Development and Validation of a Self-assessment Questionnaire as an Indicator of TOEIC® Test Score

Akihiro ITO

Seinan Gakuin University, Japan and

University of Central Lancashire, UK

Hiroshi SHIMATANI

Kumamoto University, Japan

Ken NORIZUKI

Shizuoka Sangyo University, Japan

Masayoshi KINOSHITA

Part-time lecturer, Fukuoka University, Japan

Abstract

A test should be reliable and valid enough for a test-taker to predict, assess, and/or judge his/her own test performance immediately after taking it. Based on this point of view, the present study investigated TOEIC®'s validity from the perspective of data on test-takers' post-test-taking psychological attitudes. The investigation involved the following three steps: (a) developing a self-assessment questionnaire as a tool for gathering data on participants' psychological attitudes, (b) conducting validation studies on the self-assessment questionnaire, and (c) examining the levels of "matching" between TOEIC® test scores and test-takers' judgment on their test performance. Results indicated that both versions of the TOEIC® showed high reliability coefficients and the self-assessment questionnaire showed a high reliability coefficient. In order to examine the construct validity of the self-assessment questionnaires, principal component analyses were used. On the basis of the scree plot, a single factor solution was deemed appropriate for each of the two data, implying that the factor structure underlying each of the self-assessment questionnaires is unidimensional. Moreover, there are low to moderate correlations between each part of the TOEIC® and the item self-assessment questionnaire concerned, implying that the two self-assessment questionnaires are valid indicators of TOEIC® test scores.

Introduction

In May 2006, the TOEIC® was redesigned in order to make it a more suitable test battery to measure English communicative language ability. The Educational Testing Service (ETS) says that the tasks in the assessment have been refined to make them more authentic, using various contexts from real-world settings (ETS, 2005).

Table 1 illustrates the formats of the old TOEIC® and the new TOEIC®.

Table 1
The Format of Two Versions of TOEIC®

Old TOEIC®			New TOEIC®		
Part	Name of each part	Number of Questions	Part	Name of each part	Number of Questions
Listening section (45min.)			Listening section (45min.)		
I	Photographs	20	1	Photographs	10
II	Question-Response	30	2	Question-Response	30
III	Short Conversations	30	3	Short Conversations	30
IV	Short Talks	20	4	Short Talks	30
Reading section (75min.)			Reading section (75min.)		
V	Incomplete Sentences	40	5	Incomplete Sentences	40
VI	Error Recognition	20	6	Text Completion	12
VII Reading Comprehension			Reading		
			Comprehension		
			• Single passage		
			• Double passage		

(Adapted from TOEIC® Steering Committee, 2006)

Duke (2005) summarizes the aims of the changes made in the Listening sections and Reading section in the following way.

Changes made in the Listening section:

- (1) Some longer listening stimuli – Broader measurement of ability
- (2) Short conversations as sets – More questions per conversation, fewer topics to process.
- (3) Reduced number of photo items – More testing time for other authentic tasks.
- (4) Varied accents – U.S., British, Canadian, Australian (as well as New Zealand)
 - Reflects English taught and spoken in the international environment.

Changes made in the Reading section:

- (1) Selected sentence completion items are passage-based – Allows the testing of

context-related skills.

- (2) Selected reading sets use two interrelated passages – Requires test taker to connect information from both texts, provides more authentic real-world tasks.
- (3) Some reading stimuli are longer – Provides a broad measurement of examinees' abilities, more questions per stimulus.
- (4) Error recognition items are eliminated. (p. 4)

Since 2009, the present authors have reported multiple perspectives on the difference between the two versions of the TOEIC® in terms of the participants' test performance. The data have been analyzed on the basis of classical test theory (Ito, Shimatani, Norizuki & Kinoshita, 2009a) and the Rasch model (Shimatani, Norizuki, Ito & Kinoshita, 2009a). Attempts to examine the validity of the two versions of TOEIC® as a placement test include an internal correlation study (Ito, Shimatani, Norizuki & Kinoshita, 2009b; Shimatani, Norizuki, Ito & Kinoshita, 2009b) and a predictive validation study (Ito, Shimatani & Norizuki, 2010). Furthermore, diagnostic roles of TOEIC® in Japan have been examined in the latest report (Norizuki, Ito & Shimatani, 2011). In these studies, the participants' TOEIC® scores have been focused on as a main source for exploring the validity issues of TOEIC® in multiple testing conditions and with different interpretations in Japan.

In the present paper, however, we consider the validity of TOEIC® from the perspective of test-takers' self-evaluation of their own test performance. The primary motive for this research is the fact that not much research has been conducted on test-takers' psychological attitudes after taking the test.

