Exploring Item-Examinee Response Characteristics in Search of Diagnostic Functions of TOEIC® Tests for University Students in Japan

Ken Norizuki

Shizuoka Sangyo University

Akihiro Ito

Seinan Gakuin University

Hiroshi Shimatani

Kumamoto University

Abstract

TOEIC® tests have been used extensively for placement or proficiency decisions at universities in Japan. However, the amount of diagnostic information they make available to individual candidates is fairly limited. In fact, providing such TOEIC® examinees/learners with substantial formative feedback has often been deemed impractical. The purpose of this study is to help bridge the gap between the publicized results of official TOEIC® tests and additional findings which can be derived from the analyses of individual item-examinee responses to TOEIC® Practice Tests, so as to benefit both TOEIC® examinees and English learners. Different versions of TOEIC® Practice Tests, as developed by the Institute for International Business Communication (IIBC), were administered to a sample of 136 university students in Japan and a separate sample of 86, employing a common set of anchor items to calibrate all the items from different test forms on the same latent scale. A series of analyses were conducted to delineate salient characteristics of 800 items involved. The results and implications of the study were addressed in terms of diagnostic functions of TOEIC® tests for university students in Japan.

1. Introduction

TOEIC® (henceforth, TOEIC) tests have been used extensively for placement or proficiency decisions at universities in Japan. While the number of university student examinees is steadily increasing, the amount of diagnostic information available to individual candidates is limited.

In 2006 TOEIC was redesigned in order to better align itself with current theories of language proficiency and workplace language scenarios (Powers, Kim, & Weng, 2008; Schedl, 2010). Some of the major changes made in the Redesigned TOEIC, as summarized in Table 1, include: fewer photo items in Part 1; the audio presentation of Parts 3 and 4 questions also printed on the test book; a set of three questions provided for each segment of audio input in Parts 3 and 4; the adoption of various regional accents of L1 English speakers; the replacement of error correction items by text completion items; and the adoption of double-passage questions (TOEIC Steering Committee, 2006).

Table 1

Formats of Old TOEIC (OT) and New TOEIC (NT)

Old TOEIC (OT)			New TOEIC (NT)		
Part	Type of questions	N. of Questions	Part	Type of questions	N. of Questions
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
V	Incomplete Sentences	40	5	Incomplete Sentences	40
VI	Error Recognition	20	6	Text Completion	12
VII	Reading Comprehension	40	7	Reading Comprehension	
				• Single passage	28
				• Double passage	20

(Adapted from TOEIC Steering Committee, 2006)

Alongside a series of recent ETS research reports on various aspects of psychometric quality of the New TOEIC (NT) in relation to the Old TOEIC (OT) (Powers et al., 2008; Schedl, 2010; Liao, 2010; Liao, Hatrak, & Yu, 2010), several studies have been conducted by the present authors, particularly in the light of TOEIC as a placement test (Ito, Shimatani, Norizuki, & Kinoshita, 2009a; 2009b; Ito, Shimatani, & Norizuki, 2010; Shimatani, Norizuki, Ito, & Kinoshita, 2009a; 2009b, Norizuki, Ito, Shimatani, & Kinoshita, 2009; Norizuki, Ito, & Shimatani 2010), employing a pair or two pairs of OT(s) and NT(s), all of which originated from publicized TOEIC Practice Tests.

While a major concern of the present authors has been directed towards the overall reliability and validity of the two versions of the test and the comparison of the two, their attention is also accorded to individual test items. Norizuki et al. (2009; 2010) have attempted to shed light on the quantitative and qualitative characteristics of individual test items. Norizuki et al. (2009) found that one NT was easier than one OT and analyzed a variety of item and text characteristics to identify what might affect easiness or difficulty of OT or NT items. By comparing two pairs of OTs and NTs, Norizuki et al. (2010), on the other hand, found that one NT listening section was significantly easier and all the other sections and tests were roughly at an equal difficulty level, and sought to identify what types of items were deemed easy or difficult. (See Section 4 for detailed discussions based on Norizuki et al. 2010.)

The discussion of easy or hard items should not be complete, however, without the analysis of test takers, particularly in terms of how they interact with those items. Norizuki et al. (2010) exemplified portions of examinees' comments collected through

retrospective verbal reports in conjunction with empirical findings about item and text difficulty as well as item/person fit analyses.

There has been a growing recognition of a vast array of factors that may affect test performance, which are more than pure attributes of the test taker's language ability and item difficulty. Item difficulty can be examined in comparison with a variety of other numerical and categorical item characteristics as well as the content and structure of the test text. Brindley and Slatyer (2002) focused on the effects of task characteristics and task conditions on learners' performance in ESL listening assessments and investigated 1) speech rate, 2) text type, 3) number of hearings, 4) input source and 5) item format. In Freedle and Kostin's (1999) extensive study of TOEFL listening comprehension mini-talk items, three groups of independent variables (including 11 item variables, 31 text variables and 39 text/item overlap variables) were correlated with item difficulty indices as the dependent variable.

Certain test tasks might well be relatively easy or difficult for test takers at given linguistic ability levels, depending on various item and text-related factors involved. Their task performance could also be affected significantly by their non-linguistic attributes, such as prior knowledge in the topic domain and meta-cognitive and cognitive test-taking strategies and tactics. Farr, Pritchard, & Smitten (1990) explored L1 college students' use of test-taking and reading strategies elicited from introspective and retrospective interviews on the process of taking a multiple-choice reading comprehension test. Similarly, in Cohen and Upton's (2007) study, exhaustive lists of reading and test-taking strategies were identified for ESL students' responses to multiple-choice and other selective-response-format TOEFL reading subtest items through the analysis of introspective think-aloud verbal reports.

