Measuring the Power of Learning.®

Research Memorandum
ETS RM–19-09

Field Study Statistical Analysis for the Redesigned *TOEIC Bridge*® Tests

Peng Lin

Jaime Cid

Jiayue Zhang

August 2019

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.

Field Study Statistical Analysis for the Redesigned *TOEIC Bridge*® Tests

Peng Lin, Jaime Cid, and Jiayue Zhang
Educational Testing Service, Princeton, New Jersey

August 2019

Corresponding author: P. Lin, E-mail: plin@ets.org

Suggested citation: Lin, P., Cid, J., & Zhang, J. (2019). *Field study statistical analysis for the redesigned* TOEIC Bridge® *tests* (Research Memorandum No. RM-19-09). Princeton, NJ: Educational Testing Service.

Action Editor: Don Powers

Reviewers: Timothy Davey and Rick Morgan

Abstract

To better assess a test taker's basic to intermediate English language proficiency skills in common everyday activities, ETS launched in June 2019 a comprehensive suite of redesigned *TOEIC Bridge*® tests that includes all 4 essential communication skills (listening, reading, speaking, and writing). This paper reports the results of a field study conducted in April 2018 to evaluate the statistical properties of the redesigned TOEIC Bridge tests. The evaluation of the difficulty and discrimination of the items, correlation among different parts of the test, reliability of the scores, and interrater reliability for speaking and writing, not only helped to inform the final decisions regarding the final reporting scales of the redesigned TOEIC Bridge tests, but also allowed test developers make appropriate adjustments to the test design before the official launch in June 2019.

Key words: *TOEIC Bridge*® tests, redesign, field study, statistical analysis

The *TOEIC Bridge*® tests are English language proficiency tests for nonnative speakers of English designed to measure language proficiency at the beginning and the lower-intermediate levels. Test takers may be students of English or those who need to use English for work or travel. From its inception through 2018, the original TOEIC Bridge test consisted of two separate timed sections: listening and reading, with 50 items in each section. The listening section was paced by audio recording.

In 2016, based on feedback received from clients, the Educational Testing Service (ETS) decided to redesign the original TOEIC Bridge test. The redesigned TOEIC Bridge tests were launched in June 2019. Two changes to the test occurred. First, the redesigned TOEIC Bridge tests focus on communication in the context of everyday adult life (personal, public, and familiar workplace contexts) for the beginning to lower-intermediate English language learners. Second, the redesigned TOEIC Bridge tests also measure speaking and writing communication skills. Unlike the original TOEIC Bridge test, the redesigned tests are a module-based assessment with four modules: listening, reading, speaking, and writing. It is possible to take a single module or any combination of the modules as needed. The redesigned tests measure English language listening, reading, speaking, and writing proficiency of test takers at the levels of A1, A2, and B1 of the Common European Framework of Reference[1] (CEFR; Council of Europe, 2001).

A variety of item types of the redesigned TOEIC Bridge tests were evaluated by content experts (see Everson et al., 2019). An item-level pilot study was administered in September 2017 in Japan, Korea, Taiwan, and Brazil to help specify both the appropriate item types and the appropriate number of items within each item type for all four skills (tests). Observations from the pilot study (e.g., item difficulty, format appropriateness, and testing time) were used to refine the item and test specifications for the redesigned TOEIC Bridge tests.

In April 2018, a field study was launched in three Asian countries (Japan, Korea, and Taiwan) and three non-Asian countries (Colombia, Brazil, and Mexico), in which the original Bridge test was well adopted. After the data collection was completed, statistical analyses were conducted to evaluate the statistical properties of the redesigned TOEIC Bridge tests (e.g., difficulty and discrimination of the items, correlation among different parts of the test, reliability, interrater reliability for speaking and writing). The purpose of this report is to document the results of the statistical analyses of the listening, reading, speaking, and writing tests of the field study. These results contributed to the conceptual assessment framework and assessment

implementation layers of the evidence-centered design test development process that was utilized for the development of the redesigned TOEIC Bridge tests (see Mislevy & Yin, 2012). Although not part of this report, the results from the statistical analyses of the field study informed the final decisions on the reporting scales of the redesigned TOEIC Bridge tests and the performance proficiency levels for listening, reading, speaking, and writing. The reported score scales of all four tests were set to range from 15 to 50 in increments of 1.

Background: Field Study Test Specifications

The redesigned TOEIC Bridge Listening and Reading tests contain only multiple-choice items that are scored dichotomously. As shown in Table 1, the listening test consists of four parts and the reading test consists of three parts. Unlike the original TOEIC Bridge test, which had two subscore areas for the listening section and three for the reading section, four ability measures were developed for each test (i.e., listening and reading) of the redesigned TOEIC Bridge. The four abilities are reported to test takers using a percentage correct score. Table 2 presents the number of items associated with the abilities in the listening and reading tests of the field study. The position and the number of items associated with each ability may vary across operational forms. The redesigned TOEIC Bridge Speaking test consists of six constructed-response item types. The redesigned TOEIC Bridge Writing test consists of four constructed-response item types and one multiple-selection multiple-choice item type (Build a Sentence). See Tables 3 and 4 for details.

