ETS TOEIC

Know English. Know Success.

COMPENDIUM STUDY

Statistical Analyses for the TOEIC® Speaking and Writing Pilot Study

Chi-wen Liao and Youhua Wei

January 2010

The TOEIC® Speaking and Writing tests were developed by ETS during 2005 and 2006. A pilot study conducted in December 2006 evaluated the statistical properties of the tests in order to confirm whether the planned design for the tests was achieved. The results of the study also helped fine-tune the final design of the tests before they were launched for operational use. The statistical analyses conducted for this study included determining the difficulty of the tests, establishing the correlation among different parts of the tests, and examining test score reliability and inter-rater reliability. This report documents the results of these statistical analyses.

Data Collection

Four test forms were developed for the pilot study and given to three samples of examinees in target populations in Japan, Korea and France. The target population for the TOEIC Speaking and Writing tests, the same population as for the TOEIC Listening and Reading test, is composed of adults whose first language is not English and who are interested in using and/or will use English at work. Because the TOEIC Speaking and Writing tests are direct measures of productive skills, lower ability examinees may encounter difficulty in taking the tests. Because of this, only those examinees who had scored at least a combined total score of 400[1] on the TOEIC Listening and Reading test were invited to participate in the study of the pilot forms for the TOEIC Speaking and Writing tests.

In each country, examinees were randomly assigned to take one of four forms during December 2006. The number of examinees from each country who took each form is detailed in Table 1. Smaller samples took Forms C and D, and their results were not as reliable as the results for Forms A and B, each of which had a sample size of more than 1,000. Only results for Forms A and B are presented in this report.

Test Design

The test specifications in terms of the number of questions, preparation and response times for each item or item set, evaluation criteria, and the rubrics employed for the TOEIC Speaking and Writing tests are summarized in Tables 2 and 3, respectively. See Compendium Study 7.1 for details on how the tests were based on an evidence-centered design (ECD) approach.

The TOEIC Speaking test consists of 10 questions, which are categorized into six task-types with two types for each of three ECD claims. The tasks measuring the first claim require test-takers to simply read text aloud and describe a picture. The tasks measuring the second claim are two 3-item sets of questions, and examinees are required to respond to these questions based on personal experience in the context of a telephone market survey or information from a written schedule/agenda. The tasks measuring the third claim are two extended problems for which examinees are required to provide solutions or opinions using connected and sustained discourse appropriate for typical workplace problems. See Table 4 for examples of the TOEIC Speaking test questions from the two test forms.

The TOEIC Writing test consists of eight questions, which are classified into three task-types with each type under one of three claims. The first type of task consists of five simple questions, each of which requires examinees to produce a well-formed sentence based on key words provided. The second type of task consists of two items, which present daily life or workplace problems/situations in e-mail format, and examinees are required to respond on how to deal with the problems. The third type of

---

- 1 [1]The number of 400 was suggested based on the experiences and observations of test developers and users from the field. About 25% of examinees in the worldwide TOEIC Listening and Reading test population scored below a combined total score of 400.

task has only one question, and examinees are required to write an essay to express reasons, ideas, evidence and explanations to support an opinion on an issue that could be discussed in daily life or in the workplace. See Table 5 for examples of the TOEIC Writing test questions from the two test forms.

The three claims for each of the TOEIC Speaking and Writing tests were created with an increasing level of complexity and difficulty. The tasks measuring Claim 1 are the least demanding and are assumed to be the easiest, while the tasks measuring Claim 3 are the most difficult and most demanding. Since the nature of the tasks and responses differ, rating scales with different score ranges were required. From the designer's point of view, a scoring scale with a few points (0-3) is considered to be more appropriate for the less demanding tasks, like those measuring Claim 1; a more detailed scale (0-5) is more appropriate for the more demanding and difficult tasks, like those measuring Claim 3 (Hines, 2009).

Derivation of Total Weighted Scores

The three claims were also designed to represent three hierarchical ECD tasks. For example, on the TOEIC Speaking test, examinees who can create connected, sustained discourse for Claim 3 can also carry out routine social interactions and use intelligible language, as Claims 2 and 1 require. On the other hand, examinees who can perform Claim 2 well cannot necessarily perform Claim 3 (Hines, 2009). The same situation is true for the TOEIC Writing test. This design has important implications concerning how to derive the total TOEIC Speaking and Writing scores. Since the lower level claims are measured by a greater number of items than Claim 3, simply totaling all the items in the test to achieve the total test score was ruled out. For the overall test score to reflect the test design assumptions and have the most difficult task contribute the most to the total score, the content design team suggested that weights of 1, 2, and 3 should be applied to the item scores for Claims 1, 2, and 3 (Hines, 2009) when calculating the total test score. That is, the total score is the sum of weighted scores from each claim, and the claim score is the average score of all items measuring the claim.

The reasonableness of this weighting scheme was evaluated by looking at the reliability of the total test score. Other sets of arbitrary weights such as 1, 1, 1; 1, 2, 2; and 1, 3, 4 were applied and compared to the weights of 1, 2, and 3 to check whether a higher level reliability of the total test scores could be achieved. The comparison of the reliability based on different weighting schemes is reported in the test reliability section below.

Statistical Analyses and Results

Difficulty

The difficulty of the test forms was evaluated by examining the frequency distribution of item scores, the mean difficulty of items and claims, and the weighted total test scores for Forms A and B.

