Research Memorandum
ETS RM–17-03

Statistical Analyses for the Expanded *TOEIC*® Speaking Test

Yanxuan Qu

Jaime Cid

Eric Chan

Yan Huo

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 Expanded *TOEIC*® Speaking Test

Yanxuan Qu, Jaime Cid, Eric Chan, and Yan Huo
Educational Testing Service, Princeton, New Jersey

December 2017

Corresponding author: Y. Qu, E-mail: yqu@ets.org

Suggested citation: Qu, Y., Cid, J., Chan, E., & Huo, Y. (2017). *Statistical analyses for the expanded* TOEIC® *Speaking test* (Research Memorandum No. RM-17-03). Princeton, NJ: Educational Testing Service.

Action Editor: Donald Powers

Reviewers: Sooyeon Kim and Rick Morgan

Abstract

Testing programs should periodically review their assessments to ensure that their test items or tasks are well aligned with real-world activities. For this reason, to better support communicative language learning and to discourage the use of memorization and other test-taking strategies, Educational Testing Service (ETS) expanded the existing format of some items of the *TOEIC*® Speaking test in May 2015. It is important to ensure that the new expanded item formats are comparable to the existing formats. In this paper, we report the results of a pilot study conducted in November 2013 to evaluate the comparability of items with new and existing formats in terms of difficulty, score consistency, and overall test reliability. We also summarize the operational trends observed after the implementation of the expanded item formats. The results of the pilot study suggest that even though modifications to existing item formats had a slight effect on the difficulty of items, as some items were more difficult and others were less difficult, the effects observed were within the range of variation typically observed across different forms of the test. Further monitoring of the difficulties of the new item formats based on operational testing results also indicates that items with the new formats have performed similarly to items with existing formats. This report shows that the expansion in the TOEIC Speaking item formats did not have any significant undesirable effects on item difficulty or test score reliability, indicating that with the new, more authentic tasks, the TOEIC Speaking test scores remain consistent and reliable.

Key words: ANCOVA, covariate, item difficulty, reliability, weighted kappa

Since the introduction in 2006 of the *TOEIC*® Speaking and Writing tests, the TOEIC program has periodically evaluated the test content specifications to ensure that they continue to meet the needs of test takers and test users. To better foster communicative language learning and to discourage the use of memorization and test-taking strategies, Educational Testing Service (ETS) expanded the existing format of some items of the TOEIC Speaking test in May 2015. Specifically, additional formats were added to four of the existing speaking items (Items 4, 5, 6, and 10). Items 4, 5, and 6 are often called an item set because they share the same item stem. The appendix describes the existing formats and the new formats. The purpose of the expansion was not to replace existing formats but rather to supplement them with new alternative formats. More details about the process that ETS followed to develop the expanded item formats were provided by Park and Bredlau (2014).

To ensure that these modifications would not significantly alter the difficulty of the items, a pilot study was conducted in November 2013. The purpose of the pilot study was to evaluate the comparability of existing formats with new formats in terms of difficulty and to determine if forms with the new formats had adequate reliability. In this paper, we summarize the analyses and results of the pilot study and the monitoring of the performance of the new formats in operational administrations.

The TOEIC Speaking Test

The first TOEIC Speaking test was launched in December 2006. It was designed to measure test takers' ability to communicate in spoken English in the context of daily life and the global workplace. The test has 11 items. Items 1 and 2 are each scored on two dimensions: pronunciation and intonation. Each dimension has a score scale from 0 to 3. Items 3 to 9 are rated on a scale of 0 to 3. Items 10 and 11 are rated on a scale of 0 to 5. Raw scores on each item are weighted when calculating the total test score (Qu, Liu, & Chan, 2013). The reported scaled scores range from 0 to 200 in increments of 10.

Pilot Forms

Three test forms (A, B, and C) were used in the pilot study (Table 1). Form A was selected as the base form. This existing TOEIC Speaking-only form received relatively low exposure (i.e., only a small number of test takers have taken this form). Items 4, 5, 6, and 10 in

Forms B and C used the new formats but differed in terms of content. The other seven items were common across the three forms.