Table 1. Parts of the Redesigned TOEIC Bridge Listening and Reading Tests

Part	Number of items
Listening	
Part 1. Four Pictures	6
Part 2. Question-Response	20
Part 3. Conversations	10
Part 4. Talk	14
Reading	
Part 1. Sentence Completion	15
Part 2. Text Completion	15
Part 3. Reading Comprehension	20

Table 2. Ability Measures of the Redesigned TOEIC Bridge Listening and Reading Tests

Ability	Number of items
Listening	
Appropriate Response	20
Short Dialogue or Conversation	32
Short Monologue	12
Main Idea or Stated Fact	23
Reading	
Vocabulary	14
Grammar	13
Main Idea or Stated Fact	16
Short Informational Written Texts	20

Note. The listening and reading tests each have 50 items. The sum of items for all abilities is greater than 50 as some items contribute to more than one ability.

Table 3. Item Types of the Redesigned TOEIC Bridge Speaking Test

Item	Item type	Score scale
1–2	Read a Short Text Aloud	0–3
3–4	Describe a Photograph	0–3
5	Listen and Retell	0–3
6	Short Interaction	0–3
7	Tell a Story	0–4
8	Make and Support a Recommendation	0–4

Table 4. Item Types of the Redesigned TOEIC Bridge Writing Test

Item	Item type	Score scale
1–3	Build a Sentence	0–2
4–6	Write a Sentence	0–3
7	Respond to a Brief Message	0–3
8	Write a Narrative	0–3
9	Respond to an Extended Message	0–4

Field Study Test Data Collection

Two parallel test forms for listening and reading (Form LR1 and Form LR2) and two for speaking and writing (Form SW1 and Form SW2) were assembled and administered in the field study. All items were new with no previous statistics available. The two listening and reading forms shared 20 common items in listening and 20 in reading (i.e., 40% of the total items in the test). No items were common between the two speaking and writing forms.

The test was administered in two separate sessions: one for listening and reading and one for speaking and writing. In each session, the two forms were randomly administered to test takers in order to make the test-taking groups of the two forms approximately equivalent. For listening and reading, the scores of the two forms were equated through common items and converted to scale scores. For speaking and writing, the scores were made comparable between forms through well-defined and articulated scoring rubrics and quality control procedures. Thus,

the scale scores from the two forms can be deemed comparable within each test (i.e., listening, reading, speaking, and writing) of the field study.

In total, 2,368 test takers from six countries (Japan, Korea, Taiwan, Colombia, Brazil, and Mexico) participated and took all four tests in the field study. Although an effort was made to recruit test takers from all the ability scale ranges of the target population (i.e., A1, A2, and B1), the small samples collected from some countries precluded a balanced ability distribution in all countries. In addition, the number of test takers from Colombia and Mexico was noticeably below the targeted numbers. Tables 5 and 6 summarize demographic compositions of the field study sample by country and by gender. Approximately half of the test takers were from Japan.

Table 5. Country Distributions of Test Takers in Field Study

Country	*N*	Percent
Brazil	251	11%
Colombia	18	1%
Japan	1,250	53%
Korea	391	17%
Mexico	49	2%
Taiwan	409	17%
Total	2,368	100%

Table 6. Country Distributions of Test Takers in Field Study

Gender	*N*	Percent
Female	1,118	47%
Male	1,249	53%
Unidentified	1	
Total	2,368	100%

Performance by Country and Gender

Table 7 provides the mean and standard deviation of the scale scores of the field study for the four tests by country. Recall that the redesigned TOEIC Bridge test has the scale scores of all four tests reported on a scale from 15 to 50 in increments of 1. On average, Japanese test takers were the most able group among the six countries for both listening and reading. This finding is different from what was observed in the original TOEIC Bridge test in operational settings, where Korean test takers performed better than Japanese test takers in both listening and reading. Therefore, the field study sample may not have been representative of the operational test-taking population. Scaled scores of Taiwanese test takers were close to those of Japanese test takers in listening but were noticeably lower in reading. For speaking, Colombian and Mexican test takers scored higher than the other countries, and for writing, Japan, Colombia, and Mexico were the

countries that had the highest scaled scores. Test takers from Brazil produced the lowest scaled score means in all four tests. When interpreting these results, it is important to note that Colombia and Mexico had considerably smaller samples of test takers than the other countries.

Table 7. Mean and Standard Deviation of the Test Scores by Country

Country	*N*	Listening *M*	Listening *SD*	Reading *M*	Reading *SD*	Speaking *M*	Speaking *SD*	Writing *M*	Writing *SD*
Brazil	251	21.95	9.92	25.39	9.93	24.02	11.08	26.96	11.79
Colombia	18	32.06	9.32	39.78	8.24	38.94	8.79	41.17	8.30
Japan	1,250	36.35	8.35	42.58	7.44	36.37	8.54	41.47	8.04
Korea	391	31.63	9.81	32.93	10.51	32.25	9.71	35.85	9.93
Mexico	49	31.35	12.39	36.10	10.72	38.55	9.76	41.67	8.62
Taiwan	409	36.09	10.20	39.35	10.75	35.99	11.12	38.29	11.78
Total	2,368	33.86	10.24	38.45	10.63	34.38	10.29	38.46	10.54

The mean and standard deviation of scale scores of the field study by gender are provided in Table 8. As can be seen from the table, on average, female test takers performed better than male test takers in all four tests in all countries. We observed this same trend with the original TOEIC Bridge test in operational settings.

