ETS TOEIC®

Know English. Know Success.

COMPENDIUM STUDY

Alternate Forms Test-Retest Reliability and Test Score Changes for the TOEIC® Speaking and Writing Tests

Chi-wen Liao and Yanxuan Qu

January 2010

This study evaluates the alternate forms test-retest reliability of the TOEIC® Speaking and Writing tests. These tests are constructed-response measures developed at the end of 2006 to measure non-native English speakers' productive skills in speaking and writing English. The resulting reliability estimates provide ETS and test users with evidence about the reliability of the reported scores.

There are two categories of test score reliability that differ in data collection and computation. One is the internal consistency reliability estimated by a coefficient alpha index. This method involves collecting data from a single administration and computing the reliability estimate as a ratio of estimated true variance to total variance. This is the most popular method used in operational practice as data can be directly obtained from operational administrations and no special arrangement needs to be made. The second category of reliability involves collecting data from a group of examinees who repeat the test multiple times, by taking either the same test form or alternate forms. The reliability coefficient is estimated from the correlation of the examinees' test scores from multiple test administrations. When the same test form is used in the multiple administrations, the estimated reliability is the test-retest reliability. When different test forms are used in different administrations, the estimated reliability is the alternate form test-retest reliability. The estimated reliability coefficients of test-retest and alternate form test-retest reliability are usually lower than the internal consistency reliability estimate. The reason why is that reliability of the two types of test-retest is affected by more measurement errors than the internal consistency reliability. Specifically, the test-retest reliability involves errors due to changes in individual's performance and the rater's consistency over time, plus all errors associated with the internal consistency method. The alternate form test-retest reliability is affected by errors due to content sampling in different test forms, plus all errors associated with the test-retest reliability.

The average internal consistency alpha estimated from each operational administration for the TOEIC Speaking test is about .82 with a standard error of measurement of about 1.5 raw score points or 15 scale score points. No internal consistency estimate is available for the TOEIC Writing test because of the limited number of writing tasks used in the test. This study was designed to evaluate the alternate form test-retest reliability for both the TOEIC Speaking and Writing tests using data from examinees who repeated the tests multiple times. Specifically, the following research questions were asked and investigated:

1. What are the alternate form test-retest reliability estimates for the TOEIC Speaking and Writing tests between the first and second administrations of a test (the first time an examinee took a test and the second), the second and third administrations, and the third and fourth administrations? Because many examinees took the tests more than once, the reliability of scores from different administrations might vary.

2. Did examinee scores tend to increase or decrease when examinees repeated the tests, and what were the score changes?

3. What are test-retest reliability estimates for the TOEIC Speaking and Writing tests between the first and second administrations for five groups of examinees who repeated a test within the following intervals: 1–30 days, 31–60 days, 61–90 days, 91–180 days, and 181–365 days? The purpose of this question is to examine whether the time interval between the first and the second administration impacts the reliability estimates.

4. Were the five groups of examinees equally able, and were their score increases from the first to the second administrations the same?

5. What are the distributions of the change in examinees' score levels from the first to the second administration? Are the distributions the same across the five groups of examinees?

Data Collection

The TOEIC Speaking test consists of 11 questions, with 6 different types of tasks, and the TOEIC Writing test consists of 8 questions, with 3 different types of tasks. Both tests are administered by computer, and examinees can choose to take both tests at the same time in one administration or just take one of the tests. Test items are weighted to derive the total raw scores, which range from 0 to 24 for the TOEIC Speaking test and 0 to 26 for the TOEIC Writing test. The weighted total raw scores are then transformed to scaled scores and score levels for reporting purposes (Liao & Reeder, 2008). (See the ranges for raw scores, scaled scores, and score levels in the appendix.)