Item score distributions - Examining the frequency distribution of item scores helps determine whether the rating scales and rubrics are appropriate for the population. The frequency distributions of item scores are shown in Tables 6–7 for the two forms of the TOEIC Speaking test, and in Tables 8–9 for the two forms of the TOEIC Writing test. In the tables, S represents the items in the TOEIC Speaking test, *W* indicates the items in the TOEIC Writing test, and the numbers after one of these two letters indicates the item positions within the tests. The first item in the TOEIC Speaking test measured two dimensions and produced two scores, with *S1–I* for intonation and *S1–P* for pronunciation. All the items were double rated. *R1* refers to Rating 1, and *R2* refers to Rating 2.

The majority of items had appropriate distributions; that is, the majority of examinees scored at the midpoint(s) of the scale and a smaller percentage of examinees scored at the top and low end of the scale, with some exceptions which are discussed later. Items measuring Claim 3 (i.e., S9, S10, and W8) had small proportions of examinees receiving the top score of 5, and this was expected as

this leaves room for examinee growth. The scores from the first and second ratings were, in general, consistent with each other. Overall, these findings indicate that the rating scales that were designed were fairly appropriate for the population.

However, unexpected score point distributions did occur for some items. For example, the majority of examinees scored 1 instead of the midpoint of the scale for Items 7 and 8 on Form A for the TOEIC Speaking test; Items 1, 4, and 5 on Form A for the TOEIC Writing test; and Items 1 and 5 on Form B for the TOEIC Writing test. The content design team discussed this issue and concluded that these items were simply more difficult. Also, items on Form A of the TOEIC Speaking test relating to Claim 2 were found to have a large proportion of examinees scoring zero. Items 3, 4, and 5 for the TOEIC Writing test also had relatively large proportions of examinees scoring zero. The content design team determined that the rubrics for a score of zero and for missing data for items on the TOEIC Speaking test relating to Claim 2, and rubrics for a score of zero and a score of 1 for items on the TOEIC Writing test relating to Claim 1, were not sufficiently clear for raters to follow. The rubrics were then revised to improve clarity for the operational test.

Item, claim, and total score difficulty - The raw item means and claim means for all examinees and the Japanese (J), Korean (K), and French (F) samples are shown in Tables 10–11 for the two forms of the TOEIC Speaking test and Tables 12–13 for the two forms of the TOEIC Writing test. The weighted raw total score means and standard deviations for the TOEIC Speaking and Writing tests, along with the scaled score means and standard deviations for the TOEIC Listening and Reading test, are shown in Tables 14–15. The tables use *C1*, *C2*, and *C3* for Claim 1 through Claim 3, *SP* for the TOEIC Speaking test, *WR* for the TOEIC Writing test, and *L* for the listening section and *R* for the reading section of the TOEIC Listening and Reading test. Each participant took the TOEIC Listening and Reading test within 6 months of the field study. The French group scored higher on both the TOEIC Listening and Reading test and the TOEIC Speaking and Writing tests than the other groups.

Examinees were randomly assigned to take Form A or Form B. While their scores appeared to be equivalent across forms (364 vs. 365 for the listening section and 314 vs. 315 for the reading section on the TOEIC Listening and Reading tests), their performance on the TOEIC Speaking and Writing tests were not very close to each other, especially for the TOEIC Writing test. Compared with other groups, the French sample for Form B had higher scores than the French sample for Form A on both the TOEIC Speaking and Writing tests. While this may have been due to the small sample sizes for the French group, it is likely that Form A was more difficult than Form B. The issue of form comparability was raised and discussed with the test design team. Test developers decided to review their procedures for test assembly and to seek ways to improve and strengthen the comparability of test forms.

Intercorrelations

Both the TOEIC Speaking and TOEIC Writing test consist of items measuring three distinct claims with various levels of complexity and with different rating scales. Because it is important to evaluate whether the three claims measure the same construct, intercorrelations among the three claims were examined.

The correlations calculated were determined using Pearson product-moment correlation coefficients. Both the observed score correlations and the disattenuated correlations were examined. The disattenuated correlation, known as the true-score correlation, adjusts for the random error of measurement in the variables of interest and was calculated based on the following formula:

$$R_{xy} = \frac{r_{xy}}{\sqrt{r_{xx} * r_{yy}}}$$

where r_{xy} is the correlation between two sets of measures *x* and *y*,

r_{xx} is the reliability coefficient of measure *x*, and

r_{yy} is the reliability coefficient of measure *y*.

The intercorrelations among the three claims for the two forms of the TOEIC Speaking test are shown in Table 16. Based on this table, the observed correlations between the scores on Claims 1 and 2 and Claims 1 and 3 ranged from .54 to .57. The correlations between the scores of Claims 2 and 3 were higher (.63 to .70). As mentioned above, the disattenuated correlations are corrected for the error of measurement associated with each variable. Usually, a disattenuated correlation of .95 and higher indicates that the variables of interest measure the same construct. The disattenuated correlations of scores on Claim 1 with the other two claims ranged from .74 to .86. The highest disattenuated correlations were found between Claims 2 and 3 (.91 and .92). These results are reasonable. Designed as the easiest tasks, Claim 1 does not require examinees to produce extended conversation or sustained discourse as do Claims 2 and 3. For Claim 2, examinees are required to produce dialogues, which is a task more similar to Claim 3. All of these results indicate that the three claims for the TOEIC Speaking test measured similar and correlated constructs, but not the same construct. The findings support the original test design, which has items measuring three distinct claims.