Table 8. Mean and Standard Deviation of the Test Scores by Gender

Gender	*N*	Listening *M*	Listening *SD*	Reading *M*	Reading *M*	Speaking *M*	Speaking *SD*	Writing *M*	Writing *SD*
Female	1,118	34.89	10.44	39.25	10.42	35.27	10.64	39.41	10.42
Male	1,249	32.94	9.97	37.73	10.77	33.59	9.90	37.61	10.59
Unidentified	1	38.00		38.00		23.00		32.00	
Total	2,368	33.86	10.24	38.45	10.63	34.38	10.29	38.46	10.54

Statistical Analysis Results

The analyses presented in the next sections are based on the combined field study samples, with a total of 2,050 test takers, from the countries with large numbers of test takers—Japan, Korea, and Taiwan—in operational administrations. These countries comprised 83% of the field study sample. The summary statistics of the scaled scores, including mean, standard deviation, minimum, maximum, and the correlation among tests, are presented in Table 9. On average, the reading and writing tests yielded the highest means, while the listening and speaking tests yielded the lowest means. The Pearson correlation coefficients for the four tests in Table 10 show that the four sets of test scores were moderately correlated. These correlations are similar

to the ones reported for the *TOEIC*® Listening, Reading, Speaking, and Writing tests (e.g., Liu & Costanzo, 2013).

Table 9. Summary Statistics of Test Scores for the Field Study Sample (Japan, Korea, and Taiwan)

Test	Listening	Reading	Speaking	Writing
N	2,050	2,050	2,025[a]	2,050
Mean	35.40	40.09	35.52	39.76
SD	9.21	9.56	9.45	9.54
Minimum	15	15	15	15
Maximum	50	50	50	50

[a]Twenty-five test takers with some non-scorable responses in the speaking test did not have speaking scores.

Table 10. Correlation Coefficients for the Four Tests of the Field Study Sample (Japan, Korea, and Taiwan)

Correlation	Listening	Reading	Speaking	Writing
Listening	1	.78	.68	.66
Reading		1	.66	.74
Speaking			1	.71
Writing				1

Table 11 provides the mean and standard deviation of scale scores for the two listening and reading forms (LR1 and LR2) and for the two speaking and writing forms (SW1 and SW2) for test takers from Japan, Korea, and Taiwan. One can see that the mean and standard deviation of the scale scores of the two forms within each test were very close, which indicates that the groups taking LR1 and LR2 in listening and reading and SW1 and SW2 in speaking and writing were approximately equivalent. It also indicates that the approaches to making speaking and writing scores comparable across forms appeared successful in the field study.

Table 11. Mean and Standard Deviation of the Test Scores by Form

Form	*N*	Listening *M*	Listening *SD*	Reading *M*	Reading *SD*	Speaking *M*	Speaking *SD*	Writing *M*	Writing *SD*
LR1	1,018	35.28	9.01	40.45	9.51				
LR2	1,032	35.51	9.41	39.74	9.61				
SW1	1,028					35.90	9.32	39.83	9.31
SW2	1,022					35.14	9.57	39.69	9.76

Note. LR = listening and reading; SW = speaking and writing.

Due to differences in test format across tests (i.e., multiple-choice items for listening and reading and constructed-response items for speaking and writing), we conducted separate statistical analyses for listening and reading and for speaking and writing. In the following sections, the statistical analysis results are presented in the same order.

Listening and Reading

Reliability and Standard Error of Measurement

The reliabilities of the listening and reading tests in the field study were estimated using an internal consistency method (reliability coefficient called alpha). This method assesses the consistency of test takers' responses to all of the items in the test, part, or ability. The reliability estimate ranges from 0 to 1. The higher the reliability coefficient is for the test, part, or ability, the higher the consistency of test takers' responses is to the items of the test, part, or ability. The standard error of measurement (SEM)—another indicator of score consistency—estimates the average variation expected in a test taker's score from one test form to another.

The reliability coefficients and the SEMs for the total test and different parts and abilities of the two listening and reading field study forms are reported in Tables 12 and 13, respectively. The reliabilities of listening in Form LR1 and Form LR2 were .88 and .89, respectively. Reading produced reliabilities of .93 in both forms. In listening, Four Pictures (Part 1) with six items and Question Response (Part 2) with 20 items produced the lowest reliability and highest reliability, respectively, in both forms. Likewise, Reading Comprehension (Part 3) with 20 items produced the highest reliability in reading. The reliabilities of both listening and reading of the field study forms were relatively higher than the average reliabilities of listening (.83) and reading (.85) of the original TOEIC Bridge test forms. The reliabilities of three of the four abilities in listening and all four abilities in reading were above .75 in both forms. To increase the reliabilities of Short Monologue in listening, which were .68 and .71, respectively, in the two LR forms, it was decided to add two more items to this ability in the operational forms. The SEM of total score was very close in the two forms in both listening and reading, with listening yielding a slightly higher total SEM than reading (3.0 vs. 2.5).

