Measuring the Power of Learning.®

Research Memorandum
ETS RM–17-05

Statistical Analyses for the Updated *TOEIC*® Listening and Reading Test

Jaime Cid

Youhua Wei

Sooyeon Kim

Claudia Hauck

December 2017

ETS Research Memorandum Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Memorandum series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Memorandum series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

Statistical Analyses for the Updated *TOEIC*® Listening and Reading Test

Jaime Cid, Youhua Wei, Sooyeon Kim, and Claudia Hauck
Educational Testing Service, Princeton, New Jersey

December 2017

Corresponding author: J. Cid, E-mail: jcid@ets.org

Suggested citation: Cid, J., Wei, Y., Kim, S., & Hauck, C. (2017). *Statistical analyses for the updated* TOEIC® *Listening and Reading test* (Research Memorandum No. RM-17-05). Princeton, NJ: Educational Testing Service.

Action Editor: Donald Powers

Reviewers: Rick Morgan and Hongwen Guo

Abstract

To ensure that tests continue to meet the needs of test takers and score users, it is important that testing programs periodically revisit their assessments. For this reason, in order to keep up with the continuously changing use of English and the ways in which individuals commonly communicate in the global workplace and everyday life, an updated *TOEIC*® Listening and Reading test was designed and first launched in May 2016. The update was intended to benefit both test takers and scores users. Not only would test takers be better able to demonstrate the relevant skills needed to communicate effectively in today's global workplace, but also score users would continue to be confident that test scores reflect the range of skills necessary for success in their workplaces.

Although the update to the test included some new item types, the overall quality and difficulty of the TOEIC Listening and Reading test was intended to remain the same, with no change in total testing time, number of items, test difficulty or score scale. This paper reports the results of a pilot study that contributed to ensuring that the TOEIC Listening and Reading test continues to be a fair, valid and reliable assessment of everyday and workplace English. The results of the pilot study also helped test developers make the appropriate adjustments to the test before it was launched operationally.

Since the updated test was launched, the difficulty and discrimination of the items of the updated test, reliability of its scores and scaled score values have also been closely monitored. The operational results presented in this report also suggest that that the updated TOEIC Listening and Reading test continues to have the same psychometric quality of the pre-updated TOEIC Listening and Reading test.

Key words: TOEIC, item analysis, listening, reading, statistical report

The *TOEIC*® family of products and service is designed to measure the English language proficiency of nonnative speakers of English engaged in the global workplace, where English is the language of communication. The TOEIC Listening and Reading test consists of two separately timed sections, Listening Comprehension and Reading Comprehension, with 100 items in each section. The Listening section is paced by audiotape recording.

In May 2016, the TOEIC family of products and service announced updates to the TOEIC Listening and Reading test to keep up with the changing use of English and the ways in which individuals commonly communicate in the global workplace and everyday life. New item types were included, but there was no change in total testing time, number of items, test difficulty, or score scale. The updated test included communication formats, such as text messaging and instant messaging, which are in current use. It also placed greater emphasis on connecting information across multiple sources, such as what is seen in a visual image and what is heard in a related conversation (pragmatics). A pilot study conducted in May 2015 evaluated the statistical properties of the updated TOEIC Listening and Reading test. The purpose of this report is to document the results of such statistical analyses.

Background

Table 1 presents the composition of the Listening and Reading sections, in the preupdated and updated (new) specifications. The changes in the Listening section require a greater emphasis on Short Conversations but less emphasis on Photographs and Question–Response. The changes on the Reading section require a greater emphasis on Reading Comprehension and Text Completion but less emphasis on Incomplete Sentences. Approximately one-quarter of items in each of the Listening and Reading sections are new-type items. These item types include to some extent the new features aforementioned (e.g., new communication formats in the Reading section, such as text messages, instant messages, and online chat conversations with multiple writers). The preupdated score reports included scale scores for both the Listening and Reading sections and the percentage of questions answered correctly for each of four ability claims in the Listening section and each of five ability claims in the Reading section. The updated Listening section includes an additional ability claim (Ability 5, pragmatic understanding). The reporting scale for each section of the updated test remains the same as for the preupdated test, with a score scale ranging from 5 to 495 in increments of 5.

