Comparability of TOEFL ITP and TOEIC IP: Analyzing Test Materials and Scores

Siwon Park

Yasuko Ito

Megumi Sugita

Ruriko Tsuji

Abstract

This study investigated the comparability of TOEFL ITP and TOEIC IP by analyzing test scores and test materials. We examined official test materials of TOEFL ITP and TOEIC IP to identify their content, cognitive and linguistic characteristics such as text type/genre, linguistic complexity, and cognitive operations that the tests intend to sample. We also collected score data of TOEFL ITP and TOEIC IP from 407 students who voluntarily took these tests within two weeks. The regression analyses with the data enabled us to prepare a score conversion table of TOEFL ITP and TOEIC IP. The findings indicated that 1) TOEIC ITP introduces more life-related situations examinees are likely to encounter daily than TOEFL ITP does, 2) TOEFL ITP requires applications of different strategies in answering questions, while TOEIC IP is rather limited to skills to get main ideas and details, and 3) TOEIC IP reading texts present more lexical diversity than those of TOEFL ITP. In addition, the statistical analyses revealed that TOEFL ITP and TOEIC IP may well be considered equivalent when the score range mainly concerns the low to high-intermediate levels of L2 English proficiency.

Introduction

Every year, an increasing number of institutions in Japan are adopting TOEFL and/or TOEIC for their assessment purposes, and their scores are often compared and referred to as equivalent especially in standards setting. However, it is not clear to what extent such

equivalence can be justified as the information is scarce of how the score conversion between TOEFL and TOEIC was conducted.

Comparing two or more tests for their construct equivalence has been rather discouraged in recent years, which is one reason why Educational Testing Service (ETS) has abandoned providing their earlier equation for the score conversation between the TOEFL and TOEIC tests (ETS, 2003). Yet, as the popularity of comparability studies attests (especially in the form of the benchmarking studies), the practical needs in the educational institutions for the adoption of multiple assessment options have rather grown especially for the purpose of certification and the standards setting across different language programs. Therefore, it is not too surprising to find a score/grade conversion table between tests such as TOEFL, TOEIC, IELTS, and/or STEP EIKEN.

Score conversion directly concerns test equivalence, the argument of which can only be valid when the constructs of the two tests are demonstrated equivalent (APA, 1986; AERA, APA, & NCME, 1999). Hence, test comparability entails not only statistical equivalence, but also the equivalence of content, linguistic demands, and cognitive operations that the tests intend to sample. Only after considering these multi-faceted aspects of the test equivalence, one could draw a valid argument of score equivalence.

With regards to the comparability between TOEFL and TOEIC, it is rare to find a comparability study between them (e.g., 土肥・張, 2014) for the reason mentioned earlier. Still, it is easy to find conversion tables between the same tests of different formats such as TOEFL ITP (Institutional Testing Program), CBT, and *i*BT (e.g., ETS, 2005). Rather popular are standard setting studies (Tannenbaum & Wylie, 2005, 2008) that consider the comparability of TOEFL and TOEIC and mapped their scores onto the Common European Framework (Council of Europe, 2001). Being placed onto the common framework, the scoring meanings of TOEFL and TOEIC become more or less comparable.

TOEFL ITP and TOEIC IP

While regular TOEFL tests, such as TOEFL *i*BT, are to be taken by individuals, TOEFL ITP is a test administered on an institutional basis. Its format is similar to TOEFL PBT, which is a paper-based version of TOEFL. TOEFL ITP consists of three sections, namely listening comprehension, structure and written expression, and reading comprehension, which contains 50 questions, 40 questions, and 50 questions respectively, as shown in Table 1.

Table 1

Number of Questions in Each Section of TOEFL ITP

Section	Number of questions
Listening Comprehension	50
Structure and Written Expression	40
Reading Comprehension	50

Topics in TOEFL ITP include academic, campus-life, and general ones, as presented in Table 2.

Table 2

Topics, Contents, and Settings in TOEFL ITP

Topic	Content	Setting
Academic	Arts, Humanities, Life Sciences, Physical Sciences, Social Sciences	Craft, Dance, History, Political Science, Biochemistry, Animal Behavior, Anthropology, Sociology, etc.
Campus-life	Classes, Campus Administration, Campus Activities	Class Schedule, Class Requirement, Registration, Housing, Study Abroad, Club, Committee, etc.
General	Business, Environment, Food, Language and Communication, Media, Personal, Purchases, Recreation, Transportation, etc.	Law, Weather, Nature, Restaurants, Telephone Use, TV, Health, Shopping, Sports, Travel, etc.