Cohen (1998, p.256) surmises that a major reason for doing research on test-taking strategies is to determine the demands of the learners in solving test tasks in order to better understand their proficiency levels, the quality of the test items or tasks and the relative benefits of including those items or tasks. Some research findings corroborate Cohen's claim to the extent that lower-proficiency learners may not be able to utilize some top-down processing skills and higher-proficiency learners may be able to manipulate a wider range of cognitive and meta-cognitive tactics more effectively (Goh 2000; 2002) and "the more direct assessment of what the students know and what they can do may be achieved" (Gao and Rogers, 2011, p.100).

A rationale behind our present study is to explore some of the factors discussed above or other attributes that may underlie TOEIC test performance and will lay the groundwork for discussions of how the results of TOEIC can serve as diagnostic feedback to students.

Very few studies have been reported on the diagnostic functions of TOEIC. That fact is largely attributable to a claim by Saito (2004) from the IIBC (the Institute for International Business Communication) office, the administrative body of TOEIC in

Japan, that TOEIC “is not a diagnostic test” (as cited in Chapman 2004, p.11), and hence lacks diagnostic information about “what a specific learner needs to be taught” (Childs 1995, p.74). Admittedly, the role of TOEIC as a diagnostic test is practically restricted in normal test settings; however, the cost-effectiveness of a two-hour long TOEIC as a placement test should be properly substantiated, that is, the “test content can be related to classroom tasks and so [can] inform instructional decisions” (Green & Weir 2004, p.467).

With recent advances in cognitive diagnostic assessment (CDA) models, existing achievement or proficiency tests have increasingly become a viable alternative means of diagnostic assessment due to the paucity of authentic diagnostic measures (Sheehan, 1997; Buck, Tatsuoka, & Freedle, 1997; Buck & Tatsuoka, 1998; Leighton & Gierl, 2007a; Jang, 2008; 2009a; 2009b; Sawaki, Kim, & Gentile, 2009). It is beyond the scope of this paper to delineate the theoretical underpinnings of CDA and the extent to which TOEIC test results can be translated into practical diagnostic profiles. Suffice it to notice, however, that Buck, et al. (1997) illustrate a sample diagnostic score report based on their proposed rule-space methodology (RSM) of the OT reading section, as also acknowledged by Sawaki, et al. (2009), and worthy of pedagogically-oriented discussions are how RSM-based Q-matrix, which consists of the number of cognitive attributes by the number of items and has recently been endorsed by researchers like Jang and Sawaki and her colleagues, can function as “the technical diagnosis of examinees on their mastery of each of these attributes” (Buck & Tatsuoka, 1998, p.127).

2. Purpose of this Study

The purpose of this study is to help bridge the gap between the amount of information that official TOEIC tests can provide and the additional information learners may desire. The use of publicized TOEIC Practice Tests (TOEIC Steering Committee, 2000, 2002, 2008, 2009), as compiled and authorized by IIBC, has enabled us to discuss openly the specific characteristics of individual items and examinees, which will lend themselves to the diagnostic analysis of the test results.

To that end, four research questions will be examined:

- (1) *Can different TOEIC tests function equally well as a reliable and valid measure of English proficiency for university students in Japan?*: It is the most fundamental question counted as a prerequisite for discussing more crucial questions that follow. If two pairs of NTs and OTs are all judged to be reliable and valid measures for the intended purpose, we can reasonably proceed to analyze response patterns of individual test items and examinees.
- (2) *Are there any examinees whose item response patterns do not fit well with the probabilistic pass/fail model of the data?* : Here our aim is to identify if there are any examinees who may need special diagnostic treatment. Examinees with misfitting response patterns have typically failed relatively easy items they were

expected to answer correctly and/or have passed relatively difficult items without really understanding anything.

- (3) *Are there any lexico-syntactic differences between apparently equivalent test forms or item-response formats, which might affect the difficulty levels of the tests?*: It can be assumed, for example, that two NTs are parallel-test forms. Similarly, half of NT and OT items are considered identical in item-response formats, as in Parts 1-2 (or I-II) items, Part 5 (or V) items and Part 7 (or VII) single-passage items. A question is whether listening and reading texts for these equivalent test forms or item-response formats are parallel in lexico-syntactic structures. If there are any structural variations which might be affecting test performance, they can be classified for instructional or diagnostic purposes.
- (4) *Are there any other significant factors that might provide diagnostic clues conducive to students learning to respond to the hardest items more confidently or to them solving the easiest items without unwarranted errors?*: It is our premise that TOEIC, regardless of NT or OT used, is potentially a rich source of diagnostic information, which is intended to be revealed in the quest of Research Questions 2 and 3. Here our question is to explore other important variables which are worthy of future discussions in the context of diagnostic assessment.