Table 1 Composition of Each Section Under the Updated Specification and the Preupdated Specification of the TOEIC Test

Section: Part	Updated test	Preupdated test
Listening: Part 1. Photographs	6	10
Listening: Part 2. Question–Response	25	30
Listening: Part 3. Short Conversations	39	30
Listening: Part 4. Short Talks	30	30
Reading: Part 5. Incomplete Sentences	30	40
Reading: Part 6. Text Completion	16	12
Reading: Part 7. Reading Comprehension	54	48
Reading: Part 7A. Single Passages	29	28
Reading: Part 7B. Multiple Passages	25	20

Note. Total number of items in each section was 100.

Pilot Form Design

Two parallel TOEIC Listening and Reading test pilot forms (Forms E and F) were assembled based on the updated specifications (see Table 1). Forms E and F were designed to be parallel from statistical and content perspectives. The pilot forms were randomly distributed to the test takers to make the two pilot form groups comparable in ability. To establish a strong score connection between the reference and the pilot forms, 50 Listening items and 45 Reading items were used as common items in both Form E and Form F. The common item sets were designed to be miniature versions of the reference form in terms of the content and statistical specifications.

As mentioned earlier, for both sections of the updated test, five ability claims are reported in the score report using the percentage correct score. Tables 2 and 3 present the number of items associated with each of the five abilities measured in each section. Although some of the abilities had fewer than 15 items in the pilot forms, the minimum number of items included currently in operational forms for each ability claim is 15.

Table 2 Number of Items for Each Ability Claim in the Listening Section

Ability	Form E	Form F	Reference
1. Can infer gist, purpose, and basic context based on information that is explicitly stated in short spoken texts	16	15	19
2. Can infer gist, purpose, and basic context based on information that is explicitly stated in extended spoken texts	19	16	17
3. Can understand details in short spoken texts	15	16	21
4. Can understand details in extended spoken texts	50	53	43
5. Can understand a speaker's purpose or implied meaning in a phrase or sentence (pragmatic understanding)	11	13	-

Note. Because some items measure more than one ability in the Listening section, the total number of items in each form will not be equal to 100.

Table 3 Number of Items for Each Ability Claim in the Reading Section

Ability	Form E	Form F	Reference
1. Can locate and understand specific information in tables and passages	18	20	16
2. Can connect information across multiple sentences in a single text and across texts	13	11	16
3. Can make inferences based on information in written texts	35	35	25
4. Can understand vocabulary in workplace texts	28	26	29
5. Can understand grammar in workplace texts	20	20	27

Note. Because some items measure more than one ability in the Reading section, the total number of items in each form will not be equal to 100.

Data Collection

A total of 3,673 test takers from Japan (n = 2,045) and Korea (n = 1,628) participated in the pilot study in May 2015. To evaluate the representativeness of the pilot samples, standardized mean differences[1] (SMD) were calculated based on the total score of each group. As shown in Table 4, in the reference form group, who were administered the May 2014 operational form, Korean test takers were much more able than Japanese test takers in both sections, and their ability difference was much larger in Listening (SMD = .53) than in Reading (SMD = .22). In the pilot study, a different trend emerged. The Japanese pilot form groups were more able than the Korean groups, and their ability differences were larger on the Reading section than on the Listening section. Therefore the pilot samples were not completely representative of the TOEIC population. A possible reason is that in the pilot samples, the percentage of repeaters was larger than the percentage observed in operational practice in Japan than in Korea. However, the operational trend of the Korean group performing comparatively better on the Listening section than on the Reading section was present in the pilot study (.35 and .21 better for Form E and Form F, respectively). The descriptive statistics of raw scores for Listening and Reading sections by country and form are presented in Table 5.

Table 4 Standardized Mean Differences of Groups and of Forms Based on the Total Test Score of Each Group

Difference	Listening	Reading
Form E (Korea—Japan)	−.05	−.40
Form F (Korea—Japan)	−.12	−.33
Reference (Korea—Japan)	.53	.22

Table 5 Descriptive Statistics of Raw Scores by Country and Form

Statistic	Form E: Japan	Form E: Korea	Form E: Combined	Form F: Japan	Form F: Korea	Form F: Combined	Reference: Japan	Reference: Korea	Reference: Combined
Sample size	1,019	824	1,843	1,026	804	1,830	48,745	38,500	87,245
Listening mean	65.11	63.63	64.45	66.34	63.90	65.27	66.72	73.81	69.85
Listening *SD*	15.35	18.14	16.67	15.74	18.51	17.05	15.95	16.31	16.49
Reading mean	55.96	50.99	53.74	60.31	54.97	57.96	57.05	62.42	59.42
Reading *SD*	15.1	16.98	16.16	15.87	19.02	17.52	16.95	18.14	17.69

Note. SD = standard deviation.