TOEIC IP (Institutional Program), like TOEFL ITP, is also an institutional version of TOEIC. Although a new version has been adopted in TOEIC since May 2015, an institutional version of TOEIC has always utilized the older version and thus, the test materials in the current study came from the older version. The test consists of two sections, listening and reading. Table 3 provides the details of the two sections.

Table 3

Number of Questions in Each Section of TOEIC IP

Part	Types of questions	Number of questions
Listening section (45 min.)		
1	Photographs	10
2	Question-Response	30
3	Short Conversations	30
4	Short Talks	30
	Total	100
Reading section (75 min.)		
1	Incomplete Sentences	40
2	Text Completion	12
3	Reading Comprehension	48
	Total	100

Compared to TOEFL ITP, TOEIC IP has a relatively larger number of questions in both listening and reading sections.

Purpose of research

The present study aims to examine the comparability of TOEFL ITP and TOEIC IP by analyzing test scores and test materials. We examined official test materials of TOEFL ITP and TOEIC IP to identify their content, cognitive and linguistic characteristics such as text type/genre, cognitive operations, and linguistic complexity that the tests intend to sample. We also collected score data of TOEFL ITP and TOEIC IP from university students who voluntarily took these tests within two weeks.

Method

Sources of the data and participants

There are two sets of data in this study. One is students' test scores on TOEFL ITP and TOEIC IP, and the other is test materials that were collected from different sources.

The first set of the data was students' scores on TOEFL ITP and TOEIC IP tests administered between May 2012 and October 2015. Both tests were taken by 407 university students who were majoring in English in Japan. 87% of them were either freshmen or sophomores, and 71% of them were female. Their purposes of taking the tests include, to study abroad, to fulfill course requirements, to apply for a scholarship, and to check their progress in learning English.

These students took both tests with an interval of less than two weeks. The test taking was voluntary and the students themselves paid fees to take the tests. They had taken the tests before, i.e., it was not their first trial, and therefore, they were familiar with the tests to varying extent at the time of the data collection.

The second set of the data, test materials, were used to examine the characteristics of the two tests. Materials used for the TOEFL ITP analysis were *Official Guide to the TOEFL ITP® Test* (2 sets), *TOEFL ITP® Practice Tests, Volume 1* (2 sets), and *TOEFL ITP®テスト公式テスト問題&学習ガイド* (1 set). Those used for TOEIC IP analysis were all taken from *YBM Official TOEIC Practice Book* (6 sets). All test sets were previously used by ETS, and their publication is officially endorsed by ETS.

Analysis

Test materials were analyzed in three aspects: content, cognitive, and linguistic aspects. Content analysis focused on the genre of listening and reading texts. The cognitive analysis

looked at the types of questions. Finally, the linguistic analysis involved lexical analysis, or lexical density to be more specific, of the texts used in the two tests.

In the first two analyses, content and cognitive analyses, the materials summarized in Table 4 were used.

Table 4

Materials Used for Content and Cognitive Analyses

	TOEIC	TOEFL
Number of test sets	2	2
Listening	Part 3 & 4	Part B & C
Reading	Part 7	Reading section

TOEIC Part 3 is to listen to conversations between two people, TOEIC Part 4 is to listen to talks by a single speaker, and TOEIC Part 7 is to read various types of texts, for example, magazines, newspaper articles, letters, and advertisements. In the TOEFL test, on the other hand, Part B is to listen to longer conversations between two people, Part C is to listen to lectures/talks by a single speaker, and the Reading section is to read academic texts. For the content analysis, the genre of listening texts and reading passages were analyzed to examine what contents the test takers are expected to comprehend. The cognitive analysis looked at the test questions using the following categories in Table 5. These categories were taken from *Official Guide to the TOEFL ITP Tests*.

Table 5

Categories Used in Cognitive Analysis

	Listening	Reading
Type 1	Gist Questions	Main Ideas
Type 2	Detail Questions	Factual Information
Type 3	-------	Organization and Logic
Type 4	-------	Referential Relationship
Type 5	-------	Vocabulary in Context
Type 6	-------	Inference

For the linguistic analysis, a different number of test sets were used, as summarized in Table 6.