For this study, data on examinees who took a test more than once were collected from 16,867 examinees who took the TOEIC Speaking test and 6,199 examinees who took the TOEIC Writing test multiple times from December 2006 to December 2008. These examinees repeated the tests in the public-administration program, or secure program (SP), the institutional-sponsored program (IP); or both. An examinee could take either the SP or IP administration first. In this study, the first time that an examinee took a test is labeled as the first administration, the second time as the second administration, and so on. The fourth administration was the most recent one relative to the data collection endpoint. Examinees repeated the test at different time intervals. For example, the time between the first and second administrations could be one month, but the second administration and the third administration could be more than one month apart. Some examinees took each test only twice, and others took a test up to four times.

In terms of all examinees in the study (both those examinees who took a test multiple times and those who took each test only once), 94,768 examinees took the TOEIC Speaking test and 39,897 examinees took the TOEIC Writing test in the operational SP and IP administrations from December 2006 to December 2008. The means and standard deviations of these examinees' speaking and writing scores are shown in the appendix. Slightly less than one fifth (18%) of examinees who took the TOEIC Speaking test were repeaters, and about one sixth (16%) of examinees who took that TOEIC Writing test were repeaters. The majority of the repeaters took the TOEIC Speaking test (70%) and the TOEIC Writing test (79%) only two times each. Small percentages took the TOEIC Speaking test (11%) and the TOEIC Writing test (6%) up to three times each. Tables 1 and 2 list the distribution of the examinees repeating the test.

TABLE 1

Distribution of Number of Times Examinees Took the TOEIC Speaking Test

Number of times taking the TOEIC Speaking test	Frequency	%	Cumulative frequency
2	11738	70	11738
3	3206	19	14944
4	1923	11	16867

TABLE 2

Distribution of Number of Times Examinees Took the TOEIC Writing Test

Number of times taking the TOEIC Writing test	Frequency	%	Cumulative frequency
2	4870	79	4870
3	950	15	5820
4	379	6	6199

The median time intervals between an examinee's first and the second administration were 63 days for the TOEIC Speaking test and 141 days for the TOEIC Writing test. The more an examinee repeated the test, the shorter the time interval between administrations. The time interval between the third and the fourth administrations dropped to 28 days for both the TOEIC Speaking and Writing tests. The time intervals between adjacent administrations are shown in Tables 3 and 4.

TABLE 3

Average Time Interval in Days Between the First and Second, Second and Third, and Third and Fourth Administrations for the TOEIC Speaking Test

	N	Mean	SD	Min	Max	Median
1st–2nd	16867	108	100	1	693	63
2nd–3rd	5129	78	81	1	559	45
3rd–4th	1923	48	56	1	469	28

TABLE 4

Average Time Interval in Days Between the First and Second, Second and Third, and Third and Fourth Administrations for the TOEIC Writing Test

	N	Mean	SD	Min	Max	Median
1st–2nd	6199	144	107	3	693	141
2nd–3rd	1329	116	113	2	500	56
3rd–4th	379	72	85	2	399	28

Method

In this study, the reliability of scores when two test scores are employed is estimated using the Pearson correlation coefficient. A Pearson correlation coefficient measures the direction (sign) and the strength of the linear relationship between two random variables. It has a range from -1 to +1. A correlation with a positive sign indicates the two variables are linearly related to each other in a positive way. For example, those who score high on one test form tend to score high on the second test form. A correlation with a negative sign indicates the two variables are linearly related in an opposite way. Usually a correlation coefficient larger than .8 is considered to be high, and a correlation coefficient less than .3 is considered to be low. The magnitude of the Pearson correlation coefficient (i.e., the alternate form test-retest reliability estimate in this study) depends on two important factors: restriction of range and combining groups (Allen & Yen, 2002). Restriction of range happens when the test score range is narrow, or when the examinee group is homogeneous in ability so that the test scores do not vary across the whole score range. Restriction of range will reduce the magnitude of correlation coefficients. Combining different groups of examinees can either reduce or increase the magnitude of the reliability estimates.