Research on self-assessment has mainly focused on examining the relationship between the test score and test-takers' perceptions of language proficiency (Powers, Kim, Yu, Weng & VanWinkle, 2009). In such studies, researchers are interested in examining the applicability of test scores for predicting test-takers' language performance in external contexts, such as English-speaking performance in the workplace.

Consequently, not much research has been carried out to investigate whether there is any "matching" between test-takers' self-assessment or judgment of their own test performance and their actual test scores. This kind of issue has been out of the mainstream in language testing research. However, we believe that more attention should be paid to this area, because a test-taker's post-test-taking attitude should closely parallel his/her test score.

A test should be reliable and valid enough for a test-taker to predict, assess, and/or judge his/her own test performance immediately after taking it. From this point of view, the present study investigated TOEIC®'s validity from the perspective of data on test-takers' post-test-taking psychological attitude.

To reach the goal of the present investigation, we take the following three steps:

(1) develop a self-assessment questionnaire as a tool for gathering psychological attitude data from participants, (2) conduct a validation study on the self-assessment questionnaire, and (3) employ the self-assessment questionnaire to examine the quality of the two versions of TOEIC® in terms of the levels of “matching” between test-takers’ judgment of their test performance and their test scores.

Method

Purpose

The purpose of the present study is threefold: (1) to construct a self-assessment questionnaire, (2) to conduct validation studies, and (3) to examine whether TOEIC® is a valid test from the perspective of the level of “matching” between the test-takers’ self-assessment of their test performance and their test scores, comparing the two versions of TOEIC®.

Constructing a Questionnaire on the Basis of the Score Descriptor Table

In ESL/EFL classrooms, an open-ended self-assessment questionnaire is quite often administered to test-takers who have finished taking the test in order to learn about their reaction to the test or judgment of their own English language proficiency. TOEIC® is a two-hour test with as many as 200 questions, which probably means giving the test-takers an open-ended or demographic questionnaire after they have taken it reduces the validity of the questionnaire itself. Therefore, we decided to construct an easy-to-administer and easy-to-take questionnaire that would take only a few minutes to answer.

Our first step was to examine the TOEIC® Score Descriptor Tables (SDTs) (ETS, 2007). The SDT is a description of English language proficiency based on the score in TOEIC® (Appendix). It was developed and validated in a number of studies carried out by the Educational Testing Service in the process of renewing the format of TOEIC® as a new test (Liao, 2010). Since May 2006, TOEIC® takers have received the Official Score Certificate with SDTs that show both the strong and weak points in their English language proficiency. Therefore, the SDT is an empirically validated indicator of English language proficiency based on the TOEIC® score. Obviously, the SDT is unrelated to test-takers’ self-assessment of their test performance. As mentioned, the primary aim of the present research is to examine the level of “matching” between the test-takers’ self-assessment of their test performance on TOEIC® and their actual test score. We believe that we can explore the test-takers’ post-test-taking psychological attitude if we construct questionnaire items on the basis of the SDTs. In doing so, we can examine how participants judged their own TOEIC® performance without directly asking them, “What score do you think you got today?”

As the second stage of constructing the self-assessment questionnaire, we read through the SDT and identified words and phrases that are frequently utilized for

describing strong and weak points in English language proficiency levels based on the TOEIC® score. We found the following words and phrases to be frequently and consistently used in SDTs for almost all levels of both the Listening and Reading sections in TOEIC®.

- (1) *“infer the central idea and purpose”*
- (2) *“understand grammatical structures”*
- (3) *“understand details”*

The Final Versions of the Questionnaire

The following are the final versions of the self-assessment questionnaires we constructed. For the older version of TOEIC®, the questionnaire contained seven items (four questions in the Listening section and three in the Reading section). For the new version, it contained eight: One more question was added to the Reading section as a final question (Question 8). This is because in New TOEIC®, the last sets of reading comprehension questions refer to two interrelated passages in order to require test-takers to connect information from both texts, which is a more authentic, real-world task. In reaction to this part of the test, we added the question: “Were you able to understand the interrelation between the passages in reading a written text?”

Self-assessment Questionnaire for Old TOEIC®

Listening Comprehension (Parts I, II, III, and IV)

1. Were you able to infer the central idea, purpose, and basic context of short, spoken exchanges in Part 1 and 2?
2. Were you able to infer the central idea, purpose, and basic context of short exchanges in Parts 3 and 4?
3. Were you able to understand grammatical structures and vocabulary in Parts 1 and 2?
4. Were you able to understand details in Parts 3 and 4?