The TOEIC Writing test intercorrelations among the three claims are shown in Table 17. The observed correlations among the claims for the TOEIC Writing test were considerably lower than those for the claims for the TOEIC Speaking test. For example, scores on Claims 1 and 3 had a correlation of .27; the correlations between scores on Claims 2 and 3 were slightly higher, .44 and .45. The disattenuated correlations between scores on Claims 1 and 2 were 0.50 and .58. The other two disattenuated correlations were not available due to the lack of a reliability estimate for Claim 3, which consisted of only one item. Claim 1 is simple and only requires examinees to produce five individual sentences. Claim 2 requires examinees to produce multi-sentence-length text to convey information, instruction, narratives, and so on. Claim 3 requires examinees to produce paragraphs to express complex ideas or support an opinion. Lower ability students might be able to write single sentences well, but are not necessarily able to produce extended text to express or convey ideas, as Claims 2 and 3 require. It is reasonable to observe lower correlations involving Claims 1, 2, and 3 and slightly higher correlations between Claims 2 and 3. Based on the low disattenuated score correlations among the three claims, it is reasonable to conclude that the three claims measure different aspects of speaking.

Test Score Reliability

Rationale and method - Reliability refers to the extent to which the assessment scores obtained remain consistent over repeated administrations of the same test or alternate forms of the test. Reliability also refers to the extent to which the assessment results are free from the effects of random variation caused by factors that may not be directly related to the purpose of the test (e.g., the time of administration, examinee test-taking conditions, scoring conditions).

The reliability of the forms for the TOEIC Speaking and Writing tests is based on the internal consistency method, which is estimated using data from one administration of a single form. When information on a form is available, not only is how well examinees answered items on that particular form a point of interest, but also how well the information from items on the specific form can be generalized. "One way to estimate how consistently examinees' performance on this test can be generalized to the domain of items that might have been asked is to determine how consistently the examinees performed across items or subset of items on this single test form" (Crocker & Algina, 1986). Since a test is composed of a number of items, the internal consistency method treats separate items as repeated measures for examinees, and the interrelationships among scores on the item provide information about reliability. The statistical index calculated based on this approach is called *Cronbach's alpha* or *internal alpha* (Cronbach, 1951).

As discussed earlier, the TOEIC Speaking and Writing tests contain three distinct claims. The three claims vary with respect to item types and levels of complexity. The evaluation of intercorrelations among the three claims supports the original claim of the test design that the three claims do not measure the same construct (see the intercorrelations section). An appropriate reliability estimate for a test of this kind is stratified alpha (Rajaratnam, Cronbach, & Gleser, 1965). The formula for stratified alpha is as follows,

$$stratified\,\alpha\,\rho_{XX'} = \left(1 - \frac{\sum \sigma^2_{x_i}(1 - {}_{\alpha}\rho_{x_i x_i'})}{\sigma^2_x}\right)$$

where, X is weighted total test score,

$\sigma^2_{x_i}$ is total score variance of a particular claim,

σ^2_x is the variance of the weighted total test score, and

${}_{\alpha}\rho_{x_i x_i'}$ is the internal alpha calculated based on items in a particular claim.

The formula for ${}_{\alpha}\rho_{x_i x_i'}$ is specified as follows,

$${}_{\alpha}\rho_{x_i x_i'} = \frac{n}{n-1}\left(1 - \frac{\sum \sigma^2_j}{\sigma^2_{x_i}}\right)$$

where n is the number of items (or set items) in a claim and

σ^2_j is the item (or set item) variance.

Because Item 1 was analytically scored on pronunciation and intonation, Claim 1 for the TOEIC Speaking test has two items, but three scores. When calculating the internal alpha for Claim 1, three scores were used. Claim 2 for the TOEIC Speaking test has two sets of three items. When calculating the internal alpha for Claim 2, six scores were used. The internal consistency reliability estimates for the three claims and total scores based on the first and second rating scores are shown in Table 18. The internal alpha was lowest for Claim 1 (.66 to .68) and highest for Claim 3 (.71 to .74). The stratified internal alpha for total scores ranged from .82 to .86.

To report the subscores for intonation and pronunciation in the TOEIC Speaking test more reliably, it was decided that the number of read-the-text-aloud items in Claim 1 would be increased from 1 to 2 for the final operational form.

For the TOEIC Writing test, Claim 3 consists of only one item; therefore, no internal reliability could be directly estimated for Claim 3. Table 19 shows that reliability ranged from .62 to .66 for Claim 1 and from .52 to .56 for Claim 2. Several efforts were tried to indirectly estimate the upper and lower bounds of the internal reliability of Claim 3. For example, the upper bound was assumed to be the inter-rater reliability of scores given by Rater 1 and Rater 2, and they were .86 and .84 for forms A and B (see Tables 28–29). The lower bound was calculated using the disattenuated correlation formula by assuming that (1) the true correlation between Claim 3 and the combined Claim 1 and Claim 2 scores was equal to 1; or (2) the true correlation of Claim 3 and the combined Claims 1 and 2 scores is equivalent to the true correlation between Claims 1 and 2. These two assumptions for estimating the lower bound of the reliability of Claim 3 yielded unreasonable results. They were too high (e.g., over 1) or too low (e.g., below 0.3) to be useful. Consequently, no stratified alpha could be calculated for the weighted raw total scores for the TOEIC Writing test.

Therefore, the reliability of the TOEIC Writing total scores had to be estimated using the test-retest method through a special study where examinees took both test forms. Such a study was conducted in early spring of 2009 (see Compendium Study 10), and it was found that the test-retest reliability of the TOEIC Writing test was about .83.

Evaluating the impact of weights on reliability estimates - As mentioned earlier in this report, the content developers considered the weights of 1, 2, and 3 to be most reasonable for Claims 1, 2, and 3 when deriving the total weighted scores. Other sets of weights were also considered and evaluated in terms of their impact on the reliability of total scores. The following sets of weights were tried: 1, 1, 1; 1, 2, 3; 1, 2, 2; 1, 2, 4. The results of stratified alpha reliability estimates based on the four different sets of weights are shown in Tables 20 and 21 for the two forms of the TOEIC Speaking and Writing tests. For the TOEIC Writing test, a reliability estimate of 0.73 was used for Claim 3 so that the stratified alpha could be estimated for the total weighted scores. This number was derived from the 2009 test-retest reliability study. It was used as the internal alpha estimate for Claim 3 simply for the purpose of evaluating the impact of weights on the total score reliability estimate.