3. Methodology

The present study is a composite of several interdependent analyses of two sets of test data. One set of tests, which consists of one OT (OT08) and one NT (denoted as either 08NT08 or NT08), was administered to 136 students from five universities in Japan in the autumn of 2008. Another set, comprising one OT (OT10) and two NTs (denoted as NT10 and either 10NT08 or NT08), was taken by 86 students in the spring of 2010, either in the order of OT10→NT10→10NT08 or that of NT10→OT10→10NT08, to counterbalance the order effect, thus helping to minimize the differential amounts of practice effects. The 08NT08 in the 2008 administration was used as the anchor being the name of 10NT08 (i.e., the only common set of test items used in both test administrations) in the 2010 administration for the sake of equating the two OTs and two NTs on a common latent metric. All the students took part in either 2008 or 2010 administration on a voluntary basis.

The test data were empirically scrutinized using EXCEL and SPSS, as well as WINSTEPS for a series of Rasch-model-based analyses.

To reinforce the analyses of the test data, content-related think-aloud and Tree-Tagger analyses were conducted as follows:

- (1) a think-aloud analysis (PA1) of 65 items from NT10 re-answered and self-analyzed in November 2010 by an examinee who showed 19 misfit responses at the original test administration; and

- (2) a think-aloud protocol analysis (PA2) of 63 items sampled from OT08 and 08NT08 taken by eight examinees in the summer or autumn of 2009;
- (3) a Tree-Tagger analysis of the utterances contained in Part I (or 1) from four TOEIC tests.

Theoretical frameworks and practical applications of think-aloud protocol analysis (PA) have been extensively discussed in the fields of cognitive science (Ericsson & Simon, 1993), L2 research and instruction (Jourdenais, 2001; Bowles, 2010), language assessment (Green, 1998; Cohen, 1998; Hughes, 2003; Leighton & Gierl, 2007b) and language assessment of specific skills, such as listening (Wu, 1998; Goh, 2000; 2002), reading (Farr et al., 1990; Alderson, 2000; Cohen & Upton, 2007; Jang 2008; 2009a; 2009b; Gao and Rogers, 2010), writing (Weigle 2002; Barkaoui, 2011) and vocabulary (Hoshino, 2009; Deschambault; 2011).

Closest to our adopted PA procedure may be the one for what Green (1998) labels as think-aloud, retrospective, mediated verbalizations. PA1 was administered on an ad-hoc basis to an individual student. The reason for this independent analysis was that the student who was enrolled in one of the present authors' class in the 2010 autumn semester turned out to be one of the most conspicuous misfit examinees, showing a large number of model-misfit response patterns for NT10, mostly caused by unexpected correct responses to difficult items when she participated in the 2010 spring test administration. All PA1 items were sampled from NT10, including those with her model-fit and misfit responses, which were of low, medium and high difficulty, selected in a balanced way. Her expected TOEIC score at the time of the original test sitting was probably in the range of mid-300s, and was probably nearly 400 or over at PA2 in view of her TOEIC IP score of mid-400s as recorded a few months later.

PA2 employed 63 items collected from OT08 and NT08, comprising roughly the same number of low, medium and high difficulty items, some of which had been selected from misfit (infit or outfit indices exceeding 2.0) ones. Eight students from three universities volunteered to participate in PA2 in the summer and autumn of 2009. Their proficiency levels were not certain, but were presumed to range from TOEIC scaled scores of 300s to 800s and most of them were probably around 600s or over.

Minimal prompts or probes were used in order to help learners to verbalize their ideas smoothly in a relaxed, humanistic atmosphere. Only explicitly informative comments were transcribed and summarized for seeking plausible explanations for the above-mentioned statistical data. No descriptive or inferential statistics were computed for protocol documents themselves.

Tree-Tagger is a tool for annotating text with part-of-speech and lemma information (Schmid, 1994; Fairon, Paumier, & Watrin, 2005). An online Tree-Tagger program (Goto, 2011) was applied to the analysis of Part I (or 1) total oral utterances, excluding item numbers (*1-10* or *1-20*) and the letters of choices (*A-C*), involved in the

four versions of the TOEIC. The selective application to Part 1 (or I) utterances was primarily due to the fact that salient differences were found there in item difficulty between NT08 and OT08 (Norizuki et al., 2009), and some kind of lexico-syntactic differences were alluded to by some of the examinees, aside from some variations in the clarity of visual stimuli.

4. Results

Research Question 1

Table 2 shows the descriptive statistics for the three versions of TOEIC tests (OT10, NT10, and 10NT08) from the 2010 administrations ($n = 86$) and two (OT08 and 08NT08) from the 2008 administrations ($n = 136$). The 10NT08 and 08NT08 correspond to the same NT, administered to separate groups of examinees at different times. Two Listening items needed to be deleted from the analysis for 10NT08 because all the examinees got these items correct, requiring the same items from 08NT08 to be deleted for anchoring purposes.

Table 2

Descriptive Statistics for Different TOEIC Tests

	OT10		NT10		10NT08		08NT08		OT08	
	LIS	RD	LIS	RD	LIS	RD	LIS	RD	LIS	RD
NI	100	100	100	100	98	100	98	100	100	100
NE	86	86	86	86	86	86	136	136	136	136
RWM	55.00	50.85	56.19	52.65	67.18*	55.67	67.60*	50.77	55.71	51.49
SD	10.50	13.12	12.03	14.13	12.88*	14.22	12.37*	13.23*	12.02	12.63
RMD	-0.33	-0.05	-0.37	-0.15	-1.03	-0.34	-1.03**	-0.34**	-0.31**	-0.38**

Note. NI = Number of items; NE = Number of examinees; RWM = Raw mean score; RMD = Rasch mean difficulties

Mean raw scores fluctuate considerably from test to test. When Rasch item difficulties from the four tests, with 10NT08 serving as an anchor for equating purposes, were compared on the common latent interval scale through the application of the TDAP program (Ohtomo, Nakamura, & Akiyama, 2007), only the NT08 Listening section was found distinctly easier than the other sections when analyzed by one-way ANOVA and post-hoc multiple comparison (Norizuki et al., 2010).