Table 6

Materials Used for Linguistic Analysis

	TOEFL Reading	TOEIC Reading (Part 7)
Number of test sets	5	6
Number of passages per test set	5	13

The linguistic analysis involved the lexical diversity of the listening texts and reading passages in both tests. The *vocd* function was utilized which is available on the Computerized Language Analysis (CLAN) programs, part of the Child Language Data Exchange System (CHILDES) (MacWhinney, 2000), and the parameter D values and TTR were calculated and examined.

For the statistical analyses, test scores of TOEFL ITP and TOEIC IP collected from 407 students were subjected to a series of statistical analyses for the statistical comparability of the two tests, such as correlations, regression, and confirmatory factor analysis (CFA).

Results and Discussion

Content analysis

The content analysis revealed that the TOEIC IP tests provide more texts on daily tasks than TOEFL ITP, for example, discussing where to go for lunch, an announcement at a department store, a phone conversation at a hotel, and a guest's inquiry at the front desk about room service. However, we should interpret this result with caution because TOEFL aims to measure learners' proficiency in academic English, and thus the types of English in the two tests differ in the first place. Furthermore, the content analysis did not include TOEFL listening Part A, which consists of short conversations that are likely to take place on campus. In this regard, Part A may include relatively daily topics compared to Parts B and C of TOEFL. If Part A was included in the analysis, the result may be different.

Cognitive analysis

Listening sections contained questions which involved inference of the context or the speaker(s). Among the categories for listening questions, however, there was no "inference" category as shown in Table 5, and there were only two types of questions, gist and details. To remedy this problem, when one question was identified as "inference," it was coded as "Type 2 (detail) and Inference." The reason why it was coded as "Type 2 and Inference," instead of simply creating another category of "Inference" as Type 3, was that the inference questions often entail the inference about the details. It should be pointed out, though, that the inference questions in listening sections were slightly different from those in the reading

sections. In the reading sections, particularly in TOEFL reading, the question was, for example, “What can be inferred from the text?” while in listening, it was “What will the man probably do next?” Figure 1 illustrates the result of the cognitive analysis for listening sections.

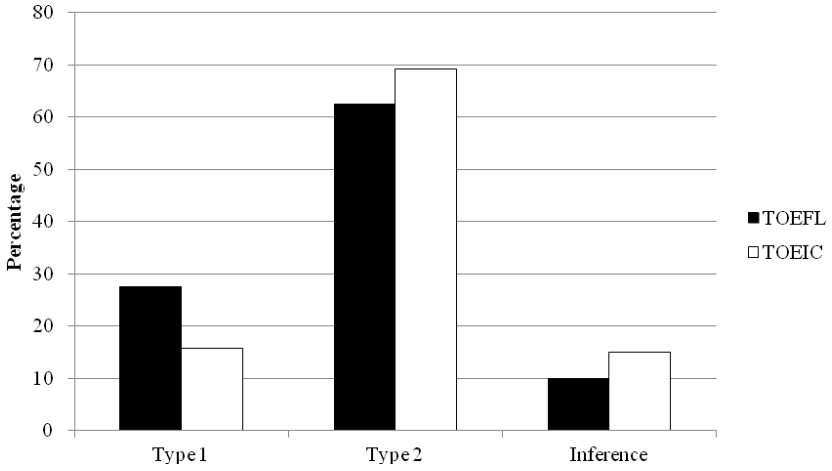

Figure 1. Cognitive analysis of listening sections

It is apparent that most of the questions ask about details in both TOEFL ITP and TOEIC IP tests. This is inevitable because one listening text is designed to have only one gist, and when there are more than one question for one text, only one of them can ask about the gist while the other may ask about the details.

The analysis of reading sections reveals a slightly different picture.