Statistical Analyses and Results

Equating

The comparability of the pilot and operational testing samples was further evaluated by examining the performance of the pilot and reference groups on the common items. Then, for each pilot form, equating—under the nonequivalent groups with anchor test design—was conducted through the common items shared in both the pilot (updated) forms and operational reference (preupdated) form. Equating is used to adjust the difficulty level of a form and derive the scaled scores from test takers' raw scores in order that the reported scaled scores obtained from different test forms are comparable, regardless of any potential differences in form difficulty. The number of common items was 50 in Listening and 45 in Reading. The equating relationship between the new forms and the operational reference form was based on the combined group of Japanese and Korean test takers. Table 6 presents the descriptive statistics of the anchor scores in Forms E and F (combined group) and the operational reference form. As indicated by the negative SMD between the new and operational reference groups in Table 6, the operational reference group was somewhat more able than the combined pilot groups in both sections. Likewise, the Form F group was somewhat more able than the Form E group.

Table 6 Summary of Anchor Statistics and Group Differences

Statistic	Form E	Form F	Reference form
Sample size	1,843	1,830	87,245
Listening number of anchor items	50	50	50
Listening mean	34.76	35.42	35.54
Listening *SD*	8.60	8.59	8.25
Listening standardized difference [a]	−0.09	−0.02	
Reading number of anchor items	45	45	45
Reading mean	25.71	26.64	26.97
Reading *SD*	8.01	8.19	7.82
Reading standardized difference [a]	−0.16	−0.04	

Note. SD = standard deviation.
[a] Denotes standardized mean difference between the pilot form (E or F) and reference form.

Table 7 provides the summary statistics (mean and standard deviation) of the scaled scores for each group taking each form. Recall that the Japanese pilot groups were more able than the Korean pilot groups. As expected, after adjusting the test form difficulty, the scaled score means of the Japanese pilot form groups were higher than the scaled score means of the Korean pilot form groups. Likewise, the scaled score mean of the combined pilot group was somewhat lower (Japan and Korea) than for the reference group. Therefore the group differences based on reported scores were consistent with the group differences based on anchor raw scores.

Table 7 Summary Statistics of Test Takers' Scale Scores

Statistic	Form E: Japan	Form E: Korea	Form E: Combined	Form F: Japan	Form F: Korea	Form F: Combined	Reference: Japan	Reference: Korea	Reference: Combined
Sample size	1,019	824	1,843	1,026	804	1,830	48,745	38,500	87,245
Listening mean	329.28	320.23	325.23	338.96	324.27	332.50	316.1	354.40	333.00
Listening *SD*	84.13	100.58	91.86	83.98	100.98	92.01	85.21	88.73	88.84
Reading mean	277.09	246.64	263.50	288.19	257.73	274.87	264.86	294.83	278.09
Reading *SD*	93.38	103.70	99.24	93.39	109.01	101.6	94.52	100.48	98.33

Note. SD = standard deviation.

Item Difficulty

The difficulty of the items was evaluated by examining two types of statistical indices: *p*-value (defined as the proportion of test takers who answer an item correctly in a given population) and delta (defined as $13 + 4z$, where z is the normal deviate corresponding to proportion correct). *P*-values range from 0 to 1, with a higher value indicating that a greater

proportion of test takers responded to the item correctly, and it was thus an easier item. Delta values typically range from 6 for a very easy item to 20 for a very difficult item, with a mean of 13 (50% correct).

Table 8 presents the *p*-values and equated deltas[2] in each section of the pilot forms and operational reference form. In Listening, the mean *p*-value for the operational reference form was .70, and the mean *p*-values for Forms E and F were .64 and .65, respectively. In Reading, the mean *p*-value for the operational reference form was .60, and the mean *p*-values for Forms E and F were .55 and .59, respectively.

The equated deltas provide us with a difficulty metric that accounts for the different ability levels among the two pilot test groups and the operational test group. The Listening sections for the pilot forms were slightly more difficult than the operational reference form. In Reading, the overall difficulty of the pilot forms was more comparable to the overall difficulty of the operational reference form. This finding is not unexpected given that test takers were not as familiar with the new item types in the pilot forms as they were with the items of the operational reference form.

