It was found that the listening section of the redesigned test, especially on form C, was made more difficult than the typical Classic TOEIC Listening and Reading test (form A9). A question was asked: How would the raw-to-scale conversion be affected given that the Redesigned TOEIC Listening test was made more difficult than the Classic TOEIC Listening and Reading test (e.g., form A9)?

When forms C and D were equated to form A9, it was found that a raw score of 87 on form C and 92 on form D reached the top TOEIC scaled score of 495. For the classic TOEIC forms, the raw scores that reached the top listening scale typically range between 92 and 95. Form C was more difficult than most of the classic TOEIC forms. This finding was shared with the redesigned content team, and the team decided that the difficulty level of the redesigned operational listening test would need to be adjusted to make it easier if the TOEIC scale was to be continued.

In summary, given the high correlations of the field test scores with form A9 scores, the current TOEIC scale could continue to be used for the Redesigned TOEIC Listening and Reading test if the listening section of the redesigned test is made easier than forms C and D.

Conclusion

The field study demonstrated that the difficulty level of the redesigned reading section in the TOEIC Listening and Reading test would be appropriate for the TOEIC population. However, the difficulty level of one listening section was harder than the listening section in the Classic TOEIC Listening and Reading test, so the TOEIC population would find it difficult. The test scores on the listening and the reading sections for both the Redesigned and Classic TOEIC Listening and Reading tests were correlated as high as .87. It was concluded that the classic TOEIC reporting scale could continue to be used for reporting the redesigned TOEIC scores as long as the difficulty level of the redesigned TOEIC listening section could be reduced so that it was closer to the level of the classic TOEIC listening section. The speededness data suggests the redesigned TOEIC reading section was not speeded. The field study results also suggest that it is possible to construct the redesigned test to maintain the high reliability of the classic test of .90 and above.