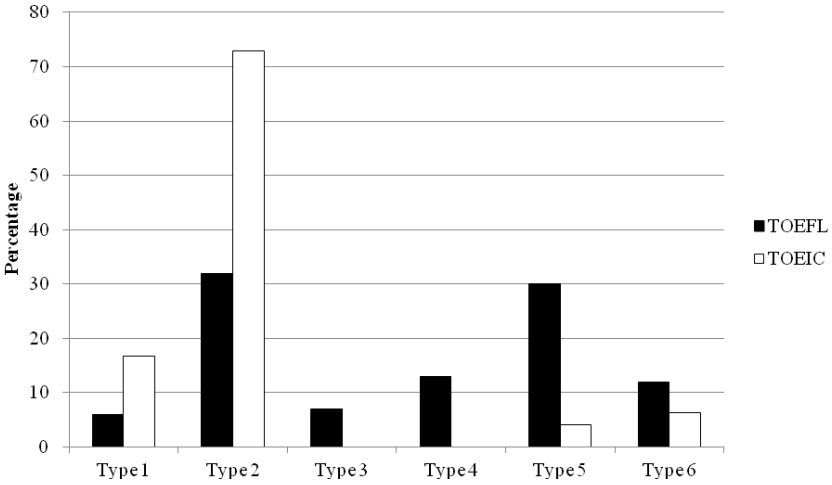

Figure 2. Cognitive analysis of reading sections

As shown in Figure 2, TOEFL ITP contains various types of questions, with Type 2 (Factual Information) and Type 5 (Vocabulary in Context) as its majority, while most of the questions in the TOEIC IP reading section were Type 2, with no question of Type 3 (Organization and Logic) and Type 4 (Referential Relationship). One possible explanation for this result is that TOEIC IP reading materials are shorter than those in TOEFL ITP and they are not long enough to ask structural details of the texts. The difference between the two reading tests in terms of the content, or the topic, may also account for this trend. TOEIC IP reading texts often include specific information needed for daily life and business purposes, and they are expected to provide such information in a straightforward manner, rather than an implied manner.

Lexical analysis

As part of the linguistic analyses, the text data were examined for their lexical characteristics. First, the number of passages and items in each test was examined, and the lexical diversity was examined in terms of D and TTR. Table 7 presents the results.

Table 7

Lexical Diversity (TOEFL Section 3 and TOEIC Part 7)

	TOEFL ITP	TOEIC IP
Passage/test	5	13
Item/test	50	48
Item/passage	10 (7-12)	3.1 (2-5)
Types/passage	191	122
Tokens/passage	342	193
D	91.47	114.02
TTR	0.56	0.63

Note that the texts (i.e., passages) under consideration all come from Section 3 of TOEFL ITP and Part 7 of TOEIC IP, both of which are to assess English reading skills. While more items are provided in TOEFL, TOEIC included more passages indicating that the reading passages of TOEIC are much shorter than those of TOEFL, which in turn affects the lexical density. TOEIC included a variety of shorter texts of differing topics compared to TOEFL, leading TOEIC reading texts to present a higher lexical density as revealed by higher D as well as TTR values.

Statistical analyses

A series of statistical analyses were performed to examine the statistical comparability of the two tests, and Table 8 shows descriptive statistics of the scores of TOEFL and TOEIC taken by the students.

Table 8

Descriptive Statistics of TOEFL and TOEIC Score Data

	TOEFL				TOEIC		
	LC	GR	RC	Total	LC	RC	Total
Mean	47.78	45.16	45.44	461.27	335.19	254.45	589.64
Median	48	45	45	457	330	245	575
SD	3.81	5.19	5.12	38.20	56.22	69.18	114.86
Min.	33	31	31	353	145	95	285
Max.	62	64	60	593	495	445	915
Range	29	33	29	240	350	350	630

As the raw responses to the tests were not available, we were not able to examine the reliability aspects of the tests; yet, the descriptive statistics confirm that both TOEFL and TOEIC data are suitable for parametric analyses in terms of their distribution.

Correlation analysis

Following the distributional considerations of the data, correlations were examined across different sections of TOEFL and TOEIC tests. The results are presented in Table 9.

Table 9

Correlations

		1	2	3	4	5	6
TOEFL	1. Listening	1.00	-	-	-	-	-
	2. Grammar	.503**	1.00	-	-	-	-
	3. Reading	.441**	.568**	1.00	-	-	-
	4. Total	.746**	.867**	.835**	1.00	-	-
TOEIC	5. Listening	**.632****	*.438***	.422**	.588**	1.00	-
	6. Reading	*.575***	*.680***	**.623****	.768**	.674**	1.00
	7. Total	.656**	.623**	.581**	.750**	.897**	.932**

Note. ** Correlations significant at $p < 0.01$.

Table 9 shows correlation coefficients between the test variables within and across the two tests. First, correlation coefficients vary significantly from the lowest, 0.441 to the highest, 0.932. The correlations between the same traits with different methods are represented in bold, and the correlations of the same method but different traits are underlined. Those italicized correlations represent coefficients of different traits and different methods.