Table 3 shows the KR-20 reliability coefficients for five sets of test data derived from four TOEIC tests, alongside correlation coefficients between them. All the reliability statistics exceed .90, whereas the correlations between the 2010 tests measure around .90. Meanwhile, pairs of Rasch item difficulties from 10NT08 and 08NT08 exhibit a very high correlation of .951.

Table 3

KR-20 for Different TOEIC Tests and Correlations between Them

	OT10	NT10	10NT08	08NT08	OT08	10&08NT
REL	.918	.933	.941	.933	.928	.951
OT10		.896				
NT10			.910			
08NT08					.831	

Note. REL = KR-20 reliability or correlation between 10NT08 & 08NT08

How well each of those tests works as an intended measure is examined in the light of the item-examinee response validity. Table 4 lists the number of items whose mean-square fit (MSF) values lie outside the range 0.75–1.3 as possible sources of data contamination, as suggested by McNamara (1996) and Bond & Fox (2007). According to this rule of thumb, a fairly large number of items would be potentially problematic. Meanwhile, Linacre (2008) argues for a less rigorous benchmark, interpreting a critical MSF value of 2.0 or more as distorting or degrading the measurement system.

On the less rigorous side, two out of all the items analyzed in the present study exceed the critical MSF value of 2.0 in terms of their outfit statistics. Although not shown in Table 4, in a similar way, three examinees from the 2010 data and one from the 2008 counterpart record outfit MSF values of over 2.0. It is likely that misfits were induced by a few unlikely responses by low- or high-performers (outfit) rather than systematic unexpected responses by medium-ability examinees (infit).

To check the effect of idiosyncratic examinee response patterns upon the quality of the tests, misfit examinees were temporarily deleted from the corresponding data sets. With a few more items to be subsequently deleted owing to their ensuing perfect scores, no remaining items were found to be misfitting with regard to Linacre's criterion, as reported in Table 5. Similarly, only one item violated McNamara's rule of thumb for infit MSF values, and significantly fewer dubious items were found for outfit MSF values.

Table 4

Infit (IF) and Outfit (OF) Statistics for Different TOEIC Tests

	OT10 (200 items)		NT10 (200 items)		10NT08 (198 items)		08NT08 (200 items)		OT08 (200 items)	
	IF	OF	IF	OF	IF	OF	IF	OF	IF	OF
>2	0	1	0	1	0	0	0	0	0	0
>1.5	0	2	0	4	0	8	0	1	0	2
1.5–1.3	0	5	1	7	3	7	0	11	0	6
0.75–0.5	0	5	0	6	0	18	0	12	0	3
<0.5	0	1	0	1	0	2	0	2	0	0

Unlike most of the expected responses observed in the protocol session, five hard items were passed again. One of them was the following item (NT10 Part 5 Item difficulty 2.10), followed by the learner's verbal report describing her process of choosing the right answer for the wrong reason.

Ex1. Science Gadgets, Inc., supports every product it sells ----- and will gladly offer a refund for any purchase a customer is dissatisfied with.

(A) conclusively (B) factually *(C) unconditionally (D) steadily
(*correct option*)

"I didn't understand the sentence well. I simply guessed a customer is not happy about something s/he bought. None of the words in the options were very familiar to me, but I thought the best choice would be (C) which might refer to some undesirable conditions."

The examinee failed to understand the meaning of the stem or that of the four choices and therefore attempted to focus on a word with some plausibly negative connotation attached to "dissatisfied" near the end of the sentence. This test-taking strategy accidentally prompted her to choose the correct option (C) because of the prefix "un," often used in words with a negative meaning.

On the whole, different versions of the TOEIC functioned well as reliable and valid measures of the intended construct. The presence of a few misfit items is largely due to unexpected responses by a few non-conforming examinees who may need special diagnostic instructions or other practical advice.

Research Question 3

Tables 6 and 7 display some of the easiest and hardest items in 800-item TOEIC tests. The majority of the easiest Listening items can be said to be from NT Part 1 and OT Part I. NT Part 1 comprises 20 items altogether from NT08 and NT10, including 2 items that were deleted from item analysis due to their "unanalyzably easy" perfect scores; 9 of these items are positioned at the easiest part of the list. OT Part I provides 40 items, of which only 6 are among the easiest Listening items.

Table 8 summarizes the results of parts-of-speech tags assigned to Part I (or 1) utterances in four versions of TOEIC through the online Tree-Tagger program. The proportions of the most frequently used lexico-syntactic categories appear to be strikingly similar across different versions of the test. (For example, either NN or DT ranks first, being the two most frequent tags, while VVG ranks third or fifth.)

However, some minor structural variations might have yielded certain sources of easiness or difficulty liable to affect test performance. The following two items illustrate typical variations.