The alternate form test-retest correlation and the internal consistency reliability are both measures of the consistency of scoring. The test-retest correlations computed in this study will be impacted by whether those choosing to repeat the tests are representative of the total population of examinees. If the repeaters are a more homogenous group, then the restriction of range will result in lower reliability estimates for the repeater sample than for a more representative sample of test takers. The average internal consistency reliability estimated from operational administrations for the TOEIC Speaking test is .82 with a standard error of measurement (SEM) of 15 scaled score points. Given the SEM of 15, the internal consistency estimate of the standard error of the score differences (SED) will be around 21 scaled score points. The standard deviations of the scaled score differences were estimated in the study, which will provide another measure of score difference consistency. Unlike the score reliability estimates, the SEDs are not directly influenced by restriction of range.

Results

Test-Retest Reliability

The estimated reliability coefficients for raw, scaled, and score levels between the first and second administrations, the second and third administrations, and the third and fourth administrations are shown in Tables 5 and 6. For the TOEIC Speaking test, the estimates range from .79 to .80 for both raw and scaled scores and .74 to .77 for the score levels. The lower reliability for the score levels are because score levels are composed of ranges of scaled scores. The reliability coefficient estimates for the TOEIC Speaking test remain very similar among the three paired administrations: first and second, second and third, and third and fourth. For the first and second administrations and the second and third administrations, the reliability coefficients estimated for the raw and scaled scores for the TOEIC Writing test ranged from .81 to .83 and from .80 to .82 for the score levels. However, certain of the reliability coefficient estimates for the raw, scaled and score levels for the TOEIC Writing test were much lower than others, ranging from .68–.69, for the third and fourth administrations; this is probably because the group in that administration period was more homogeneous in terms of score variation. See the standard deviations in Tables 7–8 below.

TABLE 5

Test-Retest Reliability for the TOEIC Speaking Test

	N	Raw scores	Scaled scores	Score levels
1st–2nd	16867	0.80	0.79	0.76
2nd–3rd	5129	0.80	0.80	0.77
3rd–4th	1923	0.79	0.79	0.74

TABLE 6

Test-Retest Reliability for the TOEIC Writing Test

	N	Raw scores	Scaled scores	Score levels
1st–2nd	6199	0.83	0.82	0.82
2nd–3rd	1329	0.81	0.81	0.80
3rd–4th	379	0.69	0.69	0.68

Change of Scores Across Multiple Administrations

The raw and scaled scores and score levels for repeaters increased when examinees repeated the TOEIC Speaking and Writing tests. Tables 7 and 8 present the means and standard deviations across the three pairs of administrations for the tests. The increase in the score level means was less than one level. The scaled score increase is summarized in the last column of the tables. For the TOEIC Speaking test, the increase ranged from 5.9 scaled score points between the first and second administrations to 3.2 scaled score points between the third and fourth administrations. For the TOEIC Writing test, the increase ranged from 6.9 scaled score points between the first and second administrations to 1.9 scaled score points between the third and fourth administrations. The larger increase tended to occur between the first and the second administrations.

The standard deviation of scaled score differences ranged from 21 to 22 for the TOEIC Speaking test and from 23 to 25 for the TOEIC Writing test. These numbers could be interpreted as the SEDs, and as mentioned earlier in the text, the SEDs are not directly influenced by restriction of range. When comparing the performance of repeaters to the combined SP and IP examinee population, it was found that repeaters (scaled score M = 115 from the first administration) are slightly less able than the operational population (scaled score M = 122) for the TOEIC Speaking test, but the score variation is comparable and both have a standard deviation of 35. For the TOEIC Writing test, the repeaters are also less able (scaled score M = 134 from the first administration) than the operational population (scaled score M = 152), but the repeaters' score variation (SD = 40) is larger than that of the operational population (SD = 37).