Table 8 Summary of Item Statistics for Each Section Based on Combined Group

Statistic	*p*-value: Form E	*p*-value: Form F	*p*-value: Reference	ED: Form F	ED: Form F	ED: Reference	*R*-biserial: Form F	*R*-biserial: Form F	*R*-biserial: Reference
Listening mean	0.64	0.65	0.70	13.1	13.2	12.7	0.48	0.50	0.47
Listening *SD*	0.16	0.15	0.13	1.6	1.5	1.3	0.11	0.10	0.11
Listening min	0.26	0.24	0.40	9.8	9.5	9.3	0.20	0.27	0.19
Listening max	0.92	0.94	0.95	16.7	17.0	15.2	0.70	0.74	0.67
Reading mean	0.55	0.59	0.60	12.5	12.3	12.3	0.45	0.49	0.47
Reading *SD*	0.18	0.18	0.16	1.9	1.8	1.7	0.14	0.13	0.11
Reading min	0.19	0.20	0.22	8.3	8.7	8.0	0.09	0.10	0.15
Reading max	0.89	0.89	0.92	16.4	16.4	16.1	0.73	0.72	0.70

Note. ED = equated delta; SD = standard deviation.

Table 9 shows *p*-values and equated delta values for the different parts of the test on the pilot forms and the operational reference form. Overall, in comparison to the operational reference form, in Listening, Short Conversations (Part 3) and Short Talks (Part 4) were more

difficult on the pilot forms than on the operational reference form. The same was observed in Reading for Multiple Passages (Part 7B). However, in general, all forms produced similar difficulty patterns. That is, in Listening, Photographs (Part 1) and Short Talks (Part 4) were, on average, the easiest and hardest parts, respectively. In Reading, as observed on the operational reference form, Incomplete Sentences (Part 5) and Multiple Passages (Part 7B) were, on average, the easiest and most difficult parts, respectively.

Table 9 Means of Item Statistics for Each Part Based on Combined Group

Section: Part	*p*-value: Form E	*p*-value: Form F	*p*-value: Reference	ED: Form E	ED: Form F	ED: Reference	*R*-biserial: Form E	*R*-biserial: Form F	*R*-biserial: Reference
Listening: Part 1	0.80	0.82	0.74	11.4	11.4	11.9	0.39	0.39	0.40
Listening: Part 2	0.67	0.70	0.68	12.9	12.7	12.8	0.45	0.47	0.44
Listening: Part 3	0.66	0.62	0.73	13.0	13.5	12.6	0.52	0.50	0.50
Listening: Part 4	0.57	0.62	0.66	13.8	13.6	13.1	0.47	0.53	0.50
Reading: Part 5	0.67	0.68	0.65	11.2	11.4	11.7	0.52	0.51	0.50
Reading: Part 6	0.55	0.57	0.51	12.5	12.5	13.0	0.42	0.46	0.40
Reading: Part 7	0.48	0.55	0.57	13.2	12.8	12.7	0.42	0.48	0.47
Reading: Part 7A	0.53	0.62	0.61	12.7	11.9	12.2	0.45	0.54	0.48
Reading Part 7B	0.42	0.45	0.51	13.8	13.7	13.4	0.39	0.42	0.45

Note. ED = equated delta; Part 1 = Photographs; Part 2 = Question–Response; Part 3 = Short Conversations; Part 4 = Short Talks; Part 5 = Incomplete Sentences; Part 6 = Text Completion; Part 7 = Reading Comprehension; Part 7A = Single Passages; Part 7B = Multiple Passages.

Item Discrimination

Item discrimination is evaluated by the *R*-biserial correlation coefficient. The *R*-biserial correlation is the relationship between test takers' scores on a particular item (e.g., 0 for an incorrect response or 1 for a correct response) with the corresponding total score (e.g., total score for a section). The *R*-biserial correlation indicates how well an item serves to discriminate between low- and high-ability test takers. Table 8 presents the summary statistics for the *R*-biserial correlations for the pilot and operational reference forms. In general, for both Listening and Reading, the means of *R*-biserial values were comparable between the pilot forms and the operational reference form. Overall, these results indicate that the three forms were, on average, equally discriminating.

Table 9 provides *R*-biserial values for the different parts of the test in Forms E and F and the reference form. Overall, the values suggest that for both Listening and Reading, on average, the items of the different parts of Forms E and F were very close in discrimination to the items of the operational reference form.

Differential Item Functioning

Differential item functioning (DIF) analyses were performed to ensure that all new item types were fair to both men and women. DIF analyses involve the statistical analysis of test items for evidence of differential item difficulty related to subgroup membership. The two groups of interest (e.g., male/female) are matched with respect to ability on a criterion (e.g., total test score) and then compared to see if the item is performing similarly in both groups. The probability that a test taker answers an item correctly should be independent of his or her group membership. The DIF analysis methodology employed (Dorans & Kulick, 1986; Holland & Thayer, 1988) uses statistics that describe the amount of DIF for each item as well as the statistical significance of the DIF effect. The DIF classification followed the ETS system as described by Zwick (2012), in which items are classified into three levels: A (least), B, and C (most). Items identified as C-level DIF should be referred to fairness committees for further evaluation. No item showed C-level DIF. Therefore no item was differentially more difficult for one gender than the other.