Table 1. Outline of the Three Forms for the Pilot Study

Item type	Form A	Form B	Form C
Read a text aloud	Item 1 and Item 2	Same items as in Form A	Same items as in Form A
Describe a picture	Item 3	Same item as in Form A	Same item as in Form A
Answer 3 questions using information provided	Item 4, Item 5, Item 6 with existing format	New item formats	Same item format as in Form B, but different content
Answer 3 questions using information provided	Item 7, Item 8, Item 9	Same items as in Form A	Same items as in Form A
Propose a solution	Item 10 with existing format	New item formats	Same item format as in Form B, but different content
State an opinion	Item 11	Same item as in Form A	Same item as in Form A

Note. Items in Form A were all in existing format.

Data Collection

Data for the pilot study were collected from multiple test administrations that took place in October and November 2013 in Korea and Taiwan. Each test taker was asked to answer a background questionnaire before taking the TOEIC Speaking test. All forms were administered according to the same test administration procedures in place for operational administrations of the TOEIC Speaking test. Test takers who had previously taken Form A were not part of the study, and no test takers took more than one form. When recruiting the pilot samples, efforts were made to represent the current test-taking population in terms of demographic characteristics. As a result, given its representation in operational samples, more than 80% of test takers came from Korea. All of the test takers’ responses were scored by two certified, trained, and calibrated TOEIC Speaking raters (Everson & Hines, 2010).

Tables 2 and 3 present summaries of the total number of test takers by gender, form, and country. Tables 4 and 5 display the sample sizes of our analysis sample by gender, form, and country. In the analysis sample, test takers with a score of zero on any of the speaking items were screened out from the total sample (except when calculating reliability indices).

Table 2. Sample Size of the Full Data Set by Gender and Form, Korea

Gender	Form A (%) *n* = 319	Form B (%) *n* = 377	Form C (%) *n* = 296	Total
Female	172 (57.9)	263 (73.5)	194 (69.8)	629 (67.4)
Male	125 (42.1)	95 (26.5)	84 (30.2)	304 (32.6)
Total	297	358	278	933

Note. Percentages of male and female test takers within each country are provided in parentheses.

Table 3. Sample Size of the Full Data Set by Gender and Form, Taiwan

Gender	Form A (%) *n* = 319	Form B (%) *n* = 377	Form C (%) *n* = 296	Total
Female	16 (72.7)	17 (89.5)	11 (61.1)	44 (74.6)
Male	6 (27.3)	2 (10.5)	7 (38.9)	15 (25.4)
Total	22	19	18	59

Note. Percentages of male and female test takers within each country are provided in parentheses.

Table 4. Sample Size of the Analysis Sample by Gender and Form, Korea

Gender	Form A (%) *n* = 278 (87)	Form B (%) *n* = 322 (85)	Form C (%) *n* = 274 (93)	Total
Female	150 (58.4)	226 (74.6)	177 (69.1)	553 (67.8)
Male	107 (41.6)	77 (25.4)	79 (30.9)	263 (32.2)
Total	257	303	256	816

Note. Column headings show percentages of test takers remaining in the analysis sample after data screening in parentheses. Percentages of male and female test takers within each country are provided in parentheses in the data cells.

Table 5. Sample Size of the Analysis Sample by Gender and Form, Taiwan

Gender	Form A (%) *n* = 278 (87)	Form B (%) *n* = 322 (85)	Form C (%) *n* = 274 (93)	Total
Female	15 (71.4)	17 (89.5)	11 (61.1)	43 (74.1)
Male	6 (28.6)	2 (10.5)	7 (38.9)	15 (25.9)
Total	21	19	18	58

Note. Column headings show percentages of test takers remaining in the analysis sample after data screening in parentheses. Percentages of male and female test takers within each country are provided in parentheses in the data cells.

Statistical Analyses

Difficulty

To compare the performance of the new and existing formats in terms of difficulty, the following statistics were calculated for each form and compared across forms administered during the pilot study.

1. Standardized score mean difference across test forms. Standardized mean differences among the pilot groups for Forms A, B, and C were calculated based on weighted raw

scores on all common items. These score mean differences on common items reflect differences in group ability.

2. Average item scores. Items 4, 5, and 6 are based on the same item stimuli and are usually considered an item set. Because the difficulty level of these three items is controlled at set level instead of item level when assembling forms, average item score is provided only at the set level (denoted as "Avg_456" in Table 6). Similarly, average item score is only provided for Items 7, 8, and 9 as a set (denoted as "Avg_789" in Table 6). In general, items with higher average scores are easier than items with lower average scores.
3. Mean and standard deviation of scaled scores.
4. Adjusted item score means by ANCOVA after controlling group differences on common items and gender. In the ANCOVA model, test form was the treatment factor, weighted common test score was the covariate, and gender was a controlling factor.