Table 6

*20 Easiest Listening Items Plus Two Items with Perfect Scores (PS) and *21 Easiest Reading Items Presented in Descending Order*

Listening	Part	I. Diff	Reading	Part	I. Diff
10NT08	1	PS	10NT08	7	-4.73
10NT08	1	PS	OT10	V	-3.54
OT10	I	-4.60	OT08	V	-3.22
OT10	I	-4.60	OT08	V	-3.09
10NT08	1	-4.02	OT08	V	-2.87
10NT08	1	-4.02	NT10	7	-2.80
10NT08	1	-4.02	OT08	V	-2.77
NT10	1	-3.93	OT08	VII	-2.68
NT10	1	-3.60	10NT08	5	-2.66
10NT08	2	-3.60	10NT08	5	-2.66
10NT08	3	-3.60	10NT08	5	-2.51
NT10	1	-3.50	NT10	5	-2.48
10NT08	2	-3.29	10NT08	5	-2.24
OT10	I	-3.16	10NT08	5	-2.24
NT10	1	-2.96	NT10	6	-2.21
NT10	2	-2.96	NT10	6	-2.21
OT08	I	-2.93	OT10	VII	-2.19
10NT08	3	-2.85	OT10	V	-2.08
NT10	2	-2.76	OT10	V	-2.08
OT10	I	-2.72	10NT08	5	-2.02
OT10	II	-2.72	10NT08	5	-2.02
OT08	I	-2.68			

Note. I. Diff = Rasch item difficulty

*Twenty-one hardest Reading items were selected because two items at the bottom of the list had the same difficulty estimate

Table 7

20 Hardest Listening Items and 20 Hardest Reading Items Presented in Ascending Order

Listening	Part	I. Diff	Reading	Part	I. Diff
10NT08	3	1.27	OT10	VI	1.45
OT08	II	1.32	OT10	VII	1.45
NT10	3	1.36	OT08	V	1.46
NT10	4	1.36	10NT08	6	1.49
OT08	III	1.37	10NT08	7	1.49
OT10	III	1.42	10NT08	7	1.49
OT10	IV	1.42	10NT08	7	1.49
OT08	III	1.42	OT10	VII	1.52
OT10	IV	1.49	OT08	VI	1.52
OT08	III	1.51	OT10	V	1.69
NT10	4	1.52	OT08	VII	1.71
OT10	III	1.66	OT08	VII	1.77
OT10	III	1.74	OT10	VII	1.78
OT10	IV	2.04	NT10	7	1.80
OT10	IV	2.04	OT08	V	2.08
OT08	III	2.27	NT10	5	2.10
OT08	III	2.35	10NT08	5	2.15
NT10	3	2.46	10NT08	7	2.26
OT10	III	2.57	OT10	V	2.32
NT10	4	2.62	OT10	VII	2.78

Ex 2. (A) She's cooking a meal.

(B) She's walking around a farm.

(C) She's shopping for food.* (*correct option*)

(D) She's eating a salad.

Ex 3. (A) Few of the tables are occupied.* (*correct option*)

(B) There's water on the floor of the restaurant.

(C) The shape of the tables is square.

(D) The restaurant has gone out of business.

Table 8

Parts-of-speech Tags Assigned to Part I (or 1) Utterances from Different TOEIC Tests

R	OT08	R	OT10	R	NT08	R	NT10
1	NN: 107 tokens (19.3%)	1	DT: 95 tokens (19.8%)	1	DT: 45 tokens (20.3%)	1	DT: 61 tokens (21.5%)
2	DT: 101 (18.2%)	2	NN: 78 (16.3%)	2	NN: 33 (14.9%)	2	NN: 55 (19.4%)
3	VVG: 51 (9.2%)	3	VBZ: 48 (10.0%)	2	VVG: 33 (14.9%)	3	NNS: 32 (11.3%)
4	IN: 47 (8.5%)	4	PP: 44 (9.2%)	4	VBP: 23 (10.4%)	4	IN: 31 (10.9%)
5	NNS: 41 (7.4%)	4	VVG: 44 (9.2%)	5	NNS: 21 (9.5%)	5	VVG: 29 (10.2%)
6	VBP: 35 (6.3%)	6	IN: 37 (7.7%)	5	PP: 21 (9.5%)	6	VBP: 22 (7.7%)
6	VBZ: 35 (6.3%)	6	NNS: 37 (7.7%)	7	VBZ: 17 (7.7%)	7	VBZ: 13 (4.6%)
8	PP: 31 (5.6%)	8	VBP: 21 (4.4%)	8	IN: 12 (5.4%)	8	VVN: 11 (3.9%)
9	JJ: 17 (3.1%)	9	JJ: 14 (2.9%)	9	VVN: 5 (2.3%)	9	PP: 8 (2.8%)
9	VVN: 17 (3.1%)	9	VVN: 14 (2.9%)	12	JJ: 2 (0.9%)	12	JJ: 2 (0.7%)
NR	24		23		15		17
NT	554		479		222		284

Note. DT = Determiner, NN = singular or mass noun; NNS = plural noun; PP = personal pronoun; IN = Preposition or subordinating conjunction; VVG = general verb, gerund or present participle; VBP = “be” verb, non-3rd person singular present; VBZ = “be” verb, 3rd person singular present; VVN = verb, past participle; JJ = adjective, R = rank; NR = number of rank categories; NT = number of tokens

Example 2 from NT08 was one of the two perfect-score items deleted from item analysis at the 2010 administration. In each option, the subject “she” is followed by a verb phrase in the present progressive form. Some of the PA2 participants noticed the conceptual simplicity of such items requiring them only to focus on the verb parts. The co-occurrence of the same personal pronouns and variable present progressive verb phrases occupy over 24% of all the tags recorded in NT08 Part 1. The first five items (50% of Part 1 items) share this feature; another three items (30% of Part 1 items) possess common noun phrases and different verb phrases.