Table 20 shows the variation in reliability estimates for the two forms of the TOEIC Speaking test when applying the different weighting schemes. Estimates are provided for each of the raters. The weighting scheme that produced the largest reliability estimates (.85 to .88) was 1, 1, 1. This reliability is .0275 higher than the 1, 2, 3 weighting scheme. However, this set of weights (1, 1, 1) was not consistent with the intention of test design from a content point of view. As a result, the weighting scheme of 1, 2, 3 was adopted for operational use. Table 21 shows the reliability estimates for the two forms of the TOEIC Writing test, which are very similar across different weighting schemes. The variation between the first and second raters was also trivial.

Rater Agreement

Because the TOEIC Speaking and Writing tests are rated by human raters, it is important to evaluate the consistency of ratings given by different raters. All of the items were rated by two raters. Two types of analyses were conducted: inter-rater agreement rates and inter-rater reliability.

Inter-rater agreement rate. The inter-rater agreement rates and correlations based on the first and second raters are shown in Tables 22–23 for the two forms of the TOEIC Speaking test and Tables 24–25 for the two forms of the TOEIC Writing test. For all the TOEIC Speaking and Writing tests items, the exact percentage of agreement ranged from 50% to 82%, meaning that at least half of the examinees received the same ratings from the first and second raters on all items. Most of the data in the *discrepancy 2 + %* column are less than 1 %, with the highest value being 1.4%, which leads to the conclusion that a very low percentage of examinees obtained a score difference of 2 or more points from the two raters. This was consistent with the inter-rater correlations, which ranged from 0.47–0.85 for the TOEIC Speaking test items and 0.70–0.87 for the TOEIC Writing test items. Therefore, the scoring results from different raters were fairly consistent for all the TOEIC Speaking and Writing test items.

However, the rating consistency also depends on the types of items within and across the TOEIC Speaking and Writing tests. Comparing the scoring results for different items within each of the tests, one can see that the more possible score points for an item yields a lower inter-rater consistency. Specifically, for the TOEIC Speaking test, Items 9 and 10 (with the maximum of 5 score points) had a lower exact percentage agreement than other items (with the maximum of 3 score points). For the TOEIC Writing test, Items 6–7 (with the maximum of 4 score points) and Item 8 (with the maximum of 5 score points) had a lower exact percentage agreement than the other items (with only 3 possible score points). The inter-rater correlations for the TOEIC Writing test Items 6, 7, and 8 were also relatively lower than those for other items on that test.

The inter-rater scoring consistency also depends on the type of characteristics of different responses, even if they are scored by the same rating scale. For the TOEIC Speaking test, items related to Claim 1 (Items 1 and 2) and Claim 2 (Items 3–8) were rated using a rating scale 0–3. Items related to Claim 1, in general, had lower exact percentage agreement and lower inter-rater correlations than items in Claim 2. The first claim in the TOEIC Writing test was also rated by a scale of 0–3; however, items in this claim received much higher exact percentage agreement rate (i.e., 70% and up), and the inter-rater correlations were also high (i.e., 0.78 and up). The discrepancies could be attributed to the complexity of rating the task itself. While judging the performance of a person's intonation and pronunciation can be difficult and more subjective, judging the grammatical soundness of a sentence appears to be easier and more objective. Given the higher inter-rater consistency for the TOEIC Writing test items, it seems that it is easier to evaluate the writing responses than the speaking responses. These results were discussed with the content development team, and the rubrics for Claim 1 for the TOEIC Speaking test were refined and the rater training was strengthened.

Inter-rater reliability - The inter-rater reliability was calculated for each item based on a generalizability study using the (p x r') model, where r' indicates rating instead of rater because we didn't have a full crossed-person-by-rater design. In this study, a very large pool of raters from the ETS Online Scoring Network (OSN) was randomly assigned to examinees' responses to enhance rating accuracy and consistency. Tables 26–29 show the breakdown of the item score variation attributed to examinees, p (i.e., how much examinees differ from each other in their abilities to respond to an item), ratings, r' (i.e., whether one rating is more lenient than the other), and their interaction p x r' (i.e., whether the relative standing of examinees differs across ratings). The generalizability index, ρ^2, is reported in the tables as the inter-rater (or strictly speaking interrating) reliability coefficient.

In general, the reliability coefficients were reasonably high, with .8 and up for items on the TOEIC Speaking test (except for items in the first claim and two items for the second claim in Form B), and .80 and up for items on the TOEIC Writing test (except for items for the second claim). Two factors can affect the reliability coefficient: the agreement rate percentage and item total variance. For example,

the reason that the first claim for the TOEIC Speaking test had low reliability is because the percentage exact agreement rates were low, meaning different raters ranked examinees differently and, accordingly, the (*p* x *r'*) error variance components tended to be large. In Form A, Item S1_I had 50% exact agreement and W1 had 79%; accordingly, the former had an inter-rater reliability of .58 while the latter .87. In addition, Item S1_I has 59% of total item variance attributable to (*p* x *r'*) error variance; Item W1 has only 23% of the total item variance attributable to (*p* x *r'*) error variance. Therefore, the two indices, inter-rater reliability and inter-rater agreement rate, produced consistent and reasonable results for raters' scoring quality.