Reliability

The reliability of the items and forms with the new formats was evaluated by examining the following statistics.

1. Pearson correlations between item raw scores and weighted total raw scores. An item with a high correlation with the total test score is a good item that can discriminate high-ability test takers from low-ability test takers and can contribute more to the overall test reliability.
2. Interrater agreement measures for evaluating scoring reliability. Because the TOEIC Speaking test is evaluated by raters, it is important to evaluate the consistency of the ratings given by the two raters. These measures included percentage of exact agreement between two raters' ratings for each item: weighted kappa (Haberman, 2012) and correlations between two raters' scores. The two scores (pronunciation and intonation) of Item 1 and Item 2 were considered independently when calculating these item level statistics.

3. Total test reliability coefficient. Reliability refers to the extent to which the assessment scores are consistent over repeated administrations of the same test or alternate forms. Stratified coefficient alpha (Rajaratnam, Cronbach, & Gleser, 1965) was used as a reliability estimate in this study. A high coefficient alpha reliability is desired because it indicates that scores obtained remain consistent over repeated administrations of the same or alternate forms of the test.

Results of the Pilot Study

Evaluation of Difficulty Level

Table 6 presents average scores for each item, item set, common items, and total test. The standardized mean difference (SMD) in weighted raw scores on all common items was 0.17 for Forms B and A and 0.30 for Forms C and A, which indicates that the three pilot groups were not equivalent in terms of ability. Therefore, it was necessary to control group ability differences before making comparisons on the difficulty levels between the new and existing formats. For this reason, an ANCOVA (Howell, 2002) was conducted to take into account group ability differences when comparing item difficulty across forms. The following section introduces the ANCOVA analyses and results.

Table 6. Comparison of Scores Across Three Forms

Scores	Form A Mean (*SD*)	Form B Mean (*SD*)	Form C Mean (*SD*)	*SMD* for Forms B and A	*SMD* for Forms C and A
Weighted score on common items	142.55 (26.22)	147.24 (27.55)	150.65 (27.67)	0.17	0.30
Weighted score on Items 4,5,6, and 10	76.91 (17.87)	77.03 (18.02)	82.01 (16.17)	0.01	0.30
Scaled score	125.5 (29.51)	128.79 (30.75)	134.93 (30.11)		
P1	2.28 (0.55)	2.39 (0.55)	2.39 (0.55)		
P2	2.36 (0.55)	2.44 (0.52)	2.50 (0.55)		
I1	2.23 (0.55)	2.28 (0.57)	2.33 (0.58)		
I2	2.17 (0.51)	2.28 (0.51)	2.30 (0.51)		
3	2.37 (0.68)	2.32 (0.63)	2.46 (0.62)		
Avg_456	2.05 (.60)	2.33 (0.47)	2.58 (0.42)		
Avg_789	1.96 (0.43)	1.97 (0.44)	2.03 (0.41)		
10	3.08 (0.76)	2.80 (0.84)	2.89 (0.78)		
11	2.98 (0.83)	3.16 (0.87)	3.22 (0.92)		

Note. P1 = Item 1 Pronunciation; P2 = Item 2 Pronunciation; I1 = Item 1 Intonation; I2 = Item 2 Intonation; SMD standardized mean difference.

Controlling Group Ability Differences by ANCOVA Analysis

A further examination of the background data revealed that the three forms had similar background distributions except on gender. Table 7 indicates that Forms B and C had higher percentages of female test takers than Form A. In addition, female test takers performed better than male test takers (see Table 8) on all items. Therefore, gender was selected as a controlling factor, and the weighted raw scores on common items was treated as the covariate in the ANCOVA model.

Two ANCOVA models were run. Both models had form as the treatment variable, gender as a controlling factor, and weighted score on common items as the covariate. The first model used the average score of Item Set 456 as the dependent variable, and the second model used the raw score of Item 10 as the dependent variable. Adjusted group means on the average score of Item Set 456 and the raw score of Item 10 were obtained in each ANCOVA analysis. Table 9 shows the results of the two ANCOVA models.