Reading Comprehension (Parts V, VI, and VII)

5. Were you able to understand grammatical structures in the text?
6. Were you able to understand vocabulary in the text?
7. Were you able to infer the central idea and purpose of the text?

Self-assessment Questionnaire for New TOEIC®

Listening Comprehension (Parts 1, 2, 3, and 4)

1. Were you able to infer the central idea, purpose, and basic context of short, spoken exchanges in Parts 1 and 2?
2. Were you able to infer the central idea, purpose, and basic context of short

exchanges in Parts 3 and 4?

3. Were you able to understand grammatical structures and vocabulary in Parts 1 and 2?

4. Were you able to understand details in Parts 3 and 4?

Reading Comprehension (Parts 5, 6, and 7)

5. Were you able to understand grammatical structures in the text?

6. Were you able to understand vocabulary in the text?

7. Were you able to infer the central idea and purpose of the text?

8. Were you able to understand the interrelation between the passages in the text?

Scale Setting

We adopted the following four-point Likert scale: (1) Didn't comprehend at all, (2) Didn't comprehend very much, (3) Comprehended somewhat, and (4) Comprehended very much. A four-point scale was used because we imagined that the test-takers might have a higher tendency to choose "(3) Undecided" to avoid dichotomous judgment if we adopted a five-point Likert scale.

Participants

The test-takers consisted of 136 university students learning English as a foreign language in Japan: 115 Japanese students and 21 foreign students. They were from five different universities and had different majors in various departments. The foreign students were originally from seven different countries: China (8), Bangladesh (4), Indonesia (3), Nepal (3), Myanmar (1), Honduras (1), and the Philippines (1). All the test-takers participated in our study voluntarily.

Test Instruments

Two TOEIC® practice tests were reproduced from two TOEIC® official preparation books: the old version of TOEIC® (TOEIC® Steering Committee, 2000) and the new version (TOEIC® Steering Committee, 2008). TOEIC® practice tests in the official preparation books contain the same number of items, overall structure, and quality as the actual TOEIC® tests administered at the test sites. Therefore, the test data obtained in our study can be regarded as identical to that obtained from an actual TOEIC® test.

TOEIC® and the Self-assessment Questionnaire Administration

Two versions of the TOEIC® were administered to the test-takers at four different universities, with an interval of one to two weeks in November and December 2008. All the test takers took the old version first, and then the new one. Immediately after the TOEIC® test, the participants took the self-assessment questionnaire in Japanese, which took one to two minutes for them to complete. The counterbalance method was

not applied in the administration of the two TOEIC® tests because of the limitations of test administration at the universities where the TOEIC® tests were conducted. However, it should be noted that another study in our research project indicated that there was no statistically significant order effect on the participants' test performance on the two versions of the TOEIC® (Norizuki, Ito & Shimatani, 2010).

Data Analysis

All the test data were analyzed using *ANOVA 4 on the Web* (Kiriki, 2002) and *BLACK BOX* (Aoki, 2002). The data analyses include (1) basic statistics, (2) analysis of variance (ANOVA), (3) factor analyses (exploratory), and (4) correlation studies.

Alpha Level Setting

Because the data does not show a normal distribution in some parts, it needs a severe treatment of the statistical analyses. Therefore, the alpha level for all statistical decisions was set at $\alpha = 0.01$.

Results and Discussion

Table 2 presents the mean scores and standard deviations of sections in the old and new TOEIC®. In authentic TOEIC® tests, test-takers' raw scores are converted to a scaled score ranging from 10 to 990, on the basis of the Rasch model and other statistical analyses. In the present investigation, however, raw scores are utilized for the following statistical analyses. This is because the present study is thoroughly based on classical test theory.

Table 2

Fundamental Statistics of Two Versions of TOEIC® (N=136)

	Old TOEIC®			New TOEIC®		
	Listening	Reading	Total	Listening	Reading	Total
	Section	Section		Section	Section	
Mean	55.71	51.49	107.21	68.22	50.77	118.99
SD	12.02	12.63	23.22	12.08	13.23	23.27

There is a high correlation coefficient between the old and new TOEIC® ($r = 0.830$, $p < 0.01$), indicating that the two tests are highly interrelated for measuring the same construct of English language proficiency, despite the fact that they differ in some points in terms of their test format.