Example 3 from OT08, on the other hand, was the most difficult item among the 60 Part I (or 1) items, with a Rasch item difficulty estimate of 0.75. Its relative difficulty was ascribed by some of the PA2 respondents to nonparallel utterance structures in the four choices; the use of two subjective adjectival words (“few” and “square”); unfamiliarity with the word “occupied” in this context and the idiomatic expression “go out of business”; as well as a rather obscure picture for this item. Below are the summary of some of the extracts. The respondents’ expected TOEIC scaled

scores would be around 600s or 700s.

PA2-A (Female)

“I chose (B). I thought (A) is wrong as there aren’t many people sitting, and (C) is also wrong as the shape of the table is not square. I didn’t understand (D). (*After confirming the answer*) I missed the word “few” so I couldn’t get the right meaning.”

PA2-B (Male)

(*After confirming the answer*) “If I had known the meaning of the word “occupied,” I would have probably been able to choose (A).”

PA-C (Female)

(*After confirming the answer*) “Even though I was able to choose the right answer, I think the question is fairly difficult because the picture is not clear, even to make a simple judgment. The (phrasal) verb in (D) is also difficult.”

The relative difficulty of OT08 as a whole may well be characterized by more frequent use of adjectival expressions (3.1% for OT08; 0.9% for NT08) and fewer parallel present progressive constructions (4 items out of 20), as compared with NT08.

Research Question 4

In our earlier research, text difficulty was examined in terms of the number of words in the text, the number of questions involved and the Guiraud index, showing fairly high correlations with the average difficulties of the passage concerned in NT08 (.813, .768, .726, respectively), but not so remarkably in OT08 (.250, .182, .059, respectively) (Norizuki et al., 2009).

Question difficulty was later addressed, and OT08 Part VII and NT08 Part 7 items were categorized into six question types. One of the major findings was that types of items tapping inferential skills were considerably more difficult than other types of items (Norizuki et al., 2010).

The issue of actual item difficulty or easiness is not always so straightforward. The topic of question types is worthy of further scrutiny, including items from other parts and those in the Listening section.

Example 4 (OT08 Part II Item difficulty 1.32) illustrates a question in a grammatically simple structure followed by cognitively simple but lexically challenging response options. One of the PA2 respondents whose TOEIC scaled score would be in the 800s explained his wrong choice of option (A) below as follows.

Ex. 4: Q: What are your total production figures for this factory?

- (A) We produce pharmaceuticals.
- (B) Over 1000 units a week* (*correct option*)
- (C) In the shipping department

“I chose (A). This question was really difficult. I thought the answer would include a noun phrase demanded by the question starting with an interrogative pronoun ‘what.’ I discarded option (C) immediately as it appeared to be an adverbial phrase. I was not sure about (B), but I didn’t think the answer would be a numerical phrase like (B). I chose (A) which seemed to include a noun phrase.

The above comment implies that the presence of some unfamiliar vocabulary in the aural input, despite the abundance of useful syntactic clues, can critically deter the advanced learner’s comprehension.

Apart from structural variability, text difficulty, question types, and vocabulary familiarity, it is our conviction that the following characteristics are some of the major determinants of item easiness or difficulty, as indicated by past studies (e.g., Freedle and Kostine, 1999; Brindley & Slatyer, 2002; Norizuki, et al., 2009; 2010).

- syntactic complexity
- directness of the correct option in relation to the notion that is described in the text
- frequency of key words used in the wrong options
- concreteness of key concepts & vocabulary
- topic familiarity
- logical or cognitive complexity
- speech rate of aural input

It is beyond the scope of the current paper to discuss them, but some of the issues will be featured in our future research.

5. Discussion

Let us briefly address the four research questions posed earlier and interpret the results of this study.

(1) *Can different TOEIC tests function equally well as a reliable and valid measure of English proficiency for university students in Japan?*

Yes. The tests worked fairly reliably, with all the KR-20 measures exceeding .80, and most centering around .90. Some misfit items with high outfit statistics are attributable to unexpected responses by low-proficiency test-takers. With the temporary deletion of such test-takers, it was found that all 800 items across four tests (except for the two items with all correct responses) were found to be within the generally acceptable level of fit statistics, meaning that all the tests share high degrees of response validity.

(2) *Are there any examinees whose item response patterns do not fit well with the probabilistic pass/fail model of the data?*

Yes. A few examinees were found to be statistically misfitting. One of them was re-examined six months later in the protocol session (PA1) with Listening and Reading items that elicited fit and misfit item responses in the original test administration. Her response pattern became more in tune with the assumed probabilistic model. Such examinees should be scrutinized as far as time allows and proper diagnostic feedback should be offered to them when necessary.

- (3) *Are there any lexico-syntactic differences between apparently equivalent test forms or item-response formats, which might affect the difficulty levels of the tests?*

Possibly yes. While the analysis of Part I (or 1) utterances revealed striking similarities across two OTs and two NTs, there were some exceptional and yet salient differences which might have affected examinees' item or test performance, as corroborated by PA2 respondents. Further exploration of the nature of the differences may lead to the systematic and scientific classification of diagnostic information so as to help students pinpoint their weaknesses. Meanwhile, different versions of the tests may not need to be precisely parallel in every small detail so long as the overall test reliability and validity are maintained, and test-takers should assume certain degrees of variability as a matter of fact.