The correlations in bold which represent the convergent validity are all significantly apart from zero, and these values represent the same trait measured by different methods. Most of the observed values are relatively high to argue for presence of convergent validity for traits across methods. Yet, the italicized correlations that measured different traits with different methods are found to be relatively low, except the one between TOEIC reading and listening sections. Therefore, while the TOEFL sections are sufficiently divergent in

terms of the assessment of differing skills, the same argument cannot be made as to the two sections of TOEIC listening and reading.

Finally, the bivariate correlation between TOEFL ITP and TOEIC IP resulted in 0.750, which is a similar finding to those of prior studies' (e.g., 土肥・張, 2014).

Regression analysis

The bivariate relationship was examined using the simple regression analysis once again, and the scatter plot of the data between TOEFL ITP and TOEIC IP is presented in Figure 3.

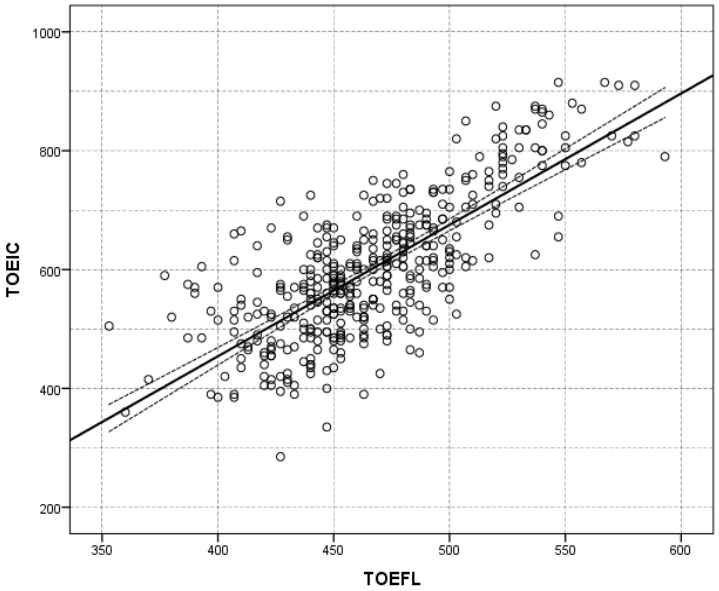

Figure 3. The Regression scatter plot of TOEFL and TOEIC data (TOEIC IP= -430.463 + 2.211*TOEFL ITP)

The regression equation is calculated for the score conversion from TOEFL to TOEIC, which is presented in the figure title. The line in the middle is the linear regression line, and the confidence interval is shown around it. The data points are mostly clustered between 400 and 500 of TOEFL and 400 and 800 of TOEIC. Consequently, the precision of measurement centers on the score levels. Likewise, as the line for the confidence intervals runs from the center of the data cluster to the extremes, the regression line loses its precision.

Confirmatory factor analysis

In order to overcome the weakness of bivariate correlation analyses that do not consider error terms in estimation, the relationships among test variables were examined using the confirmatory factor analyses (CFA). The use of CFA enabled us to examine the relationships of the latent as well as manifested variables with their measurement errors at the same time.

The model in Figure 4 shows the baseline model that presents the two test trait factors, TOEFL ITP and TOEIC IP on the left side and the five measurement variables on the right side, three of which come from the TOEFL sections and the rest from TOEIC. Between the two test traits, a correlation is specified, and each measurement variable is predicted by each relevant trait variable. The error terms are also specified for each measurement variable as from E1 to E5.

The analysis was performed using EQS 6.1 for Windows (Bentler, 2004). Using the covariance structure based on the observed scores, the baseline model was estimated for their model fit indices and individual factor loadings.