Table 7. Average Item Score by Gender, Female

Item	Form A (*N* = 165)	Form B (*N* = 243)	Form C (*N* = 188)
Pronunciation	2.42	2.48	2.53
Intonation	2.30	2.35	2.40
3	2.50	2.40	2.55
Avg_456	2.16	2.37	2.65
Avg_789	2.01	2.00	2.08
10	3.19	2.87	2.99
11	3.16	3.26	3.38

Note. Pronunciation is the average of Item 1 and Item 2 pronunciations; intonation is the average of Item 1 and Item 2 intonations.

Table 8. Average Item Score by Gender, Male

Item	Form A (*N* = 113)	Form B (*N* = 79)	Form C (*N* = 86)
Pronunciation	2.17	2.23	2.25
Intonation	2.04	2.08	2.14
3	2.17	2.08	2.24
Avg_456	1.89	2.22	2.44
Avg_789	1.89	1.86	1.91
10	2.92	2.59	2.65
11	2.73	2.86	2.88

Note. Pronunciation is the average of Item 1 and Item 2 pronunciations; intonation is the average of Item 1 and Item 2 intonations.

Table 9. Summary Results for ANCOVA (*N* = 874)

Model	R^2	Form A[a]	Form B[a]	Form C[a]	Significance test for mean difference
Avg_456 = Form + Weighted Common Test Scores + Gender	.52	2.10 (2.05)	2.32 (2.33)	2.53 (2.58)	*p* < .001 for all pairs
Item 10 = Form + Weighted Common Test Scores + Gender	.47	3.17 (3.08)	2.81 (2.80)	2.82 (2.89)	*p* < .0001 for A vs. B and A vs. C

[a] Adjusted means with unadjusted means in parentheses.

To decide how meaningful these differences in the mean scores were for Item Set 456 and Item 10, we compared the score variations for Item Set 456 and Item 10 in the pilot forms against the score variations of Item Set 456 and Item 10 across all forms administered from January 2012 through November 2013 (see Table 10). As Table 10 shows, the covariate adjusted average scores for Item Set 456 on the pilot forms varied from 2.10 to 2.53, which is within three standard deviations of the mean of Item 456 in operational administrations. The difficulty difference between existing and new formats can be considered reasonable on Item Set 456. For Item 10, the average score on the pilot forms varied from 2.81 to 3.17. The average score of Item 10 on the existing Form A (3.17) was more than three standard deviations above the operational mean. The average scores for Forms B and C were well within historical averages.

Table 10. Adjusted Item Scores Compared to Operational Scores

Item/Item set	Form A	Form B	Form C	Mean[a] (*SD*)
Avg_456	2.10	2.32	2.53	2.30 (0.16)
Item 10	3.17	2.81	2.82	2.70 (0.15)

[a] Operational data based on 46 forms administered from January 2012 to November 2013.

Evaluation of Reliability

Table 11 presents correlations of Items 4, 5, 6, and 10 with the weighted total score of the seven common items. The correlations in Forms B and C were similar to those in Form A, and all of the correlation coefficients were larger than 0.30. Items with the new formats performed as well as items with the existing formats in discriminating high- and low-ability test takers.

The total test coefficient alpha reliability information in Table 12 shows that all forms had adequately high reliability. The reliability of the forms with the new item formats (Forms B and C) were higher than the reliability of the form with the existing item formats (Form A).

Tables 13, 14, and 15 present the interrater agreement measures, including percentages of exact agreement, weighted kappas, and correlations between the two ratings. All three pilot forms had adequate to high rater agreement coefficients, indicating that the overall scoring reliability was adequately high for forms with both new and existing formats.

Table 11. Correlations Between New Format Item and Weighted Common Test Scores

Form	N	Item 4	Item 5	Item 6	Item 10
A	278	0.47	0.48	0.64	0.65
B	322	0.46	0.50	0.60	0.67
C	274	0.45	0.49	0.60	0.71

Table 12. Total Test Coefficient Alpha Reliability

Form	Reliability
A	0.87
B	0.91
C	0.91

Table 13. Interrater Reliability: Exact Agreement

Item	Form A	Form B	Form C
I1: Item 1 Intonation	62	66	64
P1: Item 1 Pronunciation	67	72	71
I2: Item 2 Intonation	71	76	72
P2: Item 2 Pronunciation	66	70	71
3	66	64	67
4	76	75	72
5	70	67	80
6	67	63	72
7	88	86	90
8	73	69	72
9	89	85	86
10	66	74	69
11	70	68	70

Sample sizes for Form A ranged from 317 to 319 across different items, from 375 to 377 for Form B, and from 294 to 296 for Form C.