Table 12. Reliability and SEM for All Scores of Listening Test by Form

Part/ability	LR1– number of items	LR1– reliability	LR1– SEM	LR2– number of items	LR2– reliability	LR2– SEM
Part						
Part 1. Four Pictures	6	.43	0.73	6	.47	0.66
Part 2. Question-Response	20	.77	1.67	20	.78	1.64
Part 3. Conversations	12	.70	1.34	12	.68	1.30
Part 4. Talk	12	.68	1.43	12	.71	1.42
Ability						
Appropriate Response	20	.77	1.68	20	.78	1.64
Short Dialogue or Conversation	32	.84	2.16	32	.85	2.10
Short Monologue	12	.68	1.43	12	.71	1.42
Main Idea and/or Stated Fact	23	.80	1.94	22	.82	1.85
Total (scale score)	50	.88	3.07	50	0.89	3.09

Note. LR = listening and reading; SEM = standard error of measurement. Form LR1: N = 1,018. Form LR2: N = 1,032. The sum of items for all abilities in a form might be greater than 50 as some items contribute to more than one ability.

Table 13. Reliability and SEM for All Scores of Reading Test by Form

Part/ability	LR1– number of items	LR1– reliability	LR1– SEM	LR2– number of items	LR2– reliability	LR2– SEM
Part						
Part 1. Sentence Completion	15	.77	1.44	15	.80	1.42
Part 2. Text Completion	15	.83	1.24	15	.77	1.43
Part 3. Reading Comprehension	20	.87	1.65	20	.89	1.66
Ability						
Vocabulary	14	.78	1.22	13	.78	1.21
Grammar	13	.78	1.32	14	.77	1.43
Main Idea or Stated Fact	16	.86	1.49	16	.87	1.48
Short Informational Written Texts	20	.87	1.65	20	.89	1.66
Total (scale score)	50	.93	2.49	50	.93	2.51

Note. LR = listening and reading; SEM = standard error of measurement. Form LR1: N = 1,018. Form LR2: N = 1,032. The sum of items for all abilities in a form might be greater than 50 as some items contribute to more than one ability.

Please note that the magnitude of reliability depends not only on the internal consistency of the items in the test but also on the homogeneity of the test takers. The reliabilities for the field study forms in this report may not be directly comparable to what one would observe in operational settings, as the field study sample may have not been representative of the operational test-taking population.

Item Difficulty

In this study, item difficulty was evaluated by examining two types of statistical indices: *p* value (defined as the proportion of the test takers who answered an item correctly in a given population) and delta (defined as 13 minus 4z, where z is the standard normal deviate corresponding to proportion correct). The *p* value ranges from 0 to 1. The higher the *p* value is, the easier the item is. The *p* value is dependent on the ability levels of the sample taking the test. That is, the *p* value of the same items based on a more able group will be higher than those based on a less able group. Therefore, *p* values are not directly comparable across forms taken by different groups of test takers to reflect the difficulty of items in different forms. The equated deltas provide a difficulty metric that accounts for the different ability levels across groups that take different forms. Delta values typically range from 6.0 for a very easy item (i.e., approximately 95% correct) to 20 for a very difficult item (i.e., approximately 5% correct); the mean of 13.0 corresponds to 50% correct. To compare the item difficulty between the two field study forms with the original TOEIC Bridge test, equated deltas, which transfer the observed delta of the field test forms to the scale of the original TOEIC Bridge test, were calculated based on a single group design. Specifically, a group of 300 test takers took both the field study Form LR1 and an operational form of the original TOEIC Bridge test. The equated deltas of items in Form LR1 were calculated based on the equated deltas of items in the operational form of the original TOEIC Bridge test. Form LR1 was then used as the reference form and the delta value of LR2 items were adjusted to the scale of LR1 through the anchor items.

The *p* value and equated delta of items for listening are summarized in Tables 14 and 15. As can be seen, on average, the listening tests of the two field study forms were comparable in difficulty. The mean of the equated delta in both forms was 11.7. Among the four parts in listening, Four Picture items were, on average, the easiest, and the Talk items were the most difficult. Among the four abilities in the listening test, Appropriate Response and Short Dialogue or Conversation were relatively easier than Short Monologue and Main Idea or Stated Fact.

Table 14. Item Difficulty in Listening Test of Form LR1

Part/ability	*p* value *M*	*p* value *SD*	Equated delta *M*	Equated delta *SD*
Part				
Part 1. Four Pictures	.88	.09	9.5	2.2
Part 2. Question-Response	.76	.14	11.7	2.0
Part 3. Conversations	.75	.08	12.0	1.0
Part 4. Talk	.70	.10	12.7	1.3
Ability				
Appropriate Response	.76	.14	11.7	2.0
Short Dialogue or Conversation	.76	.12	11.8	1.7
Short Monologue	.70	.10	12.7	1.3
Main Idea or Stated Fact	.72	.10	12.4	1.2
All 50 items	.76	.13	11.7	1.9

Table 15. Item Difficulty in Listening Test of Form LR2

Part/ability	*p* value *M*	*p* value *SD*	Equated delta *M*	Equated delta *SD*
Part				
Part 1. Four Pictures	.90	.06	9.3	1.6
Part 2. Question-Response	.77	.14	11.6	1.9
Part 3. Conversations	.74	.16	12.0	1.8
Part 4. Talk	.68	.11	12.8	1.2
Ability				
Appropriate Response	.77	.14	11.6	1.9
Short Dialogue or Conversation	.76	.15	11.8	1.9
Short Monologue	.68	.11	12.8	1.2
Main Idea or Stated Fact	.70	.14	12.6	1.6
All 50 items	.76	.15	11.7	2.0

The results for reading are summarized in Table 16 and Table 17. As is shown, on average, the reading tests of the two field study forms were comparable in difficulty. The mean of the equated delta of reading tests of the two forms were 11.3 and 11.4, respectively. Among the three parts in the reading test, Text Completion, on average, was easier than Sentence Completion and Reading Comprehension in Form LR1. But this was not the case in Form LR2, where Text Completion, on average, was as difficult as Sentence Completion and easier than Reading Comprehension. Among the four abilities in the reading test, Vocabulary was easier than the remaining three abilities, which had similar difficulties in both forms.