Conclusion

This study evaluated the statistical properties of two pilot forms for the TOEIC Speaking and Writing tests. The study found that the difficulty of and the rating scales for the tests were appropriate for the target population. The appropriateness of test difficulty, the relationships of different claims, the reliability of the total test scores, and the quality of rater agreement were all examined. In terms of item difficulty, some items displayed abnormal responses, such as too many examinees scoring zero, which resulted from a lack of clarity in the scoring rubrics. Subsequently, the rubrics were reviewed and improved for clarity. The results of intercorrelations of scores related to the three ECD claims for each test indicate that the test development team has achieved its goal of designing the tests to have three distinct claims. The reliability of total test scores was reasonably high as expected, although estimating the internal alpha was not plausible for the TOEIC Writing test. In order to establish a reliability estimate for the TOEIC Writing test, test-retest reliability was estimated from a separate study. The rater-agreement was also found to be reasonably high for the test and test forms.

In summary, the pilot study achieved its goals. Important information based on the statistical results was provided by the study, which helped in shaping and improving the final specifications for the TOEIC Speaking and Writing tests.

References

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Toronto, Ontario, Canada: Holt, Rinehart & Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.

Hines, S. (2009). *Rationale for scoring scales and weights.* Unpublished manuscript.

Rajaratnam, N., Cronbach, L. J., & Gleser, G. C. (1965). Generalizability of stratified-parallel tests.

Psychometrika, 30, 39–56.

TABLE 1

Number of Examinees From Different Countries Taking the Four Test Forms

Form	Japan	Korea	France	Total
A	719	301	96	1,116
B	703	314	88	1,105
C	188	13	0	201
D	0	115	0	115
Total	1,610	743	184	2,537

TABLE 2

Test Specifications for the TOEIC Speaking Test

Claims	Question no.	Time (in seconds)		Evaluation criteria	Rubrics
		Preparation	Response		
1	1	60	60	• Pronunciation • Intonation	(analytic) Levels 0-3 Levels 0-3
	2	30	45	• Pronunciation • Intonation • Delivery • Content • Grammar	(holistic) Levels 0-3
2	3–5	0	15,15,30	• Pronunciation • Intonation • Delivery • Vocabulary • Grammar	(holistic) Levels 0-3
	6-8	30	15,15,30	• Delivery • Vocabulary • Grammar • Info from schedule	(holistic) Levels 0-3
3	9	45	60	• Delivery • Content • Grammar • Vocabulary	(holistic) Levels 0-5
	10	15	60	• Delivery • Content • Grammar • Vocabulary	(holistic) Levels 0-5

TABLE 3

Test Specifications for the TOEIC Writing Test

Claims	Question no.	Time (in minutes)		Evaluation criteria	Rubrics
		Preparation	Response		
1	1–5	1	1	• Grammar • Key words • Contents	Levels 0-3
2	6 – 7	10	10	• Task, tone/register, and grammar and usage • Cohesion • Tone/register • Grammar and usage	Levels 0-4
3	8	30	30	• Whether your opinion is supported with reasons and examples • The quality and variety of your sentences • The range and appropriateness of your vocabulary • Your overall organization	Levels 0 - 5

TABLE 4

The TOEIC Speaking Test Questions From Forms A and B

Task	Question no.	Form A	Form B
Read text aloud	1	Some text	Some text
Talk about a picture	2	Beach (people, chairs, birds)	Market (bananas, weighing)
Respond to questions	3 – 8	(Imagine a telephone interview about …) <u>Local transportation (3 – 5)</u> 3. How do you travel from home to work or study? 4. How long does the travel take? 5. How could transportation be improved in your area? (Answer a caller's questions about a schedule) <u>Orientation schedule (6 – 8)</u> 6. Where is it and when should we be there? 7. Will someone be showing us around the building? 8. What are the activities besides paperwork?	(Imagine a telephone interview about …) <u>Television viewing (3 – 5)</u> 3. How often do you watch TV? 4. What programs do you watch? 5. Describe your favorite TV program. (Answer a caller's questions about a seminar) <u>Conference schedule (6 – 8)</u> 6. When does it start and how long it will last? 7. How much does it cost? 8. What are the activities before lunch?
Propose a solution	9	(Read a table of information) Recommend one of the two hotels using information from the chart.	(Hear a phone message) A caller complained that she could not get her banking card out of the ATM machine, propose a solution.
Express an opinion	10	Issue: Do you agree with wearing uniforms in school?	Issue: Do you prefer to take a job with a low salary but a lot of vacation time, or the other way around?

TABLE 5

The TOEIC Writing Test Questions From Forms A and B

Task	Question no.	Form A	Form B
Write a sentence based on a picture (using two words provided)	1 – 5	1. Child/push (a lady with a child in a stroller) 2. Near/building (people playing tennis) 3. Snowboard/after (a man walking with a snowboard) 4. Box/to (a man moving boxes) 5. Luggage/because (people waiting/taking luggage)	1. Next to/worker (two workers sitting next to each other) 2. Motorbike/groceries (a man riding a motorbike carrying groceries) 3. Camera/very (two men sitting, one looking at a camera, the other reading documents) 4. Airport terminal/so (many cars parking by the terminal entrance) 5. Eat/who (parents and kids sitting and eating)
Respond to an E-mail	6 – 7	1. Work schedule – explain about two times when you cannot work 2. Bill – make at least two requests for info about your bill	6. Move to a new city – make at least two requests for information 7. New computer program – make at least one request for information and explain at least two actions a user must take
Write an opinion essay	8	Employees should/should not use company equipment for personal needs.	Best way to find a job – newspaper, Internet, or personal recommendation

TABLE 6

Frequency Distribution (in %) of Item Scores – The TOEIC Speaking Form A (N = 1,116)