Table 14. Interrater Reliability: Weighted Kappa

Item	Form A	Form B	Form C
I1: Item 1 Intonation	0.40	0.50	0.47
P1: Item 1 Pronunciation	0.49	0.57	0.57
I2: Item 2 Intonation	0.49	0.59	0.53
P2: Item 2 Pronunciation	0.41	0.52	0.56
3	0.64	0.64	0.67
4	0.82	0.76	0.60
5	0.73	0.64	0.65
6	0.67	0.63	0.70
7	0.86	0.86	0.84
8	0.84	0.82	0.79
9	0.83	0.75	0.78
10	0.71	0.87	0.80
11	0.77	0.84	0.86

Note. Sample sizes for Form A ranged from 317 to 319 across different items, from 375 to 377 for Form B, and from 294 to 296 for Form C.

Table 15. Interrater Reliability: Correlation

Item	Form A	Form B	Form C
I1: Item 1 Intonation	0.40	0.50	0.47
P1: Item 1 Pronunciation	0.50	0.57	0.57
I2: Item 2 Intonation	0.49	0.60	0.53
P2: Item 2 Pronunciation	0.42	0.52	0.56
3	0.64	0.64	0.67
4	0.82	0.76	0.60
5	0.73	0.65	0.66
6	0.67	0.64	0.70
7	0.86	0.86	0.85
8	0.84	0.82	0.79
9	0.83	0.75	0.78
10	0.71	0.87	0.80
11	0.77	0.84	0.86

Note. Sample sizes for Form A ranged from 317 to 319 across different items, from 375 to 377 for Form B, and from 294 to 296 for Form C.

Difficulties of the Expanded Item Formats in Operational Administrations

The new formats of Item Set 456 and Item 10 have been used in operational practice along with the existing formats since May 2015. To monitor the difficulties of the new formats, the scores of Item Set 456 and Item 10 with the new formats were compared to the scores of Item Set 456 and Item 10 with the existing formats. In this paper, item scores were compared separately for two types of operational forms: SP (secured program) and SSP (special secured program) forms. Although SP and SSP forms have the same test specifications, SP forms are administered once a month in Korea and other Asian countries, whereas SSP forms are administered only in Korea.

Figure 1 shows plots of the scores of Item Set 456 in all SP forms administered from February 2014 through June 2016 in Asian countries. The red diamonds note the scores for Item Set 456 with the new formats. Forms administered before May 2015 were included to provide a reference for the comparison between new and existing item formats. In total, 58 SP forms were administered from February 2014 through June 2016, including the 11 forms with new formats for Item Set 456 administered after May 2015. The average score of Item Set 456 with the new formats ranged from 2.06 to 2.51, with a mean of 2.33. Similarly, the average score of Item Set 456 with the existing formats ranged from 2.02 to 2.60, with a mean of 2.31. Figure 1 shows that Item Set 456 with new formats was similar in difficulty level to those with existing formats.

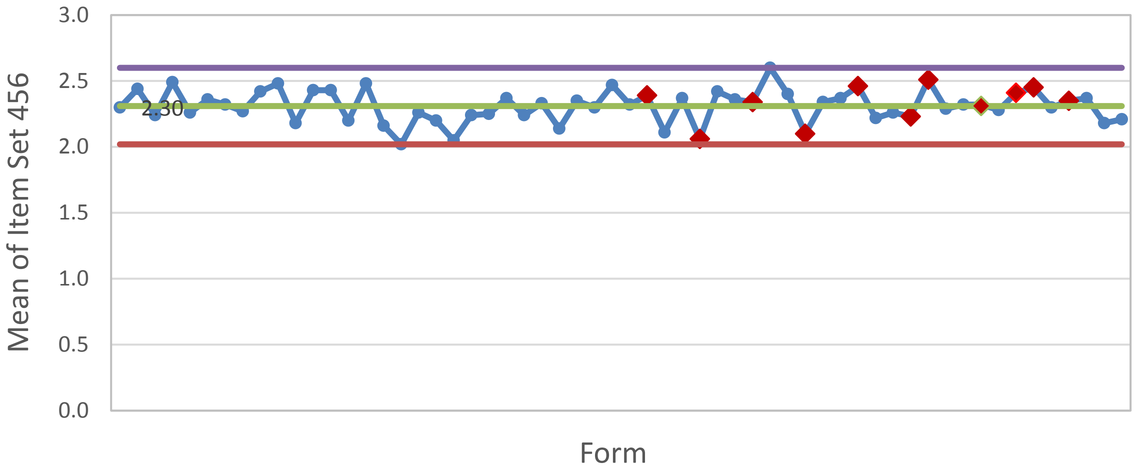

Figure 1. Comparison of the means of Item Set 456 from February 2014–June 2016 in operational administrations (SP forms). *Note.* Red diamonds denote new formats.