Test Parts and Abilities

As mentioned earlier, the Listening section of the updated test includes four parts and provides five ability scores, whereas the Reading section includes three parts and provide fives ability scores. The fifth ability of the Listening section of the updated test is a new ability claim. The correlation between each item score and its ability score measures how well each item is related to its corresponding ability claim. As shown in Table 10, the average item–ability correlations were generally moderate in the Listening and Reading sections. Forms E and F and operational reference form showed similar patterns. The newly added Listening ability claim (Ability 5, pragmatic understanding) yielded item correlations comparable to the ones observed in the other Listening claims.

Table 10 Summary of Item–Ability Correlations Based on Combined Group

Form: Ability	Listening mean	Listening *SD*	Listening min	Listening max	Reading mean	Reading *SD*	Reading min	Reading max
Form E: Ability 1	0.53	0.08	0.35	0.65	0.49	0.12	0.28	0.68
Form E: Ability 2	0.55	0.07	0.31	0.67	0.58	0.08	0.45	0.68
Form E: Ability 3	0.54	0.06	0.42	0.62	0.46	0.14	0.17	0.71
Form E: Ability 4	0.52	0.12	0.23	0.71	0.51	0.13	0.27	0.72
Form E: Ability 5	0.56	0.08	0.44	0.71	0.54	0.09	0.28	0.71
Form F: Ability 1	0.55	0.09	0.35	0.66	0.55	0.13	0.32	0.75
Form F: Ability 2	0.58	0.07	0.50	0.74	0.64	0.07	0.49	0.72
Form F: Ability 3	0.52	0.07	0.40	0.64	0.50	0.14	0.18	0.74
Form F: Ability 4	0.53	0.10	0.34	0.77	0.52	0.11	0.28	0.69
Form F: Ability 5	0.56	0.07	0.42	0.64	0.56	0.10	0.29	0.73
Reference: Ability 1	0.54	0.10	0.25	0.70	0.58	0.10	0.36	0.68
Reference: Ability 2	0.54	0.06	0.40	0.64	0.58	0.09	0.40	0.72
Reference: Ability 3	0.50	0.07	0.38	0.61	0.53	0.12	0.28	0.69
Reference: Ability 4	0.54	0.09	0.25	0.69	0.50	0.13	0.23	0.69
Reference: Ability 5	–	–	–	–	0.53	0.10	0.26	0.73

Note. Ability 5 is the new Listening ability added to the updated TOEIC test. For Listening: Ability 1, can infer gist, purpose, and basic context based on information that is explicitly stated in short spoken texts; Ability 2, can infer gist, purpose, and basic context based on information that is explicitly stated in extended spoken texts; Ability 3, can understand details in short spoken texts; Ability 4, can understand details in extended spoken texts; Ability 5, can understand a speaker's purpose or implied meaning in a phrase or sentence (pragmatic understanding). For Reading: Ability 1, can locate and understand specific information in tables and passages; Ability 2, can connect information across multiple sentences in a single text and across texts; Ability 3, can make inferences based on information in written texts; Ability 4, can understand vocabulary in workplace texts; Ability 5, can understand grammar in workplace texts.

Tables 11–14 present the intercorrelations of the different parts of the test and the abilities in Forms E and F. The lower part below the diagonal presents the correlations from the Form E group, and the upper part above the diagonal presents the correlations of the Form F group. As expected, in Listening (Tables 11–12), Photographs (Part 1), with only six items,

yielded the lowest correlations. The correlations among parts (Parts 1–4 for Listening and Parts 5–7B for Reading) and abilities (Abilities 1–5 for each section) were moderate to high. The newly added Listening ability (Ability 5, pragmatic understanding) yielded correlations comparable to those of the other abilities. Although not reported in the tables owing to space constraints, the intercorrelations of parts and abilities of the operational reference form in the Listening and Reading sections are consistent with the trends observed for the pilot forms.