- (4) *Are there any other significant factors that might provide diagnostic clues conducive to students learning to respond to the hardest items more confidently or to them solving the easiest items without unwarranted errors?*

Yes. There are a number of potential candidates for our future research, as exemplified earlier. The sheer volume of a single TOEIC Practice Test can be cherished as a rich resource for diagnostic feedback that would truly benefit university students in Japan.

6. Conclusions

The present study examined some distinctive TOEIC item-examinee response characteristics in order to illuminate potential diagnostic functions of TOEIC tests. Our findings are still fragmentary and implicational, and thus need to be substantiated and consolidated by studies with more meticulous attention to what can be done with TOEIC to diagnose strengths and weaknesses of individual learners and how.

One important question that remains is whether the results and implications of this study are of any practical relevance to the students taking an official TOEIC or IP (Institutional Program) test.

The answer is still largely conditional, depending on the timing of the use of TOEIC Practice Test (TPTs) and the nature of diagnostic feedback available to individual students in relation to their opportunities to take an official or TOEIC or IP test.

The issue of the timing should be discussed in the context of institutional and instructional demands. If TOEIC is thought to be relatively difficult for targeted students in general, easier resource books may be utilized prior to the frequent use of more authentic TPTs. If the aimed students find overall TOEIC questions fairly easy, mixed selections of difficult test items from old and new versions of TPTs or more difficult resource books may be considered. The nature of feedback may be an issue of extreme difficulty, considering a variety of factors affecting learners' test performance. The complexity of the matter is actually proof that future research is warranted. If the diagnostic functions of TOEIC are performed effectively in a practical educational setting, they would benefit enormously the test-takers of TOEIC in particular and learners of English in general.

Acknowledgments

This study was supported by the TOEIC® Research Fund. We would like to express our appreciation to anonymous reviewers on an earlier version of this paper as well as Professor Chizuko Suzuki at Nagasaki Junshin University, Mr. Mineo Mitsuhashi at the Institute for International Business Communication, and Mr. Satoshi Monzawa, a former student at Shizuoka Sangyo University for their valuable assistance and advice. We are also indebted to 222 university students who volunteered to take multiple versions of TOEIC® tests.

References

- Alderson, J.C. (2000). *Assessing reading*. Cambridge, UK: Cambridge University Press.
- Barkaoui, K. (2010). Explaining ESL essay holistic scores: A multilevel modeling approach. *Language Testing*, 27, 515-535.
- Bond, T.G., & Fox, C.M. (2007). *Applying the Rasch model*. (2nd ed.) Mahwah, NJ: Lawrence Erlbaum Associates.
- Bowles, M. A. (2010). *The think-aloud controversy in second language research*. New York: Routledge.
- Brindley, G., & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing*, 19, 369-394.
- Buck, G., Tatsuoka, K., & Kostin, I. (1997). The subskills of reading: Rule-space analysis of a multiple-choice test of second language reading comprehension. *Language Learning*, 47, 423-466.
- Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining the attributes of a free response listening test. *Language Testing*, 15, 119-157.
- Chapman, M. (2004). Insights in language testing: An interview with Kazuhiko Saito. *JALT Testing & Evaluation SIG Newsletter*, 8(2), 8-11.