In order to find out whether there is a statistically significant difference between the mean scores, a two-way ANOVA with a two-by-two factorial design (old and new versions $\times$ Listening and Reading sections) was performed. The results indicated that the interaction of TOEIC® Version and Section was statistically significant ($F = 190.89$, $p < 0.01$). A statistical test of simple main effect also indicated that the difference of version had *no* effect on Reading Sections. Therefore, we can say that there is *no* significant difference of mean scores between the two TOEIC® Reading sections (Old TOEIC® Reading section: $M = 51.49$; New TOEIC® Reading section: $M = 50.77$) ($F = 0.53$, $p = 0.49$, *n.s.*). On the other hand, another simple main effect test indicated that the difference of version had a statistically significant effect on Listening sections: The mean score of the new TOEIC® listening section is higher than that of its counterpart in the old version (New TOEIC® Listening section: $M = 68.22 >$ Old TOEIC® Listening section: $M = 55.71$) ($F = 265.88$, $p < 0.01$).

Table 3***Self-assessment Questionnaire for Old TOEIC® (N=136)***

Item	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Total
Mean	2.75	2.28	2.45	1.88	2.49	2.24	2.39	16.48
SD	0.62	0.67	0.70	0.69	0.67	0.65	0.67	3.35

Table 4***Self-assessment Questionnaire for New TOEIC® (N=136)***

Item	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Total
Mean	3.12	2.55	2.68	2.13	2.47	2.25	2.35	2.19	19.75
SD	0.63	0.67	0.71	0.71	0.69	0.64	0.76	0.82	4.13

Tables 3 and 4 present the basic statistics of the self-assessment questionnaires. The participants' mean scores varied from 1.88 to 2.75 in the self-assessment questionnaire for Old TOEIC®. In addition, the mean scores in the self-assessment questionnaire for the new version varied from 2.13 to 3.12. The values of those mean scores and SDs show that neither a ceiling effect nor a floor effect emerged in the present study.

In order to assess the participants' perceived difficulty of different versions and sections of TOEIC® test items, summary measures of these ratings were calculated. Table 5 presents the mean ratings and standard deviations of questionnaire items in the respective section groups. A two-way ANOVA with a two-by-two factorial design (old and new versions $\times$ Listening and Reading sections) was performed. The results indicated that the interaction of TOEIC® Version and Section was statistically

significant ($F = 46.83, p < 0.01$). A statistical test of simple main effect indicated that the difference of version had some effect on Listening sections ($F = 55.51, p < 0.01$) but no effect on Reading sections ($F = 1.46, p = 0.23, n.s.$). Therefore, we can say that the participants felt that the New TOEIC® Listening section was easier than the old one. However, they did not feel differently about the two Reading sections. The results of the ANOVA imply that the participants' post-test-taking attitudes parallel their test performance because the participants' test scores in listening sections differed between the two versions of TOEIC® (Listening section in Old TOEIC® < Listening section in New TOEIC®; Reading section in Old TOEIC® = Reading section in New TOEIC®). This shows that the self-assessment questionnaires are valid indicators of the TOEIC® scores in the two versions.

Table 5

Mean Scores and Standard Deviations of Questionnaire Items in the Respective Section Groups (N=136)

	Old TOEIC®		New TOEIC®	
	Listening	Reading	Listening	Reading
	Section	Section	Section	Section
Mean	2.34	2.37	2.63	2.33
SD	0.52	0.53	0.54	0.57

Reliability Estimation of the Self-assessment Questionnaire and TOEIC®

The reliability coefficient based on internal consistency, Cronbach's alpha, was calculated for the two versions of the TOEIC® and the two self-assessment questionnaires. Both versions of the test showed high reliability coefficients (Old TOEIC®: $\alpha = 0.93$; New TOEIC®: $\alpha = 0.93$) (Ito, Shimatani, Norizuki & Kinoshita, 2009a). In addition, the self-assessment questionnaire showed a high reliability coefficient (self-assessment questionnaire for Old TOEIC®: $\alpha = 0.84$, and for New TOEIC®: $\alpha = 0.88$).

Exploratory Factor Analysis for Validation Studies on the Two Self-assessment Questionnaires

Initially, the factorability of the items was examined. Several well-recognized criteria for the factorability of a correlation were used. Firstly, all the questionnaire items in the two self-assessment questionnaires correlated at least 0.3 with at least one other item, suggesting reasonable factorability. Secondly, the Kaiser-Meyer-Olkin measure

of sampling adequacy was 0.82 in the self-assessment questionnaire for Old TOEIC® and 0.88 in the self-assessment questionnaire for New TOEIC®, above the recommended value of 0.6. Finally, the communalities were all above 0.3, which further confirmed that each item shared some common variance with others. Given these overall indicators, factor analysis was conducted with all items on each self-assessment questionnaire. Principal component analyses were used because the primary purpose of the factor analyses was to identify the factors underlying each of the two self-assessment questionnaire data sets. As seen in Table 6, in the self-assessment questionnaire for Old TOEIC®, the initial eigenvalues showed that the first to third factors explained 51.25%, 11.21%, and 9.79% of the variance, respectively.