Table 11 Intercorrelations of Parts Based on Combined Group for Listening

Part	Total	Part 1	Part 2	Part 3	Part 4
Total	–	.53	.87	.94	.93
Part 1	.54	–	.47	.44	.44
Part 2	.86	.47	–	.73	.72
Part 3	.95	.45	.73	–	.82
Part 4	.91	.42	.69	.80	–

Note. Part 1 = Photographs; Part 2 = Question–Response; Part 3 = Short Conversations; Part 4 = Short Talks.

Table 12 Intercorrelations of Abilities Based on Combined Group for Listening

Ability	Total	Ability 1	Ability 2	Ability 3	Ability 4	Ability 5
Total	–	.80	.88	.80	.97	.83
Ability 1	.80	–	.65	.67	.69	.82
Ability 2	.88	.60	–	.64	.83	.70
Ability 3	.79	.63	.64	–	.69	.69
Ability 4	.97	.69	.82	.67	–	.77
Ability 5	.79	.73	.63	.63	.74	–

Note. Ability 1, can infer gist, purpose, and basic context based on information that is explicitly stated in short spoken texts; Ability 2, can infer gist, purpose, and basic context based on information that is explicitly stated in extended spoken texts; Ability 3, can understand details in short spoken texts; Ability 4, can understand details in extended spoken texts; Ability 5, can understand a speaker's purpose or implied meaning in a phrase or sentence (pragmatic understanding).

Table 13 Intercorrelations of Parts Based on Combined Group for Reading

Part	Total	Part 5	Part 6	Part 7	Part 7A	Part 7B
Total	–	.88	.81	.95	.91	.79
Part 5	.87	–	.68	.71	.71	.56
Part 6	.79	.69	–	.68	.70	.50
Part 7	.93	.66	.61	–	.92	.88
Part 7A	.88	.67	.64	.90	–	.63
Part 7B	.76	.49	.43	.87	.58	–

Note. Part 5 = Incomplete Sentences; Part 6 = Text Completion; Part 7 = Reading Comprehension; Part 7A = Single Passages; Part 7B = Multiple Passages.

Table 14 Intercorrelations of Abilities Based on Combined Group for Reading

Ability	Total	Ability 5	Ability 6	Ability 7	Ability 7A	Ability 7B
Total	–	.89	.85	.93	.91	.86
Ability 1	.83	–	.76	.86	.73	.67
Ability 2	.79	.64	–	.78	.68	.64
Ability 3	.91	.81	.72	–	.79	.69
Ability 4	.90	.66	.60	.77	–	.79
Ability 5	.85	.60	.53	.66	.78	–

Note. Ability 1, can locate and understand specific information in tables and passages; Ability 2, can connect information across multiple sentences in a single text and across texts; Ability 3, can make inferences based on information in written texts; Ability 4, can understand vocabulary in workplace texts; Ability 5, can understand grammar in workplace texts.

Reliability

Reliability provides an indication of the extent to which test scores are consistent across different conditions of administration of the same form or alternate forms. In general, when all else is equal, more items tend to lead to higher reliability. The reliability of the TOEIC Listening and Reading Test is estimated using an internal consistency method (reliability coefficient called alpha) based on the correlations between different items on the same test. The reliability estimate ranges from 0 to 1. The higher the reliability coefficient for a section, part, or test, the higher the consistency of test takers' responses to the items of that section, part, or test.

Tables 15–16 display the reliability estimates for the total test and for different parts of the test and abilities for the pilot forms and the reference form in Listening and Reading. Overall, the reliabilities of the total test were nearly the same for the pilot forms and the operational reference form (.94 on average for Listening and Reading). Photographs (Part 1) produced the lowest reliability in the pilot forms. The reliability coefficients of the other parts of the test in both Listening and Reading were aligned with the reliabilities observed in the reference form and in typical operational forms.

The reliabilities of the ability scores were moderate to high and also comparable between the pilot forms and reference form. The newly added Listening ability (Ability 5, pragmatic understanding) yielded the lowest reliabilities because the number of items included in this ability was lower than the minimum number used in operational practice (i.e., 15) in Forms E and F (see Table 2).