Figure 2 shows plots of the scores of Item 10 in all SP forms administered from February 2014 through June 2016 in Asian countries. After May 2015, nine forms used the new formats for Item 10. The scores of Item 10 with the new formats ranged from 2.40 to 2.84 with a mean of 2.62. The scores of Item 10 with the existing formats ranged from 2.32 to 2.95 with a mean of 2.62. Figure 2 shows that Item 10 with the new formats also had a difficulty level similar to those with the existing formats.

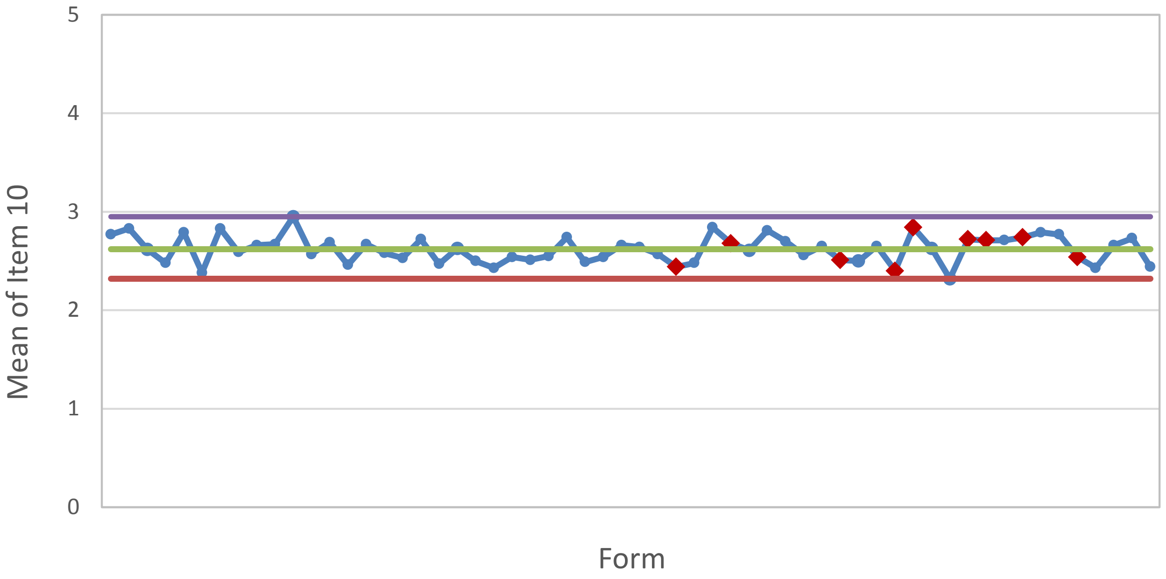

Figure 2. Comparison of the means of Item 10 from February 2014–June 2016 in operational administrations (SP forms). *Note.* Red diamonds denote new formats.

Figure 3 shows plots of the scores of Item Set 456 for all SSP forms administered in Korea from February 2014 through June 2016. Nine forms (out of 358) had Item Set 456 in the new formats. The scores of Item Set 456 with the new formats ranged from 2.31 to 2.52 with a mean of 2.41. This is within the range of the scores of Item Set 456 with the existing formats (1.81 to 2.61, with a mean of 2.32).