Item	S1 – I		S1 – P		S2		S3		S4		S5		S6		S7		S8		S9		S10	
Scale	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2
0	2	1	2	1	1	1	21	22	5	4	16	17	14	15	26	29	31	30	2	2	1	1
1	21	26	15	15	30	27	21	22	19	21	35	33	33	34	53	54	38	38	3	3	10	10
2	55	52	59	65	55	58	40	39	47	44	38	37	35	35	17	13	22	23	25	29	26	28
3	22	21	24	19	14	14	18	18	29	31	11	12	18	15	4	3	9	10	42	41	38	35
4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	24	21	19	20
5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	5	5	5	5

TABLE 7

Frequency Distribution (in %) of Item Scores – The TOEIC Speaking Form B (N = 1,105)

Item	S1 – I		S1 – P		S2		S3		S4		S5		S6		S7		S8		S9		S10	
Scale	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2
0	1	1	1	1	1	1	5	6	1	1	2	2	6	6	2	2	4	4	5	5	4	4
1	21	22	19	18	27	25	13	12	14	13	25	23	23	24	19	19	25	22	8	8	6	8
2	52	54	57	61	56	58	48	51	60	60	53	54	46	46	53	55	57	57	31	30	24	25
3	26	24	23	21	16	17	33	31	25	26	21	20	26	25	26	24	15	17	40	42	44	41
4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	15	13	19	19
5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	2	4	3

TABLE 8

Frequency Distribution (in %) of Item Scores – The TOEIC Writing Form A (N = 1116)

Item	W1		W2		W3		W4		W5		W6		W7		W8	
Scale	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2
0	1	0	1	1	6	6	10	10	14	14	1	1	2	1	1	1
1	43	45	21	20	34	36	36	34	38	37	18	21	13	14	17	16
2	38	37	32	34	41	39	31	33	33	33	29	28	27	34	39	41
3	18	18	45	45	20	19	23	22	15	15	37	35	45	40	33	32
4	.	.	.	.	.	.	.	.	.	.	14	15	13	11	8	9
5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1

TABLE 9

Frequency Distribution (in %) of Item Scores – The TOEIC Writing Form B (N = 1105)

Item	W1		W2		W3		W4		W5		W6		W7		W8	
Scale	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2
0	2	2	2	2	7	7	12	12	16	16	0	1	1	1	0	0
1	50	53	29	29	37	39	30	30	45	43	14	14	12	13	6	6
2	30	28	45	46	39	36	39	40	24	26	22	22	47	43	27	30
3	17	17	25	23	18	18	19	17	15	15	41	39	33	34	43	39
4	.	.	.	.	.	.	.	.	.	.	22	24	8	10	19	19
5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	6	5

TABLE 10

Raw Item Means by Various Groups – The TOEIC Speaking Test Form A

	%	N	S1–I	S1–P	S2	S3	S4	S5	S6	S7	S8	S9	S10	C1	C2	C3	SP
All	100	1116	1.9	2.0	1.9	1.5	2.0	1.4	1.5	0.9	1.1	2.9	2.8	1.9	1.4	2.8	13.3
J	64	719	1.9	2.0	1.8	1.4	2.0	1.3	1.4	0.8	0.9	2.8	2.6	1.9	1.3	2.7	12.8
K	27	301	1.9	2.0	1.8	1.6	1.9	1.5	1.6	1.0	1.4	2.9	2.9	1.9	1.5	2.9	13.6
F	9	96	2.1	2.1	2.0	2.1	2.3	2.0	1.8	1.4	2.1	3.5	3.4	2.1	2.0	3.5	16.4

TABLE 11

Raw Item Means by Various Groups – The TOEIC Speaking Test Form B

	%	N	S1–I	S1–P	S2	S3	S4	S5	S6	S7	S8	S9	S10	C1	C2	C3	SP
All	100	1105	2.0	2.0	1.9	2.1	2.1	1.9	1.9	2.0	1.9	2.6	2.7	2.0	2.0	2.6	13.9
J	64	703	1.9	1.9	1.9	2.1	2.1	1.9	1.9	2.0	1.8	2.4	2.6	1.9	1.9	2.5	13.3
K	28	314	2.1	2.1	1.9	2.1	2.0	2.0	1.9	2.1	1.9	2.6	2.8	2.0	2.0	2.7	14.1
F	8	88	2.4	2.3	2.3	2.3	2.2	2.4	2.1	2.1	2.3	3.4	3.5	2.4	2.2	3.5	17.2

TABLE 12

Raw Item Means by Various Groups – The TOEIC Writing Test Form A

	%	N	W1	W2	W3	W4	W5	W6	W7	W8	C1	C2	C3	WR
All	100	1116	1.7	2.2	1.7	1.7	1.5	2.4	2.4	2.4	1.8	2.4	2.4	13.7
J	64	719	1.7	2.2	1.6	1.5	1.3	2.4	2.4	2.3	1.7	2.4	2.3	13.5
K	27	301	1.8	2.3	1.9	2.0	1.9	2.5	2.4	2.3	1.9	2.5	2.3	13.7
F	9	96	1.9	2.3	1.7	1.8	1.9	2.6	2.6	2.8	1.9	2.6	2.8	15.6

TABLE 13

Raw Item Means by Various Groups – The TOEIC Writing Test Form B

	%	N	W1	W2	W3	W4	W5	W6	W7	W8	C1	C2	C3	WR
All	100	1105	1.6	1.9	1.7	1.6	1.4	2.7	2.4	2.9	1.6	2.5	2.9	15.4
J	64	703	1.5	1.9	1.6	1.5	1.2	2.8	2.4	2.8	1.5	2.6	2.8	15.2
K	28	314	1.7	1.9	1.8	1.9	1.7	2.5	2.3	2.7	1.8	2.4	2.7	14.8
F	8	88	1.7	2.0	1.9	1.9	2.0	2.9	2.7	3.7	1.9	2.8	3.7	18.7

