ETS TOEIC®

Know English. Know Success.

COMPENDIUM STUDY

The Relationships of Test Scores Measured by the TOEIC® Listening and Reading Test and TOEIC® Speaking and Writing Tests

Chi-wen Liao, Yanxuan Qu and Rick Morgan

January 2010

The TOEIC® Listening and Reading test measures a non-native speaker's listening and reading skills in English as these skills are used in the workplace. The test was developed about 30 years ago as a measure of receptive language skills and has been widely accepted and used worldwide. The TOEIC Speaking and Writing tests were launched in December 2006, and they measure speaking and writing skills in English. The purpose of this study is to examine the correlations of scores among the TOEIC Listening and Reading test and the TOEIC Speaking and Writing tests to determine whether the tests measure four separate language skills.

Research Questions

Specifically, this study seeks to answer the following four research questions:

1. What are the observed correlations among scores that measure listening, reading, speaking, and writing skills?

2. What are the disattentuated correlations among scores for listening, reading, speaking, and writing skills?

3. Are the correlations between scores for skill groups (listening and speaking, listening and writing, reading and speaking, and reading and writing, and speaking and writing) consistent regardless of the lapse of time between scores from administrations of the TOEIC Listening and Reading test and the TOEIC Speaking and Writing tests?

4. 4. What are the relationships between the combined scores for listening and reading skills and each of the score levels for speaking and writing skills?

Methodology

Data Collection

Data were based on 12,105 examinees in Korea(79%), and Japan(20%), and Taiwan (1%) who took the TOEIC Speaking and Writing tests multiple times from December 2006 to December 2008 and the TOEIC Listening and Reading test in early 2009. The TOEIC Listening and Reading test is composed of two sections, one for listening and one for reading, and examinees take both sections during the same administration, so an examinee's listening and reading scores are obtained on the same date. The TOEIC Speaking and Writing tests have three types of administrations: an administration for both the TOEIC Speaking and Writing tests, an administration for the TOEIC Speaking test only, or an administration for the TOEIC Writing test only. In this dataset, each examinee had only one record for listening and reading scores and multiple records for speaking and/or writing scores, which could be from different administrations of the TOEIC Speaking and Writing tests on different dates. To investigate the correlations among the four skills, scores from administrations that had the shortest time interval between two administrations were selected. Appendix A details the procedures used to select the final dataset. In the final dataset, each examinee had scores for at least two of the four skills. The majority of examinees (62%) had scores for all the four skills. Of the 12,105 examinees, all took the TOEIC Listening and Reading test, 12,099 took the TOEIC Speaking test, 7,483 took the TOEIC Writing test, and 7,477 took both the TOEIC Speaking and Writing tests.

The TOEIC Listening and Reading test has two scaled scores, one each for listening and reading. The listening and reading scores are reported on separate scales, but both scales range from 5 to 495 in increments of 5. The TOEIC Speaking and Writing tests report both scaled scores and score levels. The speaking and writing scales range from 0 to 200 in increments of 10. The speaking score levels

range from 1 to 8. The writing score levels range from 1 to 9. The speaking and writing score levels were created based on examinees' performance patterns, and they are directly related to the scaled scores (Liao & Reeder, 2008). Appendix B presents the range of scaled scores for each level.

Statistical Measures

The Pearson product moment correlation coefficient, or Pearson correlation coefficient, measures the degree and direction of the linear relationship between two measures. It has a range of -1 to 1, with 0 indicating no linear relationship, -1 indicating a perfect negative linear relationship, and 1 indicating a perfect positive linear relationship. A Pearson correlation coefficient with absolute value of .3 and below is usually considered a weak linear relationship, while a value of .8 and above is considered a strong linear relationship.

When determining the degree of the relationship between two constructs as measured by two tests, disattenuated correlations are often computed. The disattenuated correlation, which ranges from 0 to 1, adjusts for the random error of measurement associated with the test scores. If the size of the disattenuated correlation is 1, the two tests measure the same construct. If the constructs measured by the two tests are somewhat different, the correlation will be reduced by 1.

Results

Observed Correlations Among Listening, Reading, Speaking, and Writing Scores

Table 1 shows the Pearson correlation coefficients for the six pairs of test scores for three samples of examinees. The first sample consists of all examinees in the dataset. Not every examinee had all four scores, so the sample sizes differ depending on which two pairs of tests are used for each examinee (see the All Data column). The second sample consists of examinees from the first sample who had all four test scores. The third sample consists of examinees who took all four tests within one month. The means and standard deviations of the three samples are shown in Table 2. The smaller the group, the higher the mean score and the smaller the score variation.