Table 16. Item Difficulty in Reading Test of Form LR1

Part/ability	*p* value *M*	*p* value *SD*	Equated delta *M*	Equated delta *SD*
Part				
Part 1. Sentence Completion	.72	.18	11.8	2.9
Part 2. Text Completion	.83	.08	10.0	1.5
Part 3. Reading Comprehension	.72	.14	11.9	2.3
Ability				
Vocabulary	.80	.16	10.3	2.6
Grammar	.74	.13	11.5	2.2
Main Idea or Stated Fact	.72	.11	11.9	2.0
Short Informational Written Texts	.72	.14	11.9	2.3
All 50 items	.75	.15	11.3	2.5

Table 17. Item Difficulty in Reading Test of Form LR2

Part/ability	*p* value *M*	*p* value *SD*	Equated delta *M*	Equated delta *SD*
Part				
Part 1. Sentence Completion	.77	.08	11.1	1.6
Part 2. Text Completion	.76	.12	11.1	2.0
Part 3. Reading Comprehension	.72	.11	11.9	1.9
Ability				
Vocabulary	.82	.06	10.2	1.2
Grammar	.70	.10	10.7	1.7
Main Idea or Stated Fact	.72	.11	11.9	1.9
Short Informational Written Texts	.72	.11	11.9	1.9
All 50 items	.75	.11	11.4	1.9

Tables 18 and 19 provide a comparison of the equated deltas of the two field study forms with those of the operational forms of the original TOEIC Bridge test since 2013. As can be seen from the tables, the average equated delta of the listening test of both field study forms (mean equated delta = 11.7) was slightly higher than the maximum delta (11.6) of the operational forms. The listening test of both field study forms (mean equated delta = 11.7) was on average more difficult than the listening test of the original TOEIC Bridge (mean equated delta = 11.0). The 0.7 difference between the equated delta of the field study forms and the average of the original TOEIC Bridge forms indicates that the listening test of the field study forms was approximately 6% (3 items or points for 50 questions) more difficult than the original TOEIC Bridge test. On the other hand, the equated delta of the reading test of the two field study forms (mean equated delta = 11.3 and 11.4, respectively) was lower than the average of the operational forms (mean equated delta = 11.9) but still within the range of operational forms (11.0 – 12.8). The 0.55 average difference between the equated delta of the field study forms and the average of the original TOEIC Bridge operational forms indicates that the reading test of the field study

forms was approximately 5% (2.5 items or points for 50 questions) easier than the original TOEIC Bridge test. In addition, Tables 18 and 19 also suggest that the difficulty of listening and reading tests may be closer to one another in the redesigned TOEIC Bridge test (mean equated delta = 11.7 for listening vs. mean equated delta = 11.4 for reading in field study) than in the original TOEIC Bridge test (mean equated delta = 11.0 for listening vs. mean equated delta = 11.9 for reading). In summary, although on average the field study forms were more difficult (listening) or easier (reading) than the original TOEIC Bridge, their average difficulty was very close to the operational historical ranges in both measures. Therefore, we expect that the difficulty of the redesigned TOEIC Bridge in the operational samples will be comparable to that of the original TOEIC Bridge.

Table 18. Equated Delta of Listening and Reading of Field Study Forms

Test/field study form	*M*	*SD*
Listening–Form LR1	11.7	1.9
Listening–Form LR2	11.7	2.0
Reading–Form LR1	11.3	2.5
Reading–Form LR2	11.4	1.9

Note. LR = listening and reading. The original TOEIC Bridge test includes operational forms since 2013.

Table 19. Equated Delta of the Original TOEIC Bridge Test

Test	*M*	Minimum	Maximum
Listening	11.0	10.4	11.6
Reading	11.9	11.0	12.8

Item Discrimination

Item discrimination was evaluated by the biserial correlation coefficient. The biserial correlation is the relationship between test takers' scores on a particular item (i.e., 0 for an incorrect response or 1 for a correct response) with the corresponding total score (i.e., sum of item scores for a test). The biserial correlation indicates how well an item serves to discriminate between low- and high-ability test takers. Tables 20 and 21 present the summary statistics of the biserial correlations of items in listening and reading, respectively, in the two field study forms. For listening, the biserials were comparable between the two field study forms and across parts and abilities within the forms. For reading, the biserials, on average, were comparable between the two field study forms but were varied across parts and abilities within the forms. The overall biserial of the listening and reading of the two field study forms were both higher than the average biserial of the original TOEIC Bridge forms listening and reading, .45 and .46, respectively.