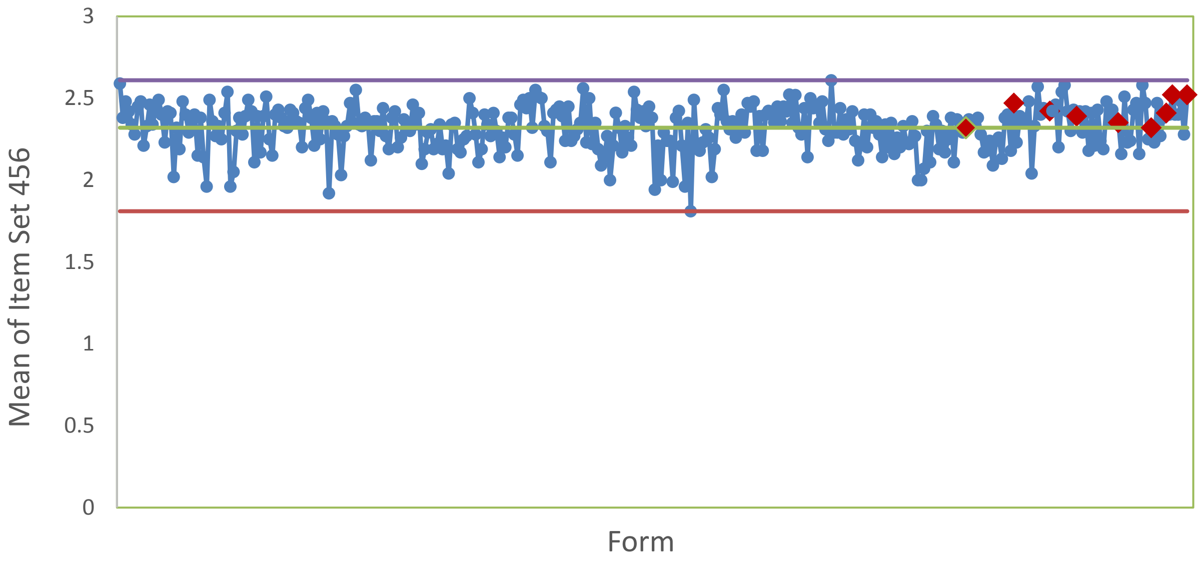

Figure 3. Comparison of the means of Item Set 456 from February 2014–June 2016 in operational administrations (SSP forms). ***Note.*** **Red diamonds denote new formats.**

Figure 4 shows the scores of Item 10 for all SSP forms administered from February 2014 through June 2016. Four forms (out of 358) had Item 10 in the new formats. The scores of Item 10 with the new formats ranged from 2.57 to 2.79 with a mean of 2.72. The scores of Item 10 with the existing formats ranged from 2.10 to 2.98 with a mean of 2.60. Therefore, in both SP and SSP administrations, the score means were similar between new and existing formats for both Item Set 456 and Item 10.

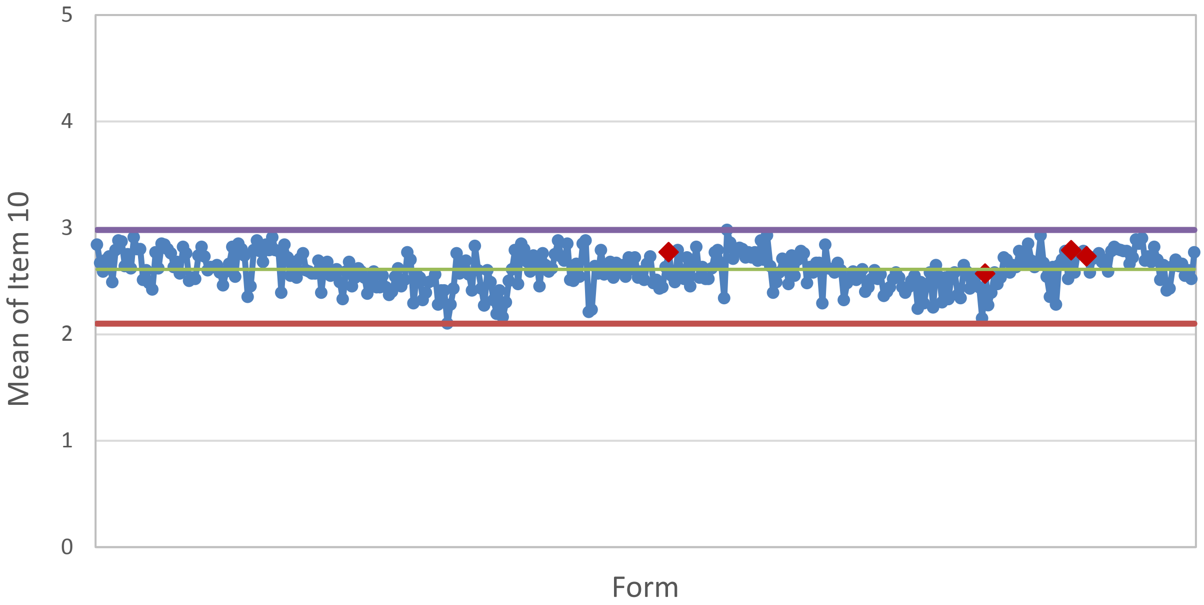

Figure 4. Comparison of the means of Item 10 from February 2014–June 2016 in operational administrations (SSP forms). ***Note.*** **Red diamonds denote new formats.**