Table 15 Reliability Estimates for Listening

Part or ability	Form E	Form F	Reference form
Total test	.94 (100)	.94 (100)	.94 (100)
Part 1. Photographs	.43 (6)	.37 (6)	.50 (10)
Part 2. Question–Response	.78 (25)	.79 (25)	.82 (30)
Part 3. Short Conversations	.88 (39)	.87 (39)	.84 (30)
Part 4. Short Talks	.82 (30)	.86 (30)	.85 (30)
Ability 1	.68 (16)	.68 (15)	.74 (19)
Ability 2	.76 (19)	.73 (16)	.70 (17)
Ability 3	.69 (15)	.68 (16)	.72 (21)
Ability 4	.89 (50)	.90 (53)	.89 (43)
Ability 5	.59 (11)	.67 (13)	–

Note. Numbers in parentheses indicate number of items. Ability 1, can infer gist, purpose, and basic context based on information that is explicitly stated in short spoken texts; Ability 2, can infer gist, purpose, and basic context based on information that is explicitly stated in extended spoken texts; Ability 3, can understand details in short spoken texts; Ability 4, can understand details in extended spoken texts; Ability 5, can understand a speaker's purpose or implied meaning in a phrase or sentence (pragmatic understanding).

Table 16 Reliability Estimates for Reading

Part or ability	Form E	Form F	Reference form
Total test	.93 (100)	.94 (100)	.94 (100)
Part 5. Incomplete Sentences	.86 (30)	.85 (30)	.88 (40)
Part 6. Text Completion	.68 (16)	.73 (16)	.60 (12)
Part 7. Reading Comprehension	.88 (54)	.91 (54)	.90 (48)
Part 7A. Single Passages	.81 (29)	.87 (29)	.83 (28)
Part 7B. Multiple Passages	.81 (25)	.82 (25)	.83 (20)
Ability 1	.64 (18)	.76 (20)	.74 (16)
Ability 2	.69 (13)	.73 (11)	.75 (16)
Ability 3	.80 (35)	.83 (35)	.80 (25)
Ability 4	.80 (28)	.80 (26)	.81 (29)
Ability 5	.77 (20)	.78 (20)	.81 (27)

Note. Numbers in parentheses indicate number of items. Ability 1, can locate and understand specific information in tables and passages; Ability 2, can connect information across multiple sentences in a single text and across texts; Ability 3, can make inferences based on information in written texts; Ability 4, can understand vocabulary in workplace texts; Ability 5, can understand grammar in workplace texts.

Speededness

The TOEIC Listening section is paced by a tape recording, and thus speededness is not a concern. Four types of statistics frequently used to evaluate the speededness of the Reading test are presented in Table 17: (a) percentage of test takers reaching all items, (b) percentage of test takers completing 75% of the items, (c) number of items reached by 80% of the test takers, and (d) ratio of not reached variance (NRV) to total variance (TV). Typically, a test is regarded as unspeeded for a group if (a) nearly all test takers complete 75% of the items, (b) at least 80% of the test takers reach all items, and (c) the ratio of NRV to TV is less than 0.15. As shown in Table 17, Reading was speeded for Forms E and F for Japan. The percentage reaching all items was 79% in both forms. Typically, this percentage in operational settings is about 95% for Japan.

In this pilot study, the last five items had nonreached rates of about 20%. The values of the speededness index (i.e., ratio of NRV to TV) for Japan were much higher than a conventional criterion of .15. For the combined group, the Reading section was slightly speeded.

Table 17 Statistics of Speededness for Reading Sections

Statistic	Form E: Japan	Form E: Korea	Form E: Combined	Form F: Japan	Form F: Korea	Form F: Combined	Reference form: Japan	Reference form: Korea	Reference form: Combined
Number of test takers	1,019	824	1,843	1,026	804	1,830	48,745	38,500	87,245
% reaching all items	79.1	96.5	86.9	79.0	96.8	86.8	96.4	99.4	97.9
% reaching 75% of items	97.2	98.9	97.9	96.7	99.5	97.9	99.6	99.8	99.7
Number of items reached by 80%	97	100	100	97	100	100	100	100	100
Ratio of NRV to TV	0.27	0.10	0.18	0.25	0.06	0.15	0.04	0.02	0.02

Note. NRV = not reached variance; TV = total variance.

One of the chief purposes of the updated TOEIC Listening and Reading test was to ensure that the psychometric properties of the updated test were comparable to those of the preupdated test. The results presented in this report suggest that the updated pilot forms were equally discriminating on average on the total test and on different parts of the test as the operational reference form. The correlations among parts and ability scores were similar to the correlations observed in the operational reference form. Likewise, the newly added Listening ability (Ability 5, pragmatic understanding) produced correlations comparable to those of the other abilities. Overall, the reliabilities of Listening and Reading, parts, and ability scores in the updated pilot forms were similar to the reliabilities of the operational reference form. However, the updated Reading pilot forms appeared to be speeded for Japan. Additionally, the results of the pilot study indicate that for both Listening and Reading, the items on the updated pilot forms were, on average, slightly more difficult than the items on the operational reference form. These findings were shared with test developers in order that they could make the appropriate adjustments to the difficulty of some items.