The four test scores are all moderately correlated. The highest correlation is between listening and reading scores (ranging from .73 to .76 across the three samples). The lowest correlation is between reading and speaking scores (ranging from .56 to .57). Listening scores are correlated higher with speaking scores (.65 to .66) than with writing scores (.58 to .59). Reading scores are correlated slightly higher with writing scores (.59 to .61) than with speaking scores (.56 to .57). The pattern of these correlations across the three samples is very consistent. For each pair of tests, the less variable the group, the smaller the correlation. For example, the correlation for listening and reading scores dropped from .76 for the total group to .73 for the group of examinees who took all of the tests within one month. The associated standard deviations for both listening and reading scores were larger for the total group.

TABLE 1

Observed Correlations Among Listening, Reading, Speaking, and Writing Scores Across Three Samples

	All data	Examinees took all 4 tests (N = 7,477)	Examinees took all 4 tests within 1 month (N = 2,160)
Listening and reading			0.73
Listening and speaking	0.66 (N = 12,099)	0.65	0.66
Listening and writing	0.59 (N = 7,483)	0.59	0.58
Reading and speaking	0.57 (N = 12,099)	0.56	0.56
Reading and writing	0.61 (N = 7,483)	0.61	0.59
Speaking and writing	0.62 (N = 7,477)	0.62	0.62

TABLE 2

The Mean and Standard Deviations for Listening, Reading, Speaking and Writing Scores Across Three Samples

	All data	Examinees took all 4 tests	Examinees took all 4 tests within 1 month
Listening			
Mean	404	415	418
SD	74	69	67
N	12,105	7,477	2,160
Reading			
Mean	363	375	377
SD	81	78	77
N	12,105	7,477	2,160
Speaking			
Mean	130	134	138
SD	30	30	29
N	12,099	7,477	2,160
Writing			
Mean	150	150	154
SD	31	31	29
N	7,483	7,477	2,160

Disattenuated Correlations of Listening, Reading, Speaking and Writing Scores

This analysis used the All Data sample, and the results of disattenuated correlations are presented above the diagonal line in Table 3. (Numbers below the diagonal line are observed correlations.) In the calculation of disattenuated correlations, the reliability for listening and reading scores, estimated from operational administrations, is assumed to be .92. The reliability for speaking and writing scores, estimated using data from a test-retest reliability study (see Compendium Study 11), was .80 and .83, respectively. The disattenuated correlation, which adjusts for the error of measurements associated with the score variables, is larger than the associated observed score correlation (unless the reliability is 1). After adjustment, the sizes of disattenuated correlations among the four scores are larger than the observed correlations, but the disattenuated correlations still are not close to 1. These results suggest that the TOEIC Listening and Reading test and the TOEIC Speaking and Writing tests measure four

different aspects of English proficiency.

TABLE 3

Observed and Disattenuated Correlations among Listening, Reading, Speaking and Writing Scores

	Listening	Reading	Speaking	Writing
Listening	1	0.83	0.77	0.68
Reading	0.76	1	0.66	0.70
Speaking	0.66	0.57	1	0.76
Writing	0.59	0.61	0.62	1

Consistency of Correlations Across Time Intervals

To consider the possible effect of the time interval between test administrations on the size of the Pearson correlation coefficients, examinees were separated into six groups according to the time interval between each test that the examinees took (or the time interval between each set of scores examinees received, such as listening and speaking scores, reading and speaking scores, and so on1). The time intervals were 1–30 days, 31–60 days, 61–90 days, 91–180 days, 181–365 days and 366–730 days. The observed score correlations are shown in Table 4, while the means and standard deviations for the scores are shown in Table 5.

The correlations between listening and speaking scores are consistent across groups, ranging from .65 to .67, which is comparable to the All Data group. The exception is the 366–730 days group, which has a correlation of .72 for listening and speaking scores. The correlations between reading and speaking scores and listening and writing scores also appear to be consistent across the six groups, except for the 366–730 days group. The correlations of reading and speaking scores range from .55 to .56, with a correlation of .65 for the 366–730 days group. The correlations of listening and reading scores range from .57 to .61, with a correlation of .63 for the 366–730 days group. The correlations between reading and writing scores also appear to be relatively consistent across the six groups, ranging from .58 to .64, but a slightly larger difference exists between this range and a correlation of .61 for the 366–730 days group. As to be expected, larger correlations are generally associated with groups with larger variances in scores (see standard deviation in Table 5). The correlation between speaking and writing scores decreases as the time intervals increase. In five of the six groups, however, the number of examinees available is small (fewer than 30) for calculating the correlations between speaking and writing scores, which means the resulting low correlation coefficients must be viewed with some caution. The correlations among the scores in the six groups are moderate, which suggests that the tests are measuring different constructs.