TABLE 14

Mean (SD) of Form A

	%	N	SP	WR	L	R
All	100	1116	13.3 (4.1)	13.7 (4.0)	364.4 (72.9)	314.0 (77.4)
J	64	719	12.8 (3.9)	13.5 (3.9)	354.8 (68.9)	306.1 (74.6)
K	27	301	13.6 (4.1)	13.7 (4.2)	381.5 (75.9)	325.6 (79.8)
F	9	96	16.4 (4.2)	15.6 (3.7)	381.9 (80.3)	336.5 (81.9)

TABLE 15

Mean (SD) of Form

	%	N	SP	WR	L	R
All	100	1105	13.9 (3.8)	15.4 (4.1)	365.1 (72.8)	315.7 (77.4)
J	64	703	13.3 (3.6)	15.2 (3.9)	354.5 (69.5)	305.7 (75.0)
K	28	314	14.1 (3.6)	14.8 (4.2)	377.6 (76.2)	325.1 (79.0)
F	8	88	17.2 (3.9)	18.7 (3.9)	405.5 (65.2)	361.4 (69.8)

TABLE 16

Correlations Among the Different Claims for the TOEIC Speaking Test

	Form A		Form B	
Claims	Observed score correlation	Disattenuated correlation	Observed score correlation	Disattenuated correlation
C1 – C2	0.54	0.74	0.57	0.86
C1 – C3	0.57	0.82	0.57	0.83
C2 – C3	0.70	0.91	0.63	0.92

TABLE 17

Correlations Among the Different Claims for the TOEIC Writing Test

	Form A		Form B	
Claims	Observed score correlation	Disattenuated correlation	Observed score correlation	Disattenuated correlation
C1 – C2	0.33	0.58	0.29	0.50
C1 – C3	0.27	NA	0.29	NA
C2 – C3	0.45	NA	0.44	NA

TABLE 18

Internal Alpha for the Claims for the TOEIC Speaking Test

Form	Claim 1		Claim 2		Claim 3		Total	
	R1	R2	R1	R2	R1	R2	R1	R2
A	0.67	0.66	0.79	0.8	0.73	0.74	0.85	0.86
B	0.68	0.66	0.67	0.66	0.74	0.71	0.83	0.82

TABLE 19

Internal Alpha for the Claims for the TOEIC Writing Test

Form	Claim 1		Claim 2		Claim 3	
	R1	R2	R1	R2	R1	R2
A	0.63	0.62	0.52	0.52	NA	NA
B	0.66	0.66	0.56	0.52	NA	NA

TABLE 20

Reliability (Alpha) of Weighted Total Scores for the TOEIC Speaking Test Forms A and B

Form	1,1,1		1,2,3		1,2,2		1,3,4	
	R1	R2	R1	R2	R1	R2	R1	R2
A	0.88	0.88	0.85	0.86	0.87	0.87	0.86	0.86
B	0.86	0.85	0.83	0.82	0.85	0.83	0.83	0.81

TABLE 21

Reliability (Alpha) of Weighted Total Scores for the TOEIC Writing Test Forms A and B

Form	1,1,1		1,2,3		1,2,2		1,3,4	
	R1	R2	R1	R2	R1	R2	R1	R2
A	0.79	0.79	0.79	0.79	0.78	0.78	0.78	0.78
B	0.80	0.79	0.80	0.79	0.79	0.78	0.79	0.78

TABLE 22

Rating Agreement Rate – Between 1st and 2nd Rating – The TOEIC Speaking Test Form A

Item	Exact %	Adjacent %	Discrepancy of 2 + %	Total N	Correlation
S1 – I	50.0	48.9	1.1	1116	0.47
S1 – P	59.4	40.3	0.3	1116	0.51
2	62.2	37.6	0.2	1116	0.56
3	65.8	34.1	0.1	1116	0.83
4	65.3	34.6	0.1	1116	0.74
5	61.8	37.6	0.5	1116	0.76
6	72.8	27.2	0.0	1116	0.85
7	80.7	19.2	0.1	1116	0.84
8	74.9	24.8	0.3	1116	0.85
9	51.3	47.7	1.0	1116	0.74
10	53.6	45.3	1.2	1116	0.78

TABLE 23

Rating Agreement Rate – Between 1st and 2nd Rating – The TOEIC Speaking Test Form B

Item	Exact %	Adjacent %	Discrepancy of 2 + %	Total *N*	Correlation
S1 – I	53.2	45.6	1.2	1105	0.50
S1 – P	62.1	37.3	0.6	1105	0.55
2	60.5	39.5	0.1	1105	0.54
3	65.5	34.5	0.0	1105	0.74
4	60.1	39.8	0.1	1105	0.53
5	64.4	35.1	0.5	1105	0.64
6	63.0	36.6	0.5	1105	0.73
7	67.3	32.5	0.2	1105	0.67
8	67.6	32.2	0.2	1105	0.69
9	59.5	40.4	0.2	1105	0.81
10	53.7	45.6	0.7	1105	0.78

TABLE 24

Rating Agreement Rate – Between 1st and 2nd Rating – The TOEIC Writing Test Form A