Table 20. Biserial Correlations of Items in the Listening Test of Form LR1 and Form LR2

Part/ability	Form LR1 *M*	Form LR1 *SD*	Form LR2 *M*	Form LR2 *SD*
Part				
Part 1. Four Pictures	.46	.10	.47	.06
Part 2. Question-Response	.51	.09	.52	.09
Part 3. Conversations	.50	.09	.52	.10
Part 4. Talk	.49	.08	.52	.09
Ability				
Appropriate Response	.51	.09	.52	.09
Short Dialogue or Conversation	.51	.09	.52	.09
Short Monologue	.49	.08	.52	.09
Main Idea or Stated Fact	.49	.09	.52	.10
All 50 items	.50	.09	.51	.09

Note. LR = listening and reading.

Table 21. Biserial Correlations of Items in Reading Test of Form LR1 and Form LR2

Part/ability	Form LR1 *M*	Form LR1 *SD*	Form LR2 *M*	Form LR2 *SD*
Part				
Part 1. Sentence Completion	.53	.12	.55	.09
Part 2. Text Completion	.60	.09	.53	.08
Part 3. Reading Comprehension	.61	.11	.61	.09
Ability				
Vocabulary	.56	.14	.56	.08
Grammar	.57	.10	.50	.09
Main Idea or Stated Fact	.62	.09	.62	.09
Short Informational Written Texts	.61	.11	.61	.09
All 50 items	.58	.11	.57	.09

Note. LR = listening and reading.

Speededness

The TOEIC Bridge Listening test is paced by an audio recording, and thus speededness could not be assessed in the traditional way. For the reading test of the field study, four types of statistics frequently used to evaluate the speededness of a test are presented in Table 22: (a) percentage of test takers completing the whole test, (b) percentage of test takers completing 75% of the test, (c) the number of items reached by 80% of the test takers, and (d) the ratio of not reached variance to the total score variance (i.e., the speededness index). If nearly all of the test takers complete 75% of the items in a test and if nearly all of the items are reached by at least 80% of the test takers and if the speededness index is .15 or less, the test is usually considered relatively unspeeded. The statistics in Table 22 indicate that the reading test of both field study forms was not speeded.

Table 22. Statistics of Speededness for Reading

Speededness	Form LR1	Form LR2	Original TOEIC Bridge Test		
			M	Minimum	Maximum
Percent completing test	97.8%	97.0%	95.0%	92.3%	97.3%
Percent completing 75%	99.8%	99.6%	99.6%	98.7%	99.9%
Number of items reached by 80%	50	50	49.9	48	50
Speededness index	0.01	0.02	0.05	0.02	0.1

Note. LR = listening and reading. The original TOEIC Bridge test includes operational forms since 2013.

Speaking and Writing

Difficulty

The difficulty of the two field study speaking and writing forms (SW1 and SW2) was evaluated at item level. The means and standard deviations of item scores and total scores (scale score) of the speaking and writing tests of the field study are presented in Tables 23 and 24. Missing responses were excluded when calculating the statistics of item scores. Therefore, the sample sizes were slightly different across items within the same form. In general, the higher the score mean was (relative to its possible score range), the easier the item was for a given group of test takers. The means of the total score of the two forms were comparable to one another in both speaking and writing. Overall, in speaking, the difficulty of items was comparable between the two forms. Among the six item types of the speaking test, Read a Short Text Aloud was relatively easier than the other item types, and Short Interaction was the most difficult item type. In the writing test, one can see a larger variability in difficulty within the same item type. For example, on both forms, the first Build a Sentence item was the easiest and the third Build a Sentence item was the hardest. On average, the Respond to a Brief Message item and the Respond to an Extended Message item were relatively easier than the other item types; Write a Narrative was the hardest. Overall, these results indicate that item types in both the speaking and the writing tests can have different levels of item difficulty. These findings were shared with test developers so they could make appropriate adjustments to the test design and were taken into account when making final decisions on the reporting scales of the redesigned TOEIC Bridge test.

Table 23. Item Difficulty for Speaking Tests of Form SW1 and Form SW2

Item	Item type	Score scale	Form SW1–*N*	Form SW1–*M*	Form SW1–*SD*	Form SW2–*N*	Form SW2–*M*	Form SW2–*SD*
1	Read a Short Text Aloud	0–3	1,012	2.58	0.58	983	2.55	0.59
2	Read a Short Text Aloud	0–3	1,016	2.60	0.57	986	2.61	0.58
3	Describe a Photograph	0–3	1,012	2.37	0.59	977	2.37	0.62
4	Describe a Photograph	0–3	1,011	2.37	0.62	986	2.31	0.61
7	Listen and Retell	0–3	921	2.16	0.74	913	2.15	0.70
5	Short Interaction	0–3	927	1.93	0.84	880	1.75	0.62
5	Short Interaction	0–3	927	1.93	0.84	880	1.75	0.62
6	Tell a Story	0–4	1,005	2.60	0.79	971	2.62	0.82
8	Make and Support a Recommendation	0–4	969	2.73	0.89	936	2.67	0.84

Note. SW = speaking and writing.