Table 6

Summary of Exploratory Factor Analysis on Self-assessment Questionnaire for Old TOEIC®

Factor	Eigenvalue	Variance (%)	Cumulative (%)
Factor 1	3.587	51.25	51.25
Factor 2	0.785	11.21	62.46
Factor 3	0.685	9.79	72.25
Factor 4	0.646	9.24	81.49
Factor 5	0.589	8.41	89.90
Factor 6	0.401	5.72	95.62
Factor 7	0.307	4.38	100.00

Table 7

Summary of Exploratory Factor Analysis on Self-assessment Questionnaire for New TOEIC®

Factor	Eigenvalue	Variance (%)	Cumulative (%)
Factor 1	4.317	53.97	53.97
Factor 2	0.836	10.44	64.41
Factor 3	0.748	9.35	73.76
Factor 4	0.557	6.94	80.72
Factor 5	0.459	5.74	86.46
Factor 6	0.429	5.37	91.83
Factor 7	0.347	4.34	96.17
Factor 8	0.306	3.83	100.00

Moreover, Table 7 demonstrates that in the self-assessment questionnaire for New TOEIC®, the initial eigenvalues showed that the first to third factors explained 53.97%, 10.44%, and 9.35% of the variance, respectively. The initial factor analysis of each of the two self-assessment questionnaire data extracted one factor with eigenvalues greater than 1.0. On the basis of the scree plot, a single factor solution was deemed appropriate for each of the two data sets. In both cases, the first factor is greater than the second and third factors, explaining more than 50% of total variance. Consequently, any rotations were not conducted because a majority of the variance was able to be explained by a single factor. In addition, as Tables 8 and 9 show, each item in each self-assessment questionnaire showed a value of factor loading greater than 0.6. This indicates that each item correlates to the total factor equally. Therefore, it may be justifiable to say that the factor structure underlying each of the self-assessment questionnaires is unidimensional, which implies their construct validity.

Table 8

Principal Component Loadings and Communalities Based on a Principal Component Analysis for 7 Items on Self-assessment Questionnaire for Old TOEIC®

Factor	Loading	Communality
Item 1	0.694	0.481
Item 2	0.616	0.380
Item 3	0.645	0.416
Item 4	0.721	0.520
Item 5	0.687	0.472
Item 6	0.629	0.396
Item 7	0.601	0.361

Table 9

Principal Component Loadings and Communalities Based on a Principal Component Analysis for 8 Items on Self-assessment Questionnaire for New TOEIC®

Factor	Loading	Communality
Item 1	0.601	0.361
Item 2	0.719	0.517
Item 3	0.760	0.577
Item 4	0.713	0.589
Item 5	0.688	0.473
Item 6	0.606	0.367
Item 7	0.768	0.589
Item 8	0.642	0.421

Concurrent Validation of the Self-assessment Questionnaire

As a next attempt to validate the two self-assessment questionnaires, we carried out concurrent validation studies calculating the Pearson's product moment correlation coefficient between each item in the self-assessment questionnaire and each part of the TOEIC®. The correlation coefficients between the TOEIC® and the self-assessment questionnaire were fairly high (Old TOEIC®: $r = 0.738$, $p < 0.01$; New TOEIC®: $r = 0.656$, $p < 0.01$); the correlation coefficients between the TOEIC® Listening section and the self-assessment questionnaire for the Listening section were moderate (Old TOEIC®: $r = 0.624$, $p < 0.01$; New TOEIC®: $r = 0.666$, $p < 0.01$); and the correlation coefficients between the TOEIC® Reading section and the self-assessment questionnaire for the Reading section were moderate (Old TOEIC®: $r = 0.651$, $p < 0.01$; New TOEIC®: $r = 0.666$, $p < 0.01$).