1 Because examinees always received listening and reading scores on the same date, correlations between those scores were not investigated.

TABLE 4

Correlations Among Listening, Reading, Speaking and Writing Scores Across Groups

	All data	Within 30 days	31–60 days	61–90 days	91–180 days	181–365 days	366–730 days
Listening and speaking[a]							
Correlation	0.66	0.65	0.66	0.67	0.66	0.65	0.72
N	12,099	4,528	1,751	1,129	1,714	2,144	830
Reading and speaking[a]							
Correlation	0.57	0.55	0.55	0.57	0.58	0.56	0.65
N	12,099	4,528	1,751	1,129	1,714	2,144	830
Listening and writing[b]							
Correlation	0.59	0.58	0.60	0.61	0.61	0.57	0.63
N	7,483	2,183	1,289	733	1,154	1,482	630
Reading and writing[b]							
Correlation	0.61	0.58	0.61	0.64	0.64	0.58	0.64
N	7,483	2,183	1,289	733	1,154	1,482	630
Speaking and writing[c]							
Correlation	0.62	0.63	0.52	0.51	0.37	0.40	0.47
N	7,477	6,717	271	133	203	121	33

[a] Based on the sample where every examinee has listening and speaking scores. [b]Based on the sample where every examinee has listening and writing scores. [c]Based on the sample where every examinee has speaking and writing scores.

TABLE 5

Means and Standard Deviations of Listening, Reading, Speaking and Writing Scores Across Groups

Scores	All data	Within 30 days	31–60 days	61–90 days	91–180 days	181–365 days	366–730 days
Listening[a]	404 (74) *N* = 12,105	407 (70) *N* = 4,528	405 (71) *N* = 1,751	403 (78) *N* = 1,129	400 (75) *N* = 1,714	403 (75) *N* = 2,144	398 (83) *N* = 830
Reading[a]	363 (81) *N* = 12,105	364 (79) *N* = 4,528	363 (79) *N* = 1,751	361 (86) *N* = 1,129	359 (82) *N* = 1,714	365 (82) *N* = 2,144	359 (89) *N* = 830
Speaking[a]	130 (30) *N* = 12,099	133 (28) *N* = 4,528	130 (30) *N* = 1,751	130 (30) *N* = 1,129	129 (30) *N* = 1,714	128 (30) *N* = 2,144	127 (34) *N* = 830
Listening[b]	415 (69) *N* = 7,483	418 (67) *N* = 2,183	410 (67) *N* = 1,289	412 (76) *N* = 733	412 (70) *N* = 1,154	418 (66) *N* = 1,482	420 (70) *N* = 630
Reading[b]	375 (78) *N* = 7483	377 (76) *N* = 2183	369 (77) *N* = 1289	370 (84) *N* = 733	372 (79) *N* = 1154	380 (75) *N* = 1482	378 (82) *N* = 630
Writing[b]	150 (31) *N* = 7,483	154 (29) *N* = 6,717	149 (31) *N* = 1,289	148 (33) *N* = 733	148 (31) *N* = 1,154	149 (30) *N* = 1,482	146 (33) *N* = 630
Speaking[c]	134 (30) *N* = 7477	134 (31) *N* = 6717	134 (27) *N* = 271	139 (26) *N* = 133	133 (24) *N* = 203	134 (27) *N* = 121	149 (24) *N* = 33
Writing[c]	150 (31) *N* = 7,477	150 (31) *N* = 6,717	151 (25) *N* = 271	149 (30) *N* = 133	147 (28) *N* = 203	145 (32) *N* = 121	151 (23) *N* = 33

[a]Based on the sample where every examinee has listening and reading and speaking scores. [b]Based on the sample where every examinee has listening and reading and writing scores. [c]Based on the sample where every examinee has speaking and writing scores.

Relationships Between the Combined Listening and Reading Scores and Speaking and Writing Score Levels

Although the TOEIC Listening and Reading test reports separate listening and reading scaled scores, it has been the custom for users in the field to use the combined listening and reading scores when

evaluating English proficiency. Users are interested in knowing the relationship between the combined listening and reading scores and the speaking and writing reported score levels. They are also interested in knowing what the predicted speaking and writing score levels are based on the listening and reading combined scaled scores.