Item	Exact %	Adjacent %	Discrepancy of 2 + %	Total *N*	Correlation
1	79.3	20.3	0.4	1116	0.80
2	82.1	17.7	0.2	1116	0.86
3	71.8	28.0	0.3	1116	0.79
4	71.9	27.2	1.0	1116	0.82
5	70.1	29.3	0.6	1116	0.81
6	58.4	41.0	0.6	1116	0.78
7	56.2	43.3	0.5	1116	0.74
8	64.2	34.8	1.1	1116	0.78

TABLE 25

Rating Agreement Rate – Between 1st and 2nd Rating – The TOEIC Writing Test Form B

Item	Exact %	Adjacent %	Discrepancy of 2 + %	Total *N*	Correlation
1	74.5	25.1	0.5	1105	0.78
2	74.1	25.7	0.2	1105	0.78
3	82.1	17.6	0.4	1105	0.86
4	80.2	19.5	0.3	1105	0.87
5	78.0	21.4	0.6	1105	0.86
6	56.3	42.5	1.2	1105	0.76
7	58.4	41.4	0.3	1105	0.70
8	55.7	43.0	1.4	1105	0.74

TABLE 26

Inter-rater Reliability for the TOEIC Speaking Test Form A

Task	SP 1_I		SP 1_P		SP 2		SP 3		SP 4		SP 5		SP 6		SP 7		SP 8		SP 9		SP 10	
Source of var	Var	%	Var	%	Var	%	Var	%	Var	%	Var	%	Var	%	Var	%	Var	%	Var	%	Var	%
P	0.21	41	0.20	46	0.23	54	0.84	81	0.48	72	0.59	72	0.73	83	0.46	81	0.74	84	0.65	66	0.87	75
r'	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0
p x *r'*	0.30	59	0.23	54	0.20	46	0.20	19	0.19	28	0.22	28	0.15	17	0.11	19	0.15	16	0.33	34	0.29	25
Total	0.51	100	0.43	100	0.43	100	1.04	100	0.67	100	0.81	100	0.88	100	0.57	100	0.89	100	0.98	100	1.17	100
N	**1116**																					
p^2	0.58		0.63		0.70		0.89		0.83		0.84		0.91		0.89		0.91		0.80		0.86	
Exp Obs SD	0.60		0.56		0.58		0.97		0.76		0.84		0.90		0.72		0.90		0.90		1.01	
SEM	0.29		0.25		0.23		0.23		0.22		0.24		0.19		0.17		0.19		0.30		0.28	

TABLE 27

Inter-rater Reliability for the TOEIC Speaking Test Form B

Task	SP 1_I		SP 1_P		SP 2		SP 3		SP 4		SP 5		SP 6		SP 7		SP 8		SP 9		SP 10	
Source of var	Var	%	Var	%	Var	%	Var	%	Var	%	Var	%	Var	%	Var	%	Var	%	Var	%	Var	%
P	.22	44	0.24	54	0.23	53	0.48	71	0.22	51	0.31	60	0.49	69	0.34	66	0.36	67	0.87	80	0.84	74
r'	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0
p x *r'*	0.28	56	0.20	46	0.20	47	0.20	29	0.22	49	0.21	40	0.22	31	0.17	34	0.17	33	0.22	20	0.29	26
Total	0.51	100	0.44	100	0.44	100	0.67	100	0.44	100	0.52	100	0.71	100	0.51	100	0.53	100	1.09	100	1.14	100
N	**1105**																					
p^2	0.61		0.70		0.69		0.83		0.67		0.75		0.82		0.80		0.81		0.89		0.85	
Exp Obs SD	0.61		0.58		0.58		0.76		0.57		0.64		0.77		0.65		0.67		0.99		0.99	
SEM	0.28		0.24		0.24		0.23		0.24		0.24		0.24		0.21		0.21		0.24		0.27	

TABLE 28

Inter-rater Reliability for theTOEIC Writing Test Form A

Task	WR 1		WR 2		WR 3		WR 4		WR 5		WR 6		WR 7		WR 8	
Source of var	Var	%	Var	%	Var	%	Var	%	Var	%	Var	%	Var	%	Var	%
P	0.44	77	0.56	85	0.55	78	0.70	80	0.65	78	0.70	64	0.57	66	0.67	75
r'	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	1	0.00	0
p x *r'*	0.13	23	0.10	15	0.16	22	0.17	20	0.18	22	0.39	36	0.29	34	0.22	25
Total	0.57	100	0.66	100	0.70	100	0.87	100	0.83	100	1.08	100	0.87	100	0.89	100
N	**1116**															
p^2	0.87		0.92		0.87		0.89		0.88		0.78		0.80		0.86	
Exp Obs SD	0.71		0.78		0.79		0.89		0.86		0.94		0.85		0.88	
SEM	0.18		0.16		0.20		0.21		0.22		0.32		0.28		0.24	

TABLE 29

Inter-rater Reliability for the TOEIC Writing Test Form B

Task	WR 1		WR 2		WR 3		WR 4		WR 5		WR 6		WR 7		WR 8	
Source of var	Var	%	Var	%	Var	%	Var	%	Var	%	Var	%	Var	%	Var	%
P	0.47	77	0.45	76	0.62	87	0.73	87	0.68	79	0.68	69	0.45	65	0.68	75
r'	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0
p x *r'*	0.14	23	0.14	24	0.09	13	0.11	13	0.19	21	0.31	31	0.24	35	0.23	25
Total	0.62	100	0.59	100	0.72	100	0.84	100	0.87	100	0.99	100	0.69	100	0.92	100
N	**1105**															
p^2	0.87		0.86		0.93		0.93		0.88		0.82		0.79		0.84	
Exp Obs SD	0.74		0.72		0.82		0.88		0.88		0.91		0.76		0.90	
SEM	0.19		0.19		0.16		0.16		0.22		0.28		0.25		0.26	