TABLE 7

Means and Standard Deviations for Examinees in the First and Second, Second and Third, and Third and Fourth Administrations for the TOEIC Speaking Test

Admins	*N*	Admins	Raw score mean (SD)	Scaled score mean (SD)	Score level mean (SD)	Difference in scaled score mean (SD)
1st–2nd	16867	1st	13.6 (3.7)	115.0 (34.7)	5.0 (1.4)	5.9 (21.9)
		2nd	14.2 (3.5)	120.8 (33.4)	5.2 (1.3)	
2nd–3rd	5129	2nd	13.7 (3.5)	116.7 (32.9)	5.0 (1.3)	3.8 (20.5)
		3rd	14.2 (3.4)	120.5 (31.8)	5.2 (1.3)	
3rd–4th	1923	3rd	13.6 (3.3)	115.5 (31.3)	5.0 (1.2)	3.2 (20.5)
		4th	13.9 (3.3)	118.6 (31.3)	5.1 (1.2)	

TABLE 8

Means and Standard Deviations for Examinees in the First and Second, Second and Third, and Third and Fourth Administrations for the TOEIC Writing Test

Admins	*N*	Admins	Raw score Mean (SD)	Scaled score mean (SD)	Score level mean (SD)	Difference in scaled score mean (SD)
1st–2nd	6,199	1st	16.3 (4.5)	133.7 (40.3)	6.3 (1.6)	6.9 (23.9)
		2nd	17.0 (4.5)	140.6 (39.9)	6.6 (1.6)	
2nd–3rd	1,329	2nd	17.4 (4.1)	144.2 (36.4)	6.7 (1.5)	5.2 (22.5)
		3rd	18.0 (4.1)	149.3 (36.2)	6.9 (1.4)	
3rd–4th	379	3rd	18.6 (3.5)	154.4 (30.6)	7.1 (1.1)	1.9 (24.5)
		4th	18.8 (3.6)	156.3 (32.0)	7.2 (1.2)	

Reliability Estimates for Examinees Repeated in Different Time Intervals

To investigate the third question, we clustered the time intervals (as seen in Tables 3 and 4) into five groups: 1–30 days, 31–60 days, 61–90 days, 91–180 days, and 181–365 days. Only a very small percentage of examinees waited more than 1 year before repeating the test; therefore, these examinees' data were not used in the reliability estimation for the five groups. The results are shown in Tables 9 and 10. The reliability estimates for the five groups within the first and second administration are lower than what is estimated for the total group (see Tables 5 and 6) except for the 181–365 days

group. Excluding the 181–365 days group, the reliability estimates for raw and scaled scores for the TOEIC Speaking test ranged from .75 to .78, and from .71 to .74 for score levels. These reliability estimates were lower than the reliability estimate for the total group because the scores of these groups were more homogenous than the scores for the total group. The positive news here is that reliability estimates seem to hold up even when the interval between administrations increases.

Also, the larger the time interval between the first and second administration, the higher the reliability. Again, this is closely related to the increase in score variation in different groups (see the standard deviations in Tables 11–12). For the TOEIC Writing test, the patterns are very similar. Excluding the 181–365 days group, the reliability estimates for raw and scaled scores for the TOEIC Writing test ranged from .59 to .73, and .56 to .70 for score levels. Also, the larger the time interval between the first and second administration, the higher the reliability. The reliability estimates for the 181–365 days group for the TOEIC Speaking and Writing tests are higher than those of the total group, because this group is more heterogeneous than the total group.

TABLE 9

Test-Retest Reliability for Five Groups of Examinees for the TOEIC Speaking Test

1st–2nd	*N*	Raw scores	Scaled scores	Score levels
1–30 days	4940	0.75	0.75	0.71
31–60 days	3017	0.75	0.75	0.72
61–90 days	1988	0.76	0.76	0.73
91–180 days	2730	0.78	0.77	0.74
181–365 days	3839	0.83	0.83	0.80

TABLE 10

Test-Retest Reliability for Five Groups of Examinees for the TOEIC Writing Test

1st–2nd	*N*	Raw scores	Scaled scores	Score levels
1–30 days	1352	0.63	0.62	0.56
31–60 days	887	0.60	0.59	0.56
61–90 days	422	0.70	0.69	0.68
91–180 days	813	0.73	0.72	0.70
181–365 days	2541	0.85	0.85	0.84

The Performance of the Five Groups of Examinees

The analysis results with regard to the fourth research question showed that examinees who waited longer to repeat the test were the least able group. For example, the raw score mean for the TOEIC Speaking test was 14.5 for the 1–30 days group, but 12.0 for 181–365 days group. The mean raw score for the TOEIC Writing test was 18.5 for the 1–30 days group, but 13.7 for the 181–365 days group. Tables 11 and 12 present the means and standard deviations from the five groups of examinees who took the first and second administrations of the TOEIC Speaking and Writing tests.