Table 10

Correlation between Each Item in the Self-assessment Questionnaire and Each Part in Old TOEIC®

	Part I	Part II	Part III	Part IV	Part V	Part VI	Part VII
Item 1	0.47	0.41					
Item 2			0.41	0.41			
Item 3	0.39	0.43					
Item 4			0.49	0.43			
Item 5					0.58	0.51	0.52
Item 6					0.50	0.34	0.48
Item 7					0.46	0.41	0.41

Table 11

Correlation between Each Item in the Self-assessment Questionnaire and Each Part in New TOEIC®

	Part 1	Part 2	Part 3	Part 4	Part 5	Part 6	Part 7	Part 8
Item 1	0.22	0.26						
Item 2			0.47	0.45				
Item 3	0.37	0.50						
Item 4			0.42	0.51				
Item 5					0.51	0.46	0.41	0.29
Item 6					0.42	0.35	0.36	0.24
Item 7					0.46	0.36	0.41	0.24
Item 8					0.43	0.34	0.46	0.49

Table 10 shows the correlations between each part of Old TOEIC® and the self-assessment questionnaire that followed it. Table 11 shows the correlations between

each part of New TOEIC® and the self-assessment questionnaire that followed it. Presented are the values of correlation between the self-assessment questionnaire item and the TOEIC® parts each item in the self-assessment questionnaire is intended to measure. The results indicate that there are low to moderate correlations between each part of the TOEIC® and the item self-assessment questionnaire concerned, implying that the two self-assessment questionnaires are valid indicators of TOEIC® test scores. It is, however, noteworthy that there is a very low correlation between Item 1 in the self-assessment questionnaire for New TOEIC® and Part 1 in New TOEIC®. Previous studies implied that Part 1 in New TOEIC® was an easier part that required less discrimination ability than other parts in the Listening section. In order to clarify the relation between the self-assessment and the test performance on Part 1 in New TOEIC®, a scatter plot was produced to display values for the two variables for a set of data. The scatter plot indicated the participants' test performance on Part 1 was high, but their self-assessment varied from low to high. Further research should address why the participants had a tendency to underestimate their test performance in Part 1 of New TOEIC®. Also, future research should consider the effects of (1) chance score and (2) confidence level in responding to TOEIC® and the self-assessment questionnaire.

The research findings of the present study showed that the two versions TOEIC® are equally reliable and valid tests and that the self-assessment questionnaires constructed on the basis of the SDT worked well as indicators of TOEIC® test scores, although there were a few low correlation coefficients between Parts 1 and 8 in New TOEIC® and the relevant items in the self-assessment questionnaire.

Conclusion

The present study reported the construction and validation processes of self-assessment questionnaires for the old and new versions of TOEIC®. The participants were 136 university students learning English as a foreign language at Japanese universities. The research findings revealed that the data gathered through the two self-assessment questionnaires accurately reflect the participants' TOEIC® scores. In addition, the self-assessment questionnaires showed high reliability, satisfactory construct validity, and concurrent validity. Given these findings, it seems acceptable to say that TOEIC® is a valid standardized test, not only on the basis of research findings from a series of previous studies, but also from the perspective of data on test-takers' post-test-taking psychological attitudes, gathered in the present investigation. One of the limitations of the present investigation is its sampling. The results can be generalized only to Japanese-dominant groups of students. However, many studies have used participants of various language backgrounds, ages, and educational backgrounds. As a result, findings have often been hard to interpret because they can only be generalized to the single situation in which the data was collected. In addition, the results of this study may be influenced by certain internal and external factors: (1) the nature of reliability

of measures in general and (2) restrictions in the range of ability that was sampled in the investigation. Generally, tests are not simply reliable and valid, but they can be reliable and valid for specific types of students with specific ranges of ability. In consideration of this, the present study should be replicated with a larger sample of participants from a much wider population.

Acknowledgements

We would like to thank Prof. Chizuko Suzuki at Nagasaki Junshin University and Mr. Mineo Mitsuhashi at the Institute for International Business Communication for their cooperation. We are also grateful to the 136 university students who participated in our research project. The project has been supported by the TOEIC® research fund. An earlier version of this paper was delivered at the 2010 International Conference on Applied Linguistics at National Chiayi University, Taiwan in November 2010. We would like to thank the participants for their helpful comments.