Reliabilities for Tests with Existing and New Formats in Operational Administrations

Reliability estimates averaged across operational forms administered from February 2014 through June 2016 are provided in Table 16 for the new and existing formats separately. The SP and SSP forms are included in the comparison. Both interrater reliability estimates (interrater correlation) at the item level and the internal consistency (coefficient alpha) reliability estimate at the test level are similar between forms with the new formats and forms with the existing formats.

Table 16. Existing Formats From February 2014 Through June 2016

Format	Item 4	Item 5	Item 6	Item 10	Internal consistency reliability estimate for total test
New formats	0.67 (*n* = 20)	0.61 (*n* = 20)	0.48 (*n* = 20)	0.71 (*n* = 13)	0.81
Existing formats	0.62 (*n* = 383)	0.62 (*n* = 383)	0.52 (*n* = 383)	0.67 (*n* = 383)	0.80

Concluding Remarks

In this report, we describe an evaluation of whether expanded item formats of the TOEIC Speaking test impacted item difficulty and test reliability. As noted at the outset, these expanded formats were intended to expand coverage in a way that was thought to foster language learning and to discourage the use of undesirable test-taking strategies. The results of this study suggest that modifications to existing item formats had a slight effect on the difficulty of items, as some items were more difficult and others were less difficult. However, the effects observed were basically within the range of variation typically observed across alternate forms of the test. Further monitoring of the difficulties of the new item formats in operational practice also indicates that items with the new formats have performed similarly to items with existing formats in operational practice. In operational administrations, forms with the new formats have also had reliability estimates similar to those with the existing formats. In conclusion, efforts to improve selected TOEIC Speaking items so as to better foster communicative language learning appears not to have had any significant undesirable effects on item difficulty or test score reliability.

References

Everson, P., & Hines, S. (2010). *How ETS scores the TOEIC Speaking and Writing test responses* (TOIEC Compendium No. TC-10-08). Princeton, NJ: Educational Testing Service. https://www.ets.org/Media/Research/pdf/TC-10-08.pdf

Haberman, S. (2012). Measures of agreement [ETS presentation handout]. Princeton, NJ: Educational Testing Service.

Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Duxbury/ Thomson Learning.

Park, E., & Bredlau, E. (2014). *Expanding the question variants of the TOEIC Speaking Test* [White paper]. Retrieved from https://www.ets.org/Media/Research/pdf/Expanding%20the%20Question%20Formats%20of%20the%20TOEIC%20Speaking%20Test.pdf

Qu, Y., Liu, J. & Chan, E. (2013). *Changing the current weights of the TOEIC Speaking test*. (Internal memorandum). Princeton, NJ: Educational Testing Service.

Rajaratnam, N., Cronbach, L. J., & Gleser G. C. (1965). Generalizability of stratified-parallel tests. *Psychometrika*, *30,* 39–56.

Appendix. Expanded Question Formats

Expanded Question Formats for Items 4–6

Task: Respond to questions	Existing formats	New formats
Familiar topics and personal experiences	x	
Begin with “Imagine that . . .”	x	
Talk on the telephone with a marketing firm	x	
Hear and read the questions	x	
Two 15-second and one 30-second response	x	
No preparation time	x	
Current rubric and scoring rules	x	
Familiar topics and personal experiences		x
Begin with “Imagine that . . .”		x
Talk on the telephone with an employee, colleague, friend, etc.		x
Hear and read the questions		x
Two 15-second and one 30-second response		x
No preparation time		x
Current rubric and scoring rules		x

Note. Bolded parts note the difference between the two formats.

Expanded Question Formats for Item 10

Task: Propose a solution	Existing formats	New formats
Single speaker	x	
Recognize the problem	x	
Propose a way of dealing with the problem	x	
Listen to the question, no reading	x	
60-second response	x	
30 seconds of preparation time	x	
Current rubric and scoring rules	x	
Two people speaking at a meeting		x
Recognize the problem		x
Propose a way of dealing with the problem		x
Listen to the question, no reading		x
60-second response		x
30 seconds of preparation time		x
Current rubric and scoring rules		x

Note. Bolded parts are the difference between the two formats.