The average scores for examinees repeating the test increased regardless of how long they waited to take the second test. The average score increase for the total group was 5.9 for the TOEIC Speaking test. The increases for the five groups ranged from 4.7 to 7.5. The group that waited the shortest time to repeat the test (1–30 days) is also the most able group in terms of their test scores, and their average score increase of 4.8 points is comparable to the group that waited the longest (181–365 days), which is also the least able group. The SEDs (i.e., the standard deviations of score increases) for these two groups are 21 vs. 23. The middle three groups of examinees have an average score increase of 7.0–7.5, and their SED is about 22.

For the TOEIC Writing test, there is a clear pattern that a longer time interval between test administrations corresponds to a greater score increase except for those examinees who had a more than 6-months interval between test administrations. The smallest score increase was 5.5 scaled score points for the 1–30 days group, which is also the most able group. The largest score increase was 8.6 scaled score points for the 91–180 days group.

TABLE 11

Means and Standard Deviations for Five Groups of Examinees for the TOEIC Speaking Test

1st–2nd administrations	*N*	Raw score mean (SD)	Scaled score mean (SD)	Score level mean (SD)	Difference in scaled score mean (SD)
1–30 days	4940	14.5 (3.2)	123.5 (30.1)	5.3 (1.2)	4.8 (21.0)
		15.0 (3.1)	128.3 (29.2)	5.5 (1.1)	
31–60 days	3017	14.0 (3.4)	118.5 (31.6)	5.1 (1.3)	7.5 (21.6)
		14.7 (3.2)	126.0 (29.9)	5.4 (1.2)	
61–90 days	1988	13.5 (3.5)	113.8 (32.9)	4.9 (1.3)	7.0 (22.4)
		14.2 (3.3)	120.8 (31.5)	5.2 (1.3)	
91–180 days	2730	13.7 (3.5)	116.5 (32.7)	5.0 (1.3)	7.0 (21.7)
		14.5 (3.3)	123.5 (31.2)	5.3 (1.2)	
181–365 days	3839	12.0 (4.2)	100.0 (39.7)	4.4 (1.6)	4.7 (23.0)
		12.5 (4.1)	104.7 (38.2)	4.6 (1.6)	

TABLE 12

Means and Standard Deviations for Five Groups of Examinees for the TOEIC Writing Test

1st–2nd administrations	*N*	Raw score mean (SD)	Scaled score mean (SD)	Score level mean (SD)	Difference in scaled score mean (SD)
1–30 days	1352	18.5(3.0)	153.9(26.9)	7.1(1.0)	5.5(23.7)
		19.2(3.1)	159.4(27.4)	7.3(1.0)	
31–60 days	887	18.3(3.0)	151.6(26.4)	7.0(0.9)	5.7(23.9)
		18.9(2.9)	157.2(26.3)	7.2(0.9)	
61–90 days	422	18.0(3.0)	149.3(27.0)	6.9(1.0)	7.1(22.3)
		18.8(3.3)	156.4(29.6)	7.2(1.1)	
91–180 days	813	17.1(3.7)	140.7(32.5)	6.6(1.2)	8.6(24.4)
		18.0(3.8)	149.3(33.0)	6.9(1.3)	
181–365 days	2541	13.7(4.9)	111.5(43.9)	5.4(1.9)	7.3(24.0)
		14.6(4.8)	118.8(43.3)	5.7(1.8)	

Distribution of Change in Score Levels Across the Five Groups of Examinees

The percentage of examinees whose score levels changed when they took the TOEIC Speaking and Writing tests a second time is shown in Tables 13 and 14. For the TOEIC Speaking test, 44% of all the examinees showed no change in score levels from the first to second administration. This percentage remains fairly consistent across the groups, except for the 180–365 days group, where the percentage was a little lower. The same pattern was found for the TOEIC Writing test, with again 44% of the examinees having no change in score level.