References

- Aoki, S. (2002). BLACK BOX. Available from <http://aoki2.si.gunma-u.ac.jp/>
- Duke, T. (2006). Modifications and features of the new TOEIC® test. *TOEIC® Newsletter*, 92, 4–7.
- Educational Testing Service. (2005). *The redesigned TOEIC® test: Rationale, evidence, and direction*. Educational Testing Service.
- Educational Testing Service. (2007). Score descriptor table. Available from http://www.toeic.or.jp/toeic_en/toeic/score_descriptors.html
- Ito, A., Shimatani, H., Norizuki, K., & Kinoshita, M. (2009a). Two versions of the TOEIC® test: A comparative study based on the classical test theory. *JLTA Journal*, 12, 26–45.
- Ito, A., Shimatani, H., Norizuki, K., & Kinoshita, M. (2009b). Two versions of the TOEIC® test: A correlation study on examinees' placement and internal construct validity. *Studies in English Teaching and Learning in East Asia*, 3, 5–17.
- Ito, A., Shimatani, H., & Norizuki, K. (2010). Two versions of the TOEIC® test: A predictive validation study. *JLTA Journal*, 13, 111–125.
- Kiriki, K. (2002). ANOVA 4 on the Web. Available from <http://www.hju.ac.jp/~kiriki/anova4/>
- Liao, C. (2010). TOEIC® listening and reading test scale anchoring study. Available from <http://www.ets.org/Media/Research/pdf/TC-10-05.pdf>
- Norizuki, K., Ito, A., & Shimatani, H. (2010, September). *Shinkyuu TOEIC® test no hikakukenshou --- 4 test set 800 mon no juken data bunseki* [A comparative analysis of 800 items from 4 TOEIC® tests]. A paper presented at the 14th JLTA Annual Conference, Toyohashi University of Technology.

- Norizuki, K., Ito, A. & Shimatani, H. (2011). Exploring item-examinee response characteristics in search of diagnostic functions of TOEIC® tests for university students in Japan. *JLTA Journal*, 14, 1-20.
- Powers, D. E., Kim, H-J., Yu, F., Weng, V. Z., & VanWinkle, W. H. (2009). The TOEIC® speaking and writing tests: Relations to test-taker perceptions of proficiency in English. Available from http://www.ets.org/research/policy_research_reports/rr-09-18
- Shimatani, H., Norizuki, K., Ito, A. & Kinoshita, M. (2009a). Two versions of the TOEIC® test: A comparative study based on the Rasch Model. *The JACET Kyushu-Okinawa Chapter Annual Review of English Learning and Teaching*, 14, 17-32.
- Shimatani, H., Norizuki, K., Ito, A. & Kinoshita, M. (2009b). Two versions of TOEIC® as a placement test: A Comparative study based on the classical test theory and Rasch model. *English Language Assessment*, 3, 39-56.
- TOEIC® Steering Committee. (2000). *TOEIC® test official preparation book (Japanese version)*. Tokyo: The Institute for International Business Communication.
- TOEIC® Steering Committee. (2006). The new TOEIC® test. *TOEIC® Newsletter*, 92, 1-3.
- TOEIC® Steering Committee. (2008). *TOEIC® test new official preparation book, vol. 3 (Japanese version)*. Tokyo: The Institute for International Business Communication.

Appendix: Score Descriptor Table

Score Descriptor Table		
Listening Section		
Score	Strength	Weakness
495	• They can infer the central idea, purpose, and basic context of short spoken exchanges across a broad range of vocabulary, even when conversational responses are indirect or not easy to predict.	Test takers who receive a score at this level typically have weaknesses only when uncommon grammar or vocabulary is used.
~		
375	• They can infer the central idea, purpose, and basic context of extended spoken texts across a broad range of vocabulary. They can do this even when the information is not supported by repetition or paraphrase and when it is necessary to connect information across the text.	

	<ul style="list-style-type: none"> • They can understand details in <i>short</i> spoken exchanges, even when negative constructions are present, when the language is syntactically complex, or when difficult vocabulary is used. • They can understand details in <i>extended</i> spoken texts, even when it is necessary to connect information across the text and when this information is not supported by repetition. They can understand details when the information is paraphrased or when negative constructions are present. 	
	<ul style="list-style-type: none"> • They can sometimes infer the central idea, purpose, and basic context of <i>short</i> spoken exchanges, especially when the vocabulary is not difficult. • They can understand the central idea, purpose, and basic context of <i>extended</i> spoken texts when this information is supported by repetition or paraphrase. • They can understand details in <i>short</i> spoken exchanges when easy or medium-level vocabulary is used. • They can understand details in <i>extended</i> spoken texts when the information is supported by repetition and when the requested information comes at the beginning or end of the spoken text. They can understand details when the information is slightly paraphrased. 	<ul style="list-style-type: none"> • They have difficulty understanding the central idea, purpose, and basic context of short spoken exchanges when conversational responses are indirect or difficult to predict or when the vocabulary is difficult. • They do not understand the central idea, purpose, and basic context of <i>extended</i> spoken texts when it is necessary to connect information within the text or when difficult vocabulary is used. • They do not understand details in <i>short</i> spoken exchanges when language is syntactically complex or when difficult vocabulary is used. They do not usually understand details that include negative constructions. • They do not understand details in <i>extended</i> spoken texts when it is necessary to connect information across the text or when the information is not supported by repetition. They do not understand most paraphrased information or difficult grammatical constructions.