Table 24. Item Difficulty for the Writing Tests of Form SW1 and Form SW2

Item	Item type	Score scale	Form SW1–*N*	Form SW1–*M*	Form SW1–*SD*	Form SW2–*N*	Form SW2–*M*	Form SW2–*SD*
1	Build a Sentence	0–2	1,014	1.65	0.48	1,018	1.96	0.20
2	Build a Sentence	0–2	1,023	1.67	0.47	1,022	1.38	0.49
3	Build a Sentence	0–2	1,024	1.22	0.42	1,013	1.16	0.37
4	Write a Sentence	0–3	999	2.22	0.77	987	2.00	0.79
5	Write a Sentence	0–3	1,019	2.23	0.71	1,001	1.57	0.71
6	Write a Sentence	0–3	1,019	2.15	0.71	1,000	2.15	0.73
7	Respond to a Brief Message	0–3	1,010	2.11	0.80	997	2.35	0.80
8	Write a Narrative	0–3	996	2.00	0.76	976	2.13	0.77
9	Respond to an Extended Message	0–4	1,001	2.73	0.93	977	2.98	0.88

Note. SW = speaking and writing.

Item Total Correlations

In order to evaluate the contribution of items to the total scores, Pearson product-moment correlations were calculated between items and the total scores. Items with a high correlation with the total test score are deemed better at discriminating high-ability test takers from low-ability test takers and, therefore, contribute more to the overall test reliability. Observed score correlation coefficients between item score and total raw score (sum of the item scores) are presented in Tables 25 and 26 for the speaking and writing tests, respectively. In the speaking test, the correlations were moderate to high (from .52 to .80). On both forms, Item 8 (Make and Support a Recommendation) and Items 1 and 2 (Read a Short Text Aloud) yielded the highest and lowest correlations, respectively. In the writing test, the item total correlation ranged from

.28 to .79 on both forms. The correlations for Items 1 through 6, especially for Item 3 (one of the Write a Sentence items) were noticeably lower than those for Items 7 through 9. As expected, the item total correlations of the Build a Sentence item type (Items 1, 2, and 3) were, on average, lower than those of the other item types on both forms because of their narrow score range (0 – 2) and extreme difficulty (i.e., items too easy or too difficult). Item 9 (Respond to an Extended Message) was the item that produced the highest item total correlation on both forms.

Table 25. Item Total Correlation for Speaking Forms SW1 and SW2

Form	Total score	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8
SW1	Speaking raw score	.52	.53	.61	.58	.60	.73	.68	.80
SW2	Speaking raw score	.57	.58	.61	.59	.54	.77	.66	.79

Note. SW = speaking and writing.

Table 26. Item Total Correlation for Writing Forms SW1 and SW2

Form	Total score	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9
SW1	Writing raw score	.40	.38	.30	.41	.48	.41	.69	.72	.79
SW2	Writing raw score	.28	.51	.34	.34	.38	.52	.67	.76	.79

Note. SW = speaking and writing.

Test Reliability and Standard Error of Measurement

The traditional coefficient alpha index—a measure of internal consistency—was used to estimate the reliability of speaking and writing tests. Table 27 displays the internal consistency reliability coefficients and SEM of test scores for the two forms of the speaking and writing tests. Although SEM was comparable between the two forms for both speaking and writing, respectively, speaking yielded smaller SEM than writing. The reliability estimates for the two speaking forms were .83 and .86, respectively. The reliability estimates for the two writing forms were .73 and .75, respectively. Based on feedback provided by test developers regarding conceptual communalities for some item types and the different levels of item difficulty observed in the field study, we also evaluated the consistency of test-taker performance on individual items within three levels of item difficulty: low, medium, and high. The coefficient alpha calculated based on the alphas within each classification is known as stratified coefficient alpha. Table 27 shows that in the speaking test the coefficient alpha and stratified alpha were quite comparable on both forms. However, in the writing test, the stratified alpha was higher than the coefficient alpha, especially on Form SW2. These findings informed the final decisions regarding the test design and the reporting scales of the redesigned TOEIC Bridge test.

Specifically, the findings were used to determine the appropriate weights for each of the item types and evaluate their impact on the reliability of total scores.

Table 27. Reliability and SEM of Speaking and Writing of Forms SW1 and SW2

Alpha	Form SW1–reliability	Form SW1–SEM	Form SW2–reliability	Form SW3–SEM
Coefficient alpha speaking	.83	3.82	.86	3.54
Coefficient alpha writing	.73	4.79	.75	4.85
Stratified alpha speaking	.84	3.73	.87	3.45
Stratified alpha writing	.78	4.37	.82	4.14

Note. SW = speaking and writing; SEM = standard error of measurement.

Interrater Agreement

Because all items in the TOEIC Bridge Speaking and Writing tests (except for Writing Items 1, 2, and 3, which are multiple choice items) were scored by two independent raters in the field study, it was important to evaluate the consistency of ratings given by different raters. The agreement rates between the first and second ratings are presented in Tables 28 and 29. In the speaking test, the percentage of exact agreement ranged from 57% to 81% on both forms, indicating that, for all items, more than half of the test takers received the same ratings from the first and second raters. The percentage of discrepancy was below 1% for most speaking items. Item 6 (Tell a Story) yielded the highest percentage of discrepancy on both forms (2.85% and 2.65% on forms SW1 and SW2, respectively), indicating that only a few test takers were given a score with a difference of two or more points from the two raters. This finding was consistent with the interrater correlation for speaking items, which ranged from .56 to .89. The interrater agreement observed in writing items, on average, was higher than that of the speaking items. In writing, the percentage of exact agreement ranged from 63% to 91%. The highest discrepancy rate (1.31%) was produced by Item 9 (Respond to an Extended Message). The interrater correlation for writing items ranged from .77 to .92.