- Childs, M. (1995). Good and bad uses of TOEIC by Japanese companies. In J. D. Brown & S. O. Yamashita (Eds.), *Language testing in Japan* (pp. 66–75). Tokyo, Japan: JALT.
- Cohen, A. (1998). *Strategies in learning and using a second language*. London: Longman.
- Cohen, A. D. & Upton, T. A. (2007). “I want to go back to the text”: Response strategies on the reading subtest of the New TOEFL. *Language Testing*, 24, 209–250.
- Deschambault, R. (2011). Thinking-aloud as talking-in-interaction: Reinterpreting how L2 lexical inferencing gets done. *Language Learning*. doi: 10.1111/j.1467-9922.2011.00653.x
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). Cambridge, MA: MIT Press.
- Fairon, C., Paumier, S. & Watrin, P. (2005). Can we parse without tagging? *Proceedings of the Language & Technology Conference: Human Language Technologies* (pp.473–477). Poznan, Poland.
- Farr, R., Pritchard, R., & Smitten, B. (1990). A description of what happens when an examinee takes a multiple-choice reading comprehension test. *Journal of Educational Measurement*, 27, 209–226.
- Freedle, R., & Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? the case for the construct validity of TOEFL's minitalks. *Language Testing*, 16, 2–32.
- Gao, L., & Rogers, W. T. (2011). Use of tree-based regression in the analyses of L2 reading test items. *Language Testing*, 28, 77–104.
- Goh, C. C. M. (2000). A cognitive perspective on language learners' listening comprehension problems. *System*, 28, 55–75.
- Goh, C. C. M. (2002). Exploring listening comprehension tactics and their interaction patterns. *System*, 30, 185–206.
- Goto, K. (2011). TreeTagger online (Version 0.7) [Computer software]. Retrieved April 26, 2011 from <http://web4u.setsunan.ac.jp/Website/TreeOnline.htm>
- Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook*. Cambridge, UK: Cambridge University Press.
- Green, A. B., & Weir, C. J. (2004). Can placement tests inform instructional decisions? *Language Testing*, 4, 467–494.
- Hoshino, Y. (2009). Reactivity of thinking aloud in EFL learners' performance in vocabulary tests. *JLTA Journal*, 12, 16–25.
- Hughes, A. (2003). *Testing for Language Teachers* (2nd ed.). Cambridge, UK: Cambridge University Press.
- Ito, A., Shimatani, H., & Norizuki, K. (2010). *Shinkyuu TOEIC® no yosokuteki datousei no kenshou* [Two versions of the TOEIC®: A Predictive validation study]. *JLTA Journal*, 13, 111–125.
- Ito, A., Shimatani, H., Norizuki, K., & Kinoshita, M. (2009a). *Kotenteki test riron ni yoru shinkyuu TOEIC® no hikaku bunseki* [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). *Shinkyuu TOEIC® no jukensha haichi ni okeru heikoudo no kentou to naiteki datousei kensho* [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.
- Jang, E. E. (2008). A framework for cognitive diagnostic assessment. In C. A. Chapelle, Y.-R. Chung, & J. Xu (Eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 117–131). Ames, IA: Iowa State University.
- Jang, E. E. (2009a). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to LanguEdge assessment. *Language Testing*, 26, 31–73.
- Jang, E. E. (2009b). Demystifying a Q-Matrix for making diagnostic inferences about L2 reading skills. *Language Assessment Quarterly*, 6, 210–238.
- Jourdenais, R. (2001). Cognition, instruction and protocol analysis. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 354–375). Cambridge, UK: Cambridge University Press.
- Leighton, J.P., & Gierl, M.J. (Eds.) (2007a). *Cognitive diagnostic assessment for education: Theory and practices*. Cambridge, UK: Cambridge University Press.
- Leighton, J.P., & Gierl, M.J. (2007b). Verbal reports as data for diagnostic assessment. In J.P. Leighton & M.J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and practices* (pp.146-172). Cambridge, UK: Cambridge University Press.
- Liao, C.-w. (2010). *Field study results for the redesigned TOEIC® Listening and Reading test* (Compendium Study TC-10-03). Princeton, NJ: Educational Testing Service.
- Liao, C.-w., Hatrak, N., & Yu, F. (2010). *Comparison of content, item statistics, and test-taker performance on the redesigned and classic TOEIC® Listening and Reading test* (Compendium Study TC-10-04). Princeton, NJ: Educational Testing Service.
- Linacre, M. (2008). *WINSTEPS Rasch measurement computer program* (Version 3.66.0). Chicago: Winsteps.com; New York: Kaplan Publishing.
- McNamara, T. (1996). *Measuring second language performance*. New York, NY: Addison Wesley Longman.
- 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., & Kinoshita, M. (2009, September). *Sakuranshi to text no bunseki ni yoru shinkyuu TOEIC® no hikaku kenshou* [Analysis of distractor and text difficulty in two versions of the TOEIC® test]. A paper presented at the 13th JLTA Annual Conference, Hokkai Gakuen University.

- Ohtomo, K., Nakamura, Y., & Akiyama, M. (2007). *Test data analysis program: TDAP Ver. 2.02*. Open source program.
- Powers, D. E., Kim, H.-J., & Weng, V. Z. (2008). *The redesigned TOEIC® (Listening and Reading) test: Relations to test-taker perceptions of proficiency in English* (ETS Research Report No. ETS RR-08-56). Princeton, NJ: Educational Testing Service.
- Sawaki, Y., Kim, H.-J., & Gentile, C. (2009). Q-Matrix construction: Defining the link between constructs and test items in large-scale reading and listening comprehension assessments. *Language Assessment Quarterly*, 6, 190–209.
- Schedl, M. (2010). *Background and goals of the TOEIC® Listening and Reading test redesign project* (Compendium Study TC-10-02). Princeton, NJ: Educational Testing Service.
- Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of International Conference on New Methods in Language Processing* (pp.44-49). Manchester, UK.
- Sheehan, K. M. (1997). A tree-based approach to proficiency scaling and diagnostic assessment. *Journal of Educational Measurement*, 34, 333-352.
- Shimatani, H., Norizuki, K., Ito, A., & Kinoshita, M. (2009a). *Rasch model ni yoru Shinkyuu TOEIC(R) no hikaku bunseki* [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, Japan: The Institute for International Business Communication.
- TOEIC® Steering Committee. (2002). *TOEIC® test official preparation book Vol. 2 (Japanese version)*. Tokyo, Japan: The Institute for International Business Communication.
- TOEIC® Steering Committee. (2006, April). The new TOEIC® test. *TOEIC® Newsletter*, 92, 1–3.
- TOEIC® Steering Committee. (2008). *TOEIC® test new official preparation book Vol. 3 (Japanese version)*. Tokyo, Japan: The Institute for International Business Communication.
- TOEIC® Steering Committee. (2009). *TOEIC® test new official preparation book Vol. 4 (Japanese version)*. Tokyo, Japan: The Institute for International Business Communication.
- Weigle, S.C. (2002). *Assessing writing*. Cambridge, UK: Cambridge University Press.
- Wu, Y. (1998). What do tests of listening comprehension test? - A retrospection study of EFL test-takers performing a multiple-choice task. *Language Testing*, 15, 21-44.