For both tests, higher percentages of the repeaters had increases of 1 or 2 levels than had decreases of 1 or 2 levels. A very small percentage (less than 0.5%) of the repeaters had score level decreases of 3 or more levels. Roughly 1% of the repeaters had score level increases of 3 or more levels for the TOEIC Speaking test, and slightly more than 1% for the TOEIC Writing test.

TABLE 13

Percentage of Examinees Whose Speaking Score Level Changed the Second Time They Took the TOEIC Speaking Test

Difference in score levels between 1st and 2nd administrations	All examinees	1-30 days	31-60 days	61-90 days	91-180 days	181-365 days
-6	0.0	0.0	0.0	0.0	0.0	0.1
-5	0.0	0.0	0.0	0.0	0.0	0.0
-4	0.0	0.0	0.0	0.0	0.1	0.1
-3	0.1	0.1	0.0	0.3	0.0	0.2
-2	2.2	2.1	1.6	2.1	1.8	3.1
-1	16.5	16.5	15.0	14.6	14.8	19.0
0	44.1	46.8	43.9	43.9	45.0	40.8
1	29.2	28.3	31.3	30.5	29.1	28.3
2	6.9	5.5	7.1	7.4	8.2	7.4
3	0.8	0.6	1.0	1.1	1.0	0.8
4	0.1	0.0	0.1	0.2	0.0	0.2
5	0.0	0.0	0.0	0.1	0.0	0.0
6	0.0	0.0	0.0	0.0	0.0	0.0
N	16,867	4,940 (29%)	3,017 (18%)	1,988 (12%)	2,730 (16%)	3,839 (23%)

TABLE 14

Percentage of Examinees Whose Speaking Score Level Changed the Second Time They Took the TOEIC Writing Test

Difference in score levels between 1st and 2nd administrations	All examinees	1-30 days	31-60 days	61-90 days	91-180 days	181-365 days
-7	0.0	0.0	0.1	0.0	0.0	0.0
-6	0.0	0.1	0.0	0.0	0.1	0.0
-5	0.1	0.2	0.0	0.0	0.0	0.0
-4	0.1	0.1	0.0	0.0	0.1	0.0
-3	0.2	0.2	0.1	0.0	0.4	0.3
-2	2.2	2.1	1.7	2.1	1.4	2.7
-1	16.3	16.1	17.6	14.7	15.1	16.5
0	44.1	46.5	45.2	46.2	45.4	41.8
1	29.3	29.2	30.4	30.8	30.6	28.2
2	6.5	5.1	4.6	5.7	5.7	8.4
3	1.0	0.4	0.2	0.5	1.0	1.6
4	0.2	0.1	0.0	0.0	0.0	0.3
5	0.1	0.0	0.0	0.0	0.1	0.1
6	0.1	0.0	0.0	0.0	0.1	0.1
N	6,199	1,352 (22%)	887 (14%)	422 (7%)	813 (13%)	2,541 (41%)

Conclusions

This study evaluated the alternate form test-retest reliability for the TOEIC Speaking and Writing tests. Reliability coefficients were estimated for examinees who repeated the test up to four times. Detailed analyses were also applied to groups of examinees who waited different lengths of time before repeating the test and to the resulting changes in their scaled and score levels. The major findings from this study are summarized below.

1. The repeaters in this study are slightly less able (as measured by their test scores) than the total test population who took the TOEIC Speaking and Writing tests from December 2006 to December 2008. The score variation in the TOEIC Speaking test for repeaters is comparable to that of the total test population, while the score variation in the TOEIC Writing test for the repeaters is slightly larger than that of the total population.