Table 28. Interrater Agreement and Reliability of Speaking for Forms SW1 and SW2

Item	Form SW1–*N*	Form SW1–exact %	Form SW1–adj. %	Form SW1–dis. %	Form SW1–interrater	Form SW2–*N*	Form SW2–exact %	Form SW2–adj. %	Form SW2–dis. %	Form SW2–interrater
1	1,019	73	27	0.2	.62	1,005	69	30	0.3	.67
2	1,019	73	27	0.1	.59	1,006	72	28	0.4	.67
3	1,019	66	33	0.2	.56	1,006	68	32	0.2	.70
4	1,019	67	33	0.5	.59	1,006	62	37	0.4	.59
5	1,019	78	21	0.8	.87	1,006	74	25	1.1	.78
6	1,019	60	37	2.9	.67	1,006	57	41	2.7	.70
7	1,019	81	19	0.1	.89	1,006	79	21	0.0	.87
8	1,019	63	36	1.3	.82	1,006	64	35	0.9	.83

Note. SW = speaking and writing; exact % = the percentages of exact agreement between two ratings; adj. % = the percentages of adjacent ratings between the two raters; dis. % = the percentages of ratings with a discrepancy of 2 or more score points.

Table 29. Interrater Agreement and Reliability of Writing for Forms SW1 and SW2

Item	Form SW1–*N*	Form SW1–exact %	Form SW1–adj. %	Form SW1–dis. %	Form SW1–interrater	Form SW2–*N*	Form SW2–exact %	Form SW2–adj. %	Form SW2–dis. %	Form SW2–interrater
4	1,008	91	9	0.5	.92	998	84	14	1.9	.84
5	1,021	89	11	0.6	.88	1,007	86	13	0.6	.85
6	1,023	87	12	0.9	.85	1,007	86	14	0.7	.85
7	1,017	74	25	0.5	.80	1,005	80	20	0.3	.84
8	1,002	73	26	0.9	.76	989	73	26	0.7	.77
9	1,011	70	29	1.2	.82	989	63	35	1.3	.78

Note. SW = speaking and writing; exact % = the percentages of exact agreement between two ratings; adj. % = the percentages of adjacent ratings between the two raters; dis. % = the percentages of ratings with a discrepancy of 2 or more score points.

Conclusion

The redesigned TOEIC Bridge tests were launched in June 2019. Before the official launch, a field study, with two parallel forms for listening and reading and two for speaking and writing, was administered in April 2018 to evaluate the statistical properties of the redesigned TOEIC Bridge tests. Test takers from six countries (Japan, Korea, Taiwan, Colombia, Brazil, and Mexico) participated the field study. The statistical analysis in this report focused on the test takers from Japan, Korea, and Taiwan, who comprised 83% of the field study sample.

Overall, the reliabilities of the listening and reading tests in the field study were relatively higher than the average reliabilities of the original TOEIC Bridge Listening and Reading tests. Although on average the field study forms were harder (listening) or easier (reading) than the original TOEIC Bridge test, their average difficulty was very close to the operational historical

ranges in both measures. Therefore, we expect that the overall difficulty of the redesigned TOEIC Bridge Listening and Reading tests will not differ much from that of the original TOEIC Bridge Listening and Reading test. The overall item discrimination of both listening and reading of the field study was relatively higher than the original TOEIC Bridge test, with no speediness issues identified.

Because speaking and writing are not part of the original TOEIC Bridge test, it is not possible to compare the statistics of the redesigned TOEIC Bridge tests to those of the original TOEIC Bridge test. The difficulty of items varied across different item types for both the speaking and writing tests. Overall, these results indicate that item types in both speaking and writing tests can have different levels of item difficulty. In speaking, reliability—measured by coefficient alpha and stratified alpha—were quite comparable on both forms (over .80). In writing, although stratified alpha was higher than the coefficient alpha, reliabilities were lower than speaking on both forms (over .70). The interrater agreement rates were found to be reasonably high for both tests.

All in all, these findings not only helped to inform the final decisions regarding the final reporting scales of the redesigned TOEIC Bridge tests, but also allowed test developers make appropriate adjustments to the test design. Given that the findings of this study were based on a field study sample, which may have not been fully representative of the operational test-taking population, additional analyses will be conducted once sufficient operational data are gathered after the redesigned TOEIC Bridge tests are officially launched.

References

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge, UK: Press Syndicate of the University of Cambridge.

Everson, P., Duke, T., Garcia Gomez, P., Carter Grissom, E., Park, E., & Schmidgall, J. (2019). *Development of the redesigned TOEIC Bridge® tests* (Research Memorandum No. RM-19-10). Princeton, NJ: Educational Testing Service.

Liu, J., & Costanzo, K. (2013). The relationship among *TOEIC®* listening, reading, speaking, and writing skills. In D. E. Powers (Ed.), *The research foundation for the TOEIC® tests: A compendium of studies* (Vol. II, pp. 2.1–2.25). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., & Yin, C. (2012). Evidence-centered design in language testing. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 208–222). New York, NY: Routledge.

Notes

[1] The CEFR describes a progression of language proficiency in listening, reading, speaking, and writing on a six-level scale clustered in three bands: A1–A2 (basic user), B1–B2 (independent user), and C1–C2 (proficient user).