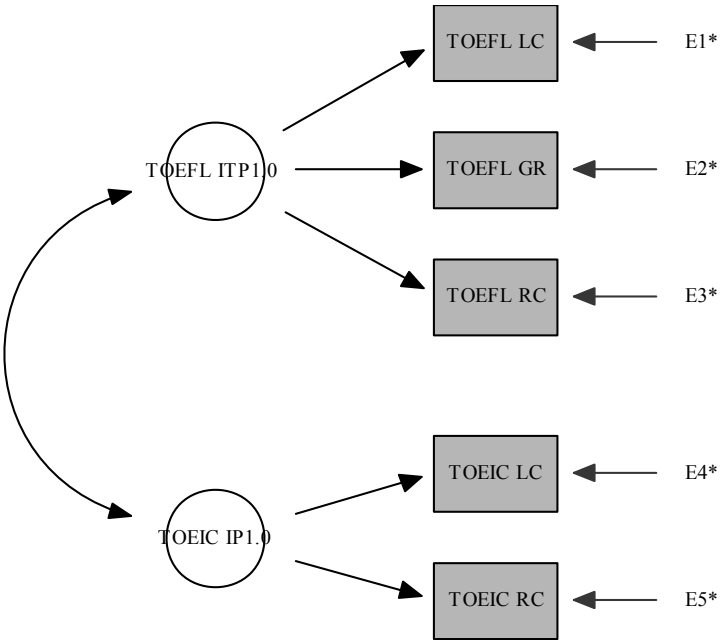

Figure 4. The baseline factor model of TOEFL and TOEIC tests

Note. LC: Listening Comprehension; GR: Grammar; RC: Reading Comprehension

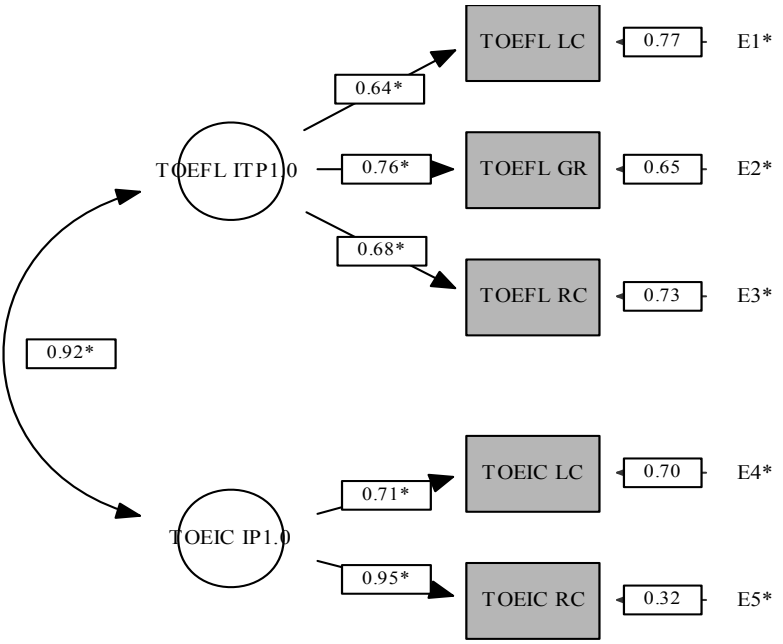

Figure 5. The baseline model with factor loadings (χ^2 (4) = 72.28. *CFI*=.948, *NFI* = .945, *RMSEA* = .053)

The baseline model in Figure 5 presents the factor loadings and model fit indices. The conventional fit indices, CFI, NFI, and RMSEA, indicate that the baseline model fits the data relatively well.

The TOEFL trait demonstrates similar factor loadings onto the three skills measurements ranging between 0.64 and 0.76, the loadings between the TOEIC trait and their corresponding measurement variables are not consistent with the loading between the TOEIC trait and the RC measurement being too high, 0.95 and the other loading with the

listening being relatively low, 0.71. This means that much of the TOEIC trait could be explained only by the TOEIC RC alone making the TOEIC LC a rather redundant in predicting examinee performance on the TOEIC test.

One reason for these inconsistent factor loadings between the TOEIC trait and its measurement variables may be due partly to the fact that TOEIC RC is a section that requires both grammar knowledge and reading skills from the examinees. Whether it to be the reason or not, this inconsistency should be considered an undesirable finding for the use of the two separate skills sections in the TOEIC test.

Another noteworthy finding on the factor structure in Figure 5 is the correlation of 0.92 between TOEFL and TOEIC. Such a high correlation coefficient indicates that the two factors share a large amount of common variance. That is, this high correlation coefficient suggests a possibility that the two test traits represented by their individual measurement variables can be considered equivalent at least statistically. Note however that this statistical equivalence does not entail the interpretive equivalence of score meanings.