2. The test-retest reliability estimate is .80 for the raw score for the TOEIC Speaking test and .83 for the raw score for the TOEIC Writing test from the first and second administrations. Comparable estimates for the scaled scores are .79 for the TOEIC Speaking test and .82 for the TOEIC Writing test. The reliability of the score levels is .76 for the TOEIC Speaking test and .82 for the TOEIC Writing test. The reliability of the score levels is lower because score levels are based on ranges of scaled scores. The reliability estimates from the second administration to the third and from the third administration to the fourth, in general, remain similar to the reliability estimates from the first administration to the second (unless there was a large reduction in score variation).

3. For the TOEIC Writing test, more able examinees tended to repeat the test more times. For the TOEIC Speaking test, examinees who repeated the test a second or third time appeared to have similar abilities. However, the largest score gain is observed from the first administration to the second for both the TOEIC Speaking test (5.9 scaled score) and the TOEIC Writing test (6.9 scaled score). The more times an examinee repeated a test, the less the gain that was achieved. The average scaled score increases from the third administration to the fourth are 3.2 and 1.9 for the TOEIC Speaking test and the TOEIC Writing test, respectively.

4. The examinees who waited longer to repeat the test for the second time tended to be those who were less able. This pattern is the same for the TOEIC Speaking test and the TOEIC Writing test. The reliability estimates were slightly different among different groups of examinees who waited different lengths of time to repeat the test. The magnitude of the reliability estimates are closely related to the score variation in the groups. Also, the reliability estimates for the groups are lower than that of the total group. Again, this is related to the homogeneity of the groups.

5. When examinees repeat the test for the second time, about half of the time their score level will remain unchanged, about one third of the time their score level will increase 1 score level, and about one sixth (17%) of the time their score will decrease 1 score level. The likelihood that their score level will decrease by 3 or more score levels is slight.

In conclusion, the alternate form test-retest reliability coefficients estimated for the TOEIC Speaking and Writing tests from this study are considered reasonably high and acceptable for their intended purposes. Factors such as the reduction of score range and homogeneity of groups, however, should always be considered when trying to interpret and use the data. The authors would like to point out that all of the reliability coefficients calculated in this study were based on the data that also included outliers, as we considered that every second score an examinee achieved was a reasonable score regardless of how much the second score differed from the first one. If outliers were removed, all of the reliability coefficients would have been slightly higher.

References

Allen, M., & Yen, W. (2002). *Introduction to measurement theory*. Prospect Heights, IL: Waveland Press, Inc.

Liao, C. (2009). The *TOEIC Speaking and Writing reliability*. Unpublished manuscript.

Liao, C., & Reeder, J. (2008). *Scale and levels for the TOEIC Speaking and Writing tests*. Unpublished manuscript.

Appendix

Means and Standard Deviations for Examinees From the SP and IP Administrations of the TOEIC Speaking Test During December 2006 and December 2008

	N	Raw scores	Scaled scores	Score levels
SP	46757	15.7 (3.4)	130 (31.7)	5.6 (1.2)
IP	48011	13.9 (3.8)	114 (35.6)	5.0 (1.4)
SP + IP	94768	14.8 (3.7)	122 (34.7)	5.3 (1.4)

Means and Standard Deviations for Examinees From the SP and IP Administrations of the TOEIC Writing Test during December 2006 and December 2008

	N	Raw scores	Scaled scores	Score levels
SP	21225	18.7 (3.7)	152 (32.6)	7.0 (1.2)
IP	18672	16.7 (4.5)	134 (39.9)	6.3 (1.6)
SP + IP	39897	17.8 (4.2)	144 (37.3)	6.7 (1.5)

Possible Score Ranges for the TOEIC Speaking and Writing Tests

	Raw scores	Scaled scores	Score levels
Speaking	0 – 24	0 – 200 (increment of 10)	1 – 8
Writing	0 – 26	0 – 200 (increment of 10)	1 – 9