The observed (and disattenuated) correlations of the listening and reading combined scores with speaking and writing score levels were .64 (.75) and .62 (.70) respectively. Again, the correlations indicate that the TOEIC Listening and Reading test and the TOEIC Speaking and Writing tests are related, but do not assess identical skills. The internal consistency reliability of combined listening and reading scores used in calculating the disattenuated correlations was assumed to be .95, which was estimated from operational administrations. The test-retest reliability of speaking and writing score levels is .76 and .82, respectively (Liao & Qu, in press). Table 6 provides the means and standard deviations of listening and reading and speaking and writing scores and their correlations.

TABLE 6

Means, Standard Deviations, and Correlations for Listening and Reading Combined Scaled Scores and Speaking and Writing Score Levels

	Mean (SD)	Observed correlation	Disattenuated correlation	*N*
Listening and reading	766 (145)	0.64	0.75	12,099
Speaking	5.58 (1.2)			
Listening and reading	790 (137)	0.62	0.70	7,483
Writing	6.9 (1.1)			

The listening and reading combined scaled scores have many possible data points (i.e., 198 points) while the speaking or writing score level has only 8 or 9 data points. When predicting the speaking or writing score levels, it was considered to be more appropriate to group the listening and reading combined scaled scores into several score ranges (as can be seen in Tables 7 and 8). Each range contains about 50 combined listening and reading scaled score points, and 13 score ranges are formed. The listening and reading combined scaled scores below 400 are formed into only one range, as very few examinees scored below 400. The percentages of scoring within a particular speaking or writing score level based on an examinee's listening and reading combined scaled score range are shown in Tables 7 and 8. These percentages can be interpreted as probabilities of scoring a particular speaking or writing level given an examinee's listening and reading combined scores.

For each listening and reading score range, probabilities of scoring different speaking and writing levels are shown. The yellow-colored cell indicates the highest probability of an examinee receiving a score within a particular speaking or writing level. The light yellow cells in each listening and reading score range indicates the second and third highest probability of an examinee receiving a score within a particular speaking and writing level. No percentages less than 5% are shown in the tables.

TABLE 7

Percentage of Examinees Receiving Each TOEIC Speaking Score Level for Ranges of Combined TOEIC Listening and Reading Scaled Scores

Combined score	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6	Level 7	Level 8	N
< 400	9	23	21	31	10				217
400–445		8	23	46	15	7			179
450–495		6	19	42	24	7			274
500–545			9	46	29	13			413
550–595			7	32	36	22			539
600–645			6	28	38	25			769
650–695				23	33	35	5		947
700–745				17	33	40	7		1,296
750–795				11	27	51	10		1,525
800–845				7	22	54	16		1,779
850–895					15	55	25		1,860
900–945					8	47	38	5	1,537
950–990						27	47	22	764

Note. ***Total*** *N* ***= 12,099.***

TABLE 8

Percentages of Receiving Each TOEIC Writing Score Level for Ranges of Combined TOEIC Listening and Reading Scaled Scores

Combined scores	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6	Level 7	Level 8	Level 9	N
< 400	12	11	19	26	17	9				110
400–445				30	29	20	11			70
450–495				23	22	33	17			109
500–545				5	21	48	21			165
550–595					15	46	30	5		250
600–645					12	43	35			396
650–695					6	40	44	6		511
700–745					7	30	47	13		767
750–795						25	55	16		932
800–845						16	56	22		1,130
850–895						11	49	34	5	1,288
900–945						5	38	47	9	1,113
950–990							19	51	28	642

Note. ***Total*** *N* ***= 7,483.***

The most likely speaking levels as predicted from the listening and reading combined scaled ranges are clustered around levels 4 to 7. The most likely speaking level corresponding to a listening and reading combined scaled score from 650 to 945 is 6. The listening and reading combined scaled score did not spread out the predicted speaking levels widely, and this is expected as the correlation between the combined listening and reading and the speaking score levels is only .64. The same situation was observed in the predicted writing levels. The most likely writing levels as predicted from the listening and reading combined scaled ranges are clustered around levels 4 to 8. The most likely writing score level corresponding to a combined listening and reading scaled score from 650 to 895 is 7. This is again because the correlation between the combined listening and reading score and the writing score levels is only .62. These results reinforce the need to assess writing and speaking skills instead of relying on predictions from listening and reading scores.

Conclusion

Both the observed and disattenuated Pearson correlation coefficients were estimated for the four English language skills as measured by the TOEIC Listening and Reading test and the TOEIC Speaking and Writing tests. The test scores were found to be moderately correlated with one another, with the highest correlation between listening and reading scores (.73–.7), and the lowest between reading and speaking scores (.56–.57). The correlations among the four skills were, in general, consistent across the groups of examinees with varying numbers of days between tests. The results from the disattenuated correlations confirm that there are four separate language skills measured by the TOEIC tests. It is natural that different language skills are correlated with each other to a certain degree; however, each test measures distinct aspects of English language proficiency that cannot be adequately assessed by the other tests. Examinees should take all of the TOEIC tests in order to gain a full understanding of the complete spectrum of their language proficiency skills.