Operational Results

Since the launch of the updated TOEIC Listening and Reading test, the difficulty of the updated test and reliability of its scores have been closely monitored. To illustrate how the TOEIC test has continued to maintain the psychometric properties of the preupdated test, Table 18 provides a test performance comparison between preupdated and updated operational TOEIC Listening and Reading forms based on Japan. The difference in average equated delta between preupdated and updated forms is .23 for both Listening and Reading. This difficulty difference is considered small.[3] In this regard, it is important to note that operational data have shown that, unlike for the Reading pilot forms, the percentage of reaching all items for Japan has been the same as the percentage observed for the preupdated forms (about 95%). Test discrimination and reliability have also not changed since the updates to the TOEIC test. The test continues to be equally discriminating (*R*-biserial ranges from .45 to .47 in Listening and Reading) and equally reliable (average reliability of .93). The average scaled scores are also quite stable. After forms are equated and the test scores are adjusted based on the difficulty levels of the forms, the average scaled scores for each section are relatively close.

Table 18 Summary Statistics of Preupdated and Updated TOEIC Listening and Reading Forms

Statistic	Preupdated form: Equated delta mean	Preupdated form: *R*-biserial mean	Preupdated form: Reliability	Preupdated form: Scale score mean	Updated form: Equated delta mean	Updated form: *R*-biserial mean	Updated form: Reliability	Updated form: Scale score mean
Listening mean	12.66	0.47	0.93	320.46	12.89	0.47	0.93	317.35
Listening *SD*	0.15	0.01	0.01	4.75	0.12	0.01	0.00	5.14
Listening min	12.30	0.44	0.92	312.89	12.60	0.45	0.92	306.50
Listening max	13.00	0.49	0.94	331.93	13.30	0.49	0.94	325.89
Reading mean	12.23	0.47	0.93	263.37	12.46	0.45	0.93	261.98
Reading *SD*	0.21	0.02	0.01	4.32	0.17	0.01	0.01	4.75
Reading min	11.90	0.44	0.92	253.35	12.00	0.42	0.91	252.37
Reading max	12.80	0.51	0.94	271.38	12.80	0.49	0.94	273.87

Note. $N = 49$. Preupdated forms are forms administered between November 2013 and April 2016. Updated forms are forms administered between May 2016 and May 2017. SD = standard deviation.

Similar trends were observed in a test performance comparison of 23 preupdated and 23 updated operational forms based in Korea. Average difficulty, discrimination, and reliability of the forms were also consistent between the preupdated and updated forms.

In summary, given the difficulty, discrimination, reliability, and scaled score values observed in operational practice, one can say that the updated TOEIC test continues to have the same psychometric quality as the preupdated TOEIC test.

Conclusion

Beginning with the public test in May 2016, the TOEIC Listening and Reading Test included some updates to the question formats to reflect the changing use of English and the ways in which individuals commonly communicate in everyday social and workplace situations around the world. A pilot study conducted in May 2015 to evaluate the statistical properties of the updated TOEIC Listening and Reading Test demonstrated that the psychometric properties of the updated pilot forms were comparable to those of the preupdated reference form. Overall, discrimination of items and sections; correlations among parts and ability scores; and reliabilities of sections, parts, and ability scores were similar to the ones observed in the operational reference form. The slight differences in difficulty levels observed in the pilot study were addressed by test developers, who made appropriate adjustments to the difficulty levels of some items. Operational data gathered after the launch of the updated test suggest that the TOEIC Listening and Reading Test continues to have the same appropriate psychometric properties (e.g., difficulty, discrimination, reliability) as the preupdated test.

References

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing differential item functioning on the Scholastic Aptitude Test. *Journal of Educational Measurement*, *23*, 355–368. https://doi.org/10.1111/j.1745-3984.1986.tb00255.x

Holland, P. W., & Thayer, D. T. (1988). *An alternative definition of the ETS delta scale of item difficulty* (Research Report No. RR-85-43). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1985.tb00128.x

Zwick, R. (2012). *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement* (Research Report No. RR-12-08). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2012.tb02290.x

Notes

[1] A summary statistic that expresses the mean difference between two groups in standard deviation units.

[2] Type of delta that indicates how difficult an item would be after placed on the same scale for all forms.

[3] A difference of .23 in equated delta converts approximately to a difference of .02 in *p*-value or proportion correct or to a difference of 2% in percentage correct.