270 ~5	<ul style="list-style-type: none">• They can understand <i>short</i> (single-sentence) descriptions of the central idea of a photograph.• They can sometimes understand the central idea, purpose, and basic context of <i>extended</i> spoken texts when this information is supported by a lot of repetition and easy vocabulary.• They can understand details in <i>short</i> spoken exchanges and descriptions of photographs when the vocabulary is easy and when there is only a small amount of text that must be understood.• They can understand details in <i>extended</i> spoken texts when the requested information comes at the beginning or end of the text and when it matches the words in the spoken text.	<ul style="list-style-type: none">• They do not understand the central idea, purpose, or basic context of <i>short</i> spoken exchanges, even when the language is direct and no unexpected information is present.• They do not understand the central idea, purpose, and basic context of <i>extended</i> spoken texts when it is necessary to connect information across the text or when the vocabulary is somewhat difficult.• They do not understand details in <i>short</i> spoken exchanges when somewhat difficult vocabulary is used or when the language is syntactically complex. They do not understand details that include negative constructions.• They do not understand details in <i>extended</i> spoken texts when the requested information is heard in the middle of the text. They do not understand paraphrased information or difficult grammatical constructions.
	*Test takers who score below 170 may have some of the same strengths as test takers who score around 200, but their performance is likely to be less consistent. Test takers who score around 200 <i>typically</i> have the strengths described in the box above.	

Reading Section

Score	Strength	Weakness
495 ~ 425	<ul style="list-style-type: none">• They can infer the central idea and purpose of a written text, and they can make inferences about details.• They can read for meaning. They can understand factual information, even when it is paraphrased.• They can connect information across an entire text, and they can make connections between two related texts.• They can understand a broad range of	Test takers who score around 450 <i>typically</i> have weaknesses only when the information tested is particularly dense or involves difficult vocabulary.

vocabulary, unusual meanings of common words, and idiomatic usage. They can also make distinctions between the meanings of closely related words.

- They can understand rule-based grammatical structures. They can also understand difficult, complex, and uncommon grammatical constructions.

- They can infer the central idea and purpose of a written text, and they can make inferences about details.

- They can read for meaning. They can understand factual information, even when it is paraphrased.

- They can connect information across a small area within a text, even when the vocabulary and grammar of the text are difficult.

- They can understand medium-level vocabulary.

They can sometimes understand difficult vocabulary in context, unusual meanings of common words, and idiomatic usage.

- They can understand rule-based grammatical structures. They can also understand difficult, complex, and uncommon grammatical constructions.

- They do not connect information across a wide area within a text.

- They do not consistently understand difficult vocabulary, unusual meanings of common words, or idiomatic usage. They usually cannot make distinctions between the meanings of closely related words.

420

~

325

320 ~ 225	<ul style="list-style-type: none"> • They can make simple inferences based on a limited amount of text. • They can locate the correct answer to a factual question when the language of the text matches the information that is required. They can sometimes answer a factual question when the answer is a simple paraphrase of the information in the text. • They can sometimes connect information within one or two sentences. • They can understand easy vocabulary, and they can sometimes understand medium-level vocabulary. • They can understand common, rule-based grammatical structures. They can make correct grammatical choices, even when other features of language, such as difficult vocabulary or the need to connect information, are present. 	<ul style="list-style-type: none"> • They do not understand inferences that require paraphrase or connecting information. • They have a very limited ability to understand factual information expressed as a paraphrase using difficult vocabulary. They often depend on finding words and phrases in the text that match the same words and phrases in the question. • They usually do not connect information beyond two sentences. • They do not understand difficult vocabulary, unusual meanings of common words, or idiomatic usage. They usually cannot make distinctions between the meanings of closely related words. • They do not understand more difficult, complex, or uncommon grammatical constructions.
220 ~5	<ul style="list-style-type: none"> • They can locate the correct answer to a factual question when not very much reading is necessary and when the language of the text matches the information that is required. • They can understand easy vocabulary and common phrases. • They can understand the most-common, rule-based grammatical structures when not very much reading is necessary. 	<ul style="list-style-type: none"> • They cannot make inferences about information in written texts. • They do not understand paraphrased factual information. They rely on matching words and phrases in the text to answer questions. • They are often unable to connect information even within a single sentence. • They understand only a limited range of vocabulary. • They do not understand even easy grammatical constructions when other language features, such as difficult vocabulary or the need to connect information, are also required.

*Test takers who score below 120 may have some of the same strengths as test takers who score around 150, but their performance is likely to be less consistent. Test takers who score around 150 *typically* have the strengths described in the box above.