References

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

Appendix A

In the original dataset we received from the field, each examinee had only one record of listening and reading scores, but multiple scores of speaking and/or writing scores from different administrations. The number of speaking or writing scores for each individual varies and they could range from 1 to 4. To investigate the correlations among the four skills, we hoped to select one speaking and/or writing (if available) score for each individuals so that his or her four (or three) scores were taken within the shortest time intervals for any combinations of two tests (e.g., the TOEIC Listening and Reading test and the TOEIC Speaking test, the TOEIC Listening and Reading test and the TOEIC Writing test, and the TOEIC Speaking test and the TOEIC Writing test). The following selection rules were decided.

1. To determine which speaking score to use for an examinee to correlate with the examinee's listening and reading scores, the date that the examinee took the TOEIC Listening and Reading test was first compared with each of the dates that he/she took the speaking test. The date for the TOEIC Speaking test that was closest to the date for the TOEIC Listening and Reading test was selected, along with the examinee's speaking score for that date.

2. To determine which writing score to use for an examinee to correlate with the examinee's listening and reading scores, the date that the examinee took the TOEIC Listening and Reading test was compared with each of the dates that he/she took the TOEIC Writing test. The date for the TOEIC Writing test that was closest to the datt for the TOEIC Listening and Reading test was selected, along with the examinee's writing score for that date.

While the selection rules described above found the administration dates for each of the TOEIC Speaking and Writing tests that were closest to the administration date for the TOEIC Listening and Reading test, these administration dates may not have been the closest for each individual between his/her speaking and writing dates. Some analyses were done to check whether this was an issue to be concerned about.

Altogether, the final data selected contained 12,105 examinees that took the TOEIC Listening and Reading test. Among these examinees, 12,099 took the TOEIC Speaking test, and 7,483 took the TOEIC Writing test, and 7,477 took both the TOIEC Speaking and Writing tests. Analyses were performed to check the distributions of the time intervals between the TOEIC Listening and Reading test and the TOEIC Speaking test dates (see Table A1), TOEIC Listening and Reading test and the TOEIC Writing test (see Table A), and the TOEIC Speaking test and the TOEIC Writing test dates (see Table A3). The dates were classified into seven subgroups.

The analyses results shown in Table 3A indicated that the aforementioned selection rules happened to select 86% of speaking and writing scores that were also the closes too each. Indeed, they were from the same administration date. The rest of the 14% of examinees have taken speaking and writing from different dates; however, we can't conclude whether those were or were not closest to each other. Since we have already obtained the 86% closest speaking and writing scores, we believe that the aforementioned selection rules are appropriate.

TABLE A1

Distribution of Groups of Examinees Who Took the TOEIC Listening and Reading Test and the TOEIC Speaking Test

Time interval between tests	Frequency	Percent	Cumulative frequency
1–30 days	4,528	37	4,528
31–60 days	1,751	15	6,279
61–90 days	1,129	9	7,408
91–180 days	1,700	14	9,108
181–365 days	2,142	18	11,250
366–730 days	832	7	12,082
Beyond 730 days	17	0.1	12,099

TABLE A2

Distribution of Groups of Examinees Who Took the TOEIC Listening and Reading Test and the TOEIC Writing Test

Time interval between tests	Frequency	Percent	Cumulative frequency
1–30 days	2,183	29	2,183
31–60 days	1,289	17	3,472
61–90 days	733	10	4,205
91–180 days	1,152	15	5,357
181–365 days	1,481	20	6,838
366–730 days	631	8	7,469
Beyond 730 days	14	0.2	7,483

TABLE A3

Distribution of Groups of Examinees Who Took the TOEIC Speaking test and the TOEIC Writing Test

Time interval between tests	Frequency	Percent	Cumulative frequency
Same day	6,344	86	6,344
1–30 days	373	5	6,717
31–60 days	271	4	6,988
61–90 days	133	2	7,121
91–180 days	202	3	7,323
181–365 days	121	2	7,444
366–730 days	33	0.4	7,477

Appendix B

Speaking level	Speaking scaled score
1	0–30
2	40–50
3	60–70
4	80–100
5	110–120
6	130–150
7	160–180
8	190–200

Writing level	Writing scaled score
1	0–30
2	40
3	50–60
4	70–80
5	90–100
6	110–130
7	140–160
8	170–190
9	200