In an attempt to account for this high correlational relationship between the two test traits of TOEFL ITP and TOEIC IP, the comparability of TOEFL and TOEIC should be considered with reference to their relationships in terms of common or shared variance. As Figure 6 indicates, the common variance between the two tests may be designated as General English Proficiency, and the unshared part of each test may be considered as their unique test variance that comes from their content or method aspects. In other words, test takers' performance on either of the tests must be largely comparable due to the substantial amount of the shared variance. Alternatively, the incomparability between the tests must come from their unique characteristics as represented by the specifics of TOEFL and TOEIC in Figure 6. This incomparable aspect of the two tests was examined and discussed

in terms of content, cognitive, and lexical characteristics in the earlier analysis section of this paper.

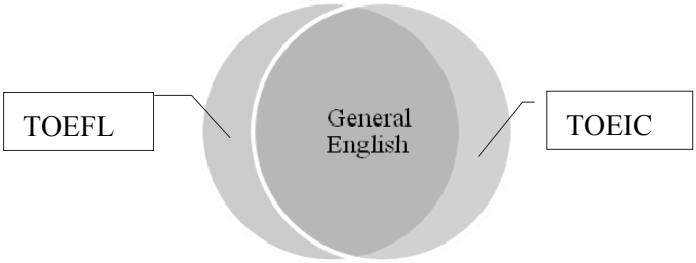

Figure 6. The common variance of TOEFL ITP and TOEIC IP

Conclusions

The current study examined how comparable TOEFL ITP and TOEIC IP are in terms of content, cognitive, and linguistic aspects through text analyses, and in their scores using statistical analyses. As the analytical findings from the earlier sections suggested, the results of the current study can be summarized as follows.

Our text analyses revealed that:

1) In terms of the content of the tests, TOEIC contains more situations that examinees are likely to encounter in daily lives than TOEFL, such as discussing where to go for lunch. However, TOEFL listening Part A, which was excluded from analysis in the current study, may also include such contexts.
2) For the cognitive aspects, TOEFL requires more strategies in answering questions, while TOEIC is rather limited to skills to get main ideas and details.
3) As for the lexical diversity, D values revealed that TOEIC IP reading texts present more lexical diversity than those of TOEFL ITP. Yet, the TOEFL ITP reading sections

may suggest more lexical challenges to test takers as they include more types and tokens per passage.

Through the statistical analyses, we also found that TOEFL ITP and TOEIC IP may well be considered equivalent, when the score range mainly concerns the low to high-intermediate levels of L2 English proficiency.

Finally, this paper reported what we have found so far in our research project as a study in-progress. As the future research avenues, we recognize that more fine-tuned text analyses are necessary especially for investigating the linguistic complexity of the texts and the question stems. In addition, for a more complete picture of the latent relationships between and among trait and measurement variables, more data from upper level students and also from different student populations need to be collected and entered to the analyses.

Acknowledgement

We would like to thank the Research Institute of Language Studies & Language Education at Kanda University of International Studies for their support for this project. In addition, we would like to express our sincere gratitude to Ms. Satsuki Tomita and Mr. Mitsuo Hirahara who helped deal with the large amount of data for our project.

References

土肥充・張智君 (2014)「千葉大学における TOEIC IP と TOEFL ITP のスコア分析と経年調査」, 『言語文化論叢』, 8, 15-32. Retrieved from http://f.chiba-u.jp/about/plc08/plc08-02.pdf.

American Psychological Association Committee on Professional Standards and Committee on Psychological Tests and Assessments (APA) (1986). *Guidelines for computer-based tests and interpretations*. Washington, DC: Author.

American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing.* Washington, DC: AERA.

Bentler, P. M. (2004). *EQS 6 structural equations program manual.* Encino, CA: Multivariate Software, Inc.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge: Cambridge University Press.

ETS (2003). *TOEFL Institutional Testing Program (ITP) and TOEIC Institutional Program (IP): Two On-Site Testing Tools from ETS at a Glance*. Handout Berlin Conference 2001. Princeton: Educational Testing Service.

ETS (2005). *TOEFL® Internet-based test: Score comparison tables.* Princeton: Educational Testing Service.

MacWhinney, B. (2000). *The* CHILDES *Project: Tools for Analyzing Talk.* 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Tannenbaum, R. J., & Wylie, E. C. (2005). *Mapping English-language proficiency test scores onto the Common European Framework* (TOEFL Research Report No. RR-80). Princeton, NJ: Educational Testing Service.

Tannenbaum, R. J., & Wylie, E. C. (2008). *Linking English-language test scores onto the Common European Framework of Reference: An application of standard setting methodology* (TOEFL iBT Series Report No. 06). Princeton, NJ: Educational Testing Service.