

Can TOEIC test be an alternative to the National Center Test?

— Comparison of test difficulties reflected in multiple-choice distributions —

Kei Miyazaki

1. Introduction

In recent years, a great deal of attention has been given to external language examinations which are recognized as objective and consistent measures of foreign language assessment. The considerable influence of these examinations on English as a Foreign Language (EFL) in Japan has been spreading rapidly among private sectors and academic institutions. In fact, the score measured by external commercial exams such as the Test of English for International Communication (TOEIC[®]) has already been used as a means of screening for employment and promotion in many corporations. Accordingly, in an academic situation, the Ministry of Education, Culture, Sports, Science and Technology (MEXT) proposed clear exit criteria based on external exams for average high school learners (MEXT, 2010).¹ Furthermore, MEXT announced that the university entrance exam reforms would include using external language exams such as the TOEIC[®] and the Test of English as a Foreign Language (TOEFL[®]) as substitutes for the National Center Test for University Entrance Exam (Center Test)² (Sankei msn, 2013). With this trend, many high schools have incorporated special training courses for external exams to meet the needs of the students who plan to take the Center Test. The students also mention the significant role the external exams play in gaining entrance to university and for succeeding in their future careers.

While attempts at using the external exams as an alternative to entrance examinations might be beneficial, there is a problem that must be addressed. The problem is whether the results obtained on two different exams are similar to each other for high school learners. In other words, the test used as an alternative to the entrance examination must show the same result as the entrance exam, not only in terms of the score itself, but also in terms of its difficulty and test taker's ability. For example, if the difficulty levels of the Center Test and the TOEIC[®] are different, it is inappropriate to accept both tests as criteria for measuring the determination of university admission. If these two tests are to be used in the same criteria, test difficulty should be carefully examined (Bachman & Palmer, 2010).

Moreover, if there is a difficulty gap between these two tests, the influence of the discriminating power for test takers' abilities should also be confirmed, because the factors affecting the item difficulties, in most cases, arise from the question types (Alderson, 2000). In other words, it should be considered whether each question item in the test follows the theoretical foundation for discriminating between test takers' abilities. For example, in the case of multiple-choice item tests, the function of the incorrect items is to serve as distractors, which should be selected by test takers who did not achieve the learning outcome but ignored by test takers who did. Items that are implausible do not serve as functional distractors and thus should not be used.

Given the widespread use of the external tests as criteria of university entrance admission, the comparison of

two different tests is inevitable. The present study focused on the Center Test and the TOEIC[®] because these two tests are currently taken by a large number of high school learners. The study aims to investigate whether TOEIC test result can be an indicator of the Center Test by comparing their difficulties and then by comparing their multiple-choice distributions. For the researcher to interpret the collected data more easily, the study employed the Rasch measurement model, which produces estimates of each test taker's ability and each item's difficulty simultaneously measured in mathematical units.

2. Theoretical Background

2-1. Center Test and TOEIC[®] as NRT

It is well known that the external exams are administered and scored with high standards of consistency and reliability. Since the external exams are used as a way to rank or compare test takers, they are categorized as norm-referenced test (NRT) in the field of language testing (Brown, 2010). Unlike criterion-reference test (CRT) which is designed to give test takers feedback on specific learning objectives through diagnostic and term exams, scores for NRT are interpreted in relation to percentile rank. Each learner's performance is compared to other test takers. The score is usually reported back to the test takers in the form of a numerical score or a percentile rank.

The Center Test and TOEIC test are examples of NRT. The Center Test is a high stakes test which significantly influences decisions on university admission and the scores of the test are often the sole criteria for the determination of admission to public universities or even some private universities in Japan. This test is the only nationwide national standardized exam in Japan (Guest, 2008).

TOEIC test measures English language proficiency in the international work environment, unlike CRT that is used in academic settings such as university and high school. The number of correct answers on the listening and reading sections is represented as a standard score with values ranging from 10 to 990 (ETS, 2013).³

While the Center Test has long been used as a university entrance criterion, MEXT encourages high school learners to take external exams such as the TOEFL[®] and the TOEIC[®] as substitutes for the Center Test. MEXT is exploring the possibilities for an idea that high school learners who gain a high score on TOEFL (iBT) or TOEIC test, are considered able to gain a high score of the Center Test (Sankei msn, 2013). In this respect, there are some research findings on the validity and reliability issues of TOEFL[®] and the Center Test (Saida, Kobayashi, & Noguchi, 2009 ; Kato, 2011), which suggest little relevancy between these two tests. However, very little research on the relevancy of the Center Test and TOEIC test has been conducted to date, though a large number of universities in Japan have adopted these two tests as admission criteria. In fact, the impact of TOEIC test in Japan is more significant than the one of TOEFL (iBT) because of the variability of the population. ETS (2013) provides demographic information on the population of TOEIC test takers. According to the report, 463 university level institutions (including junior college and technical college) are currently using TOEIC scores as part of the tools for their admission decisions (ETS report, 2013). It is thus worth while comparing the Center Test and TOEIC test.

Both the tests have been improved for the past 10 years. According to Guest (2008), the 2006 version of the Center Test was superior to the earlier versions, in that it tested a variety of cognitive abilities and language skills through a wide range of topics and tasks. The 2009 version was more difficult than the earlier versions (Underwood, 2010).

TOEIC test was also revised considerably in 2006. The test was redesigned in order to better align with the current communicative situation and workplace language proficiency needs. The test adopted many integrative communicative questions such as presentation summaries and double-passage questions (Schedl, 2010).

The use of improved Center Test and TOEIC test as NRT seems to offer a reliable measure of language proficiency. As a result, MEXT might be over reliant on TOEIC test as a reasonable reliable measure for university admission. However, it has not been announced yet that test difficulty of TOEIC test used as an indicator of the ability in the Center Test was taken into consideration.

2-2. Test Difficulty Research

Factors affecting language test difficulty have been studied by Alderson (2000), who states that the factors affecting the difficulty of test items are the language of the question stems and the test types. Of them, the influences of distractors on test items are pointed out in terms of test difficulty in multiple-choice questions (Coombe et al., 2012). The most used methods in investigating the difficulty of a test item have been item analyses, which generally consist of item difficulty and item discrimination. These analyses provide information on how efficiently the items differentiate between test takers' abilities (Brown, 2010). However, these sorts of analyses do not reveal much about the appropriateness of item distractors. The effectiveness of the multiple-choice test in both the Center Test and TOEIC test is largely dependent on its distractors. If some distractors are implausible, the question items are, in effect, not meaningful questions, which consequently do not differentiate between test takers' abilities. It therefore makes sense to confirm how each distractor works, and it helps test developers produce better multiple-choice items. Furthermore, analyzing choice distributions gives researchers some pieces of information about the test appropriateness for the test takers. The present study will employ the Rasch measurement model as an analytic device. This model allows researchers to examine the difficulty of specific test question items in relation to each test taker's ability. It also makes it possible to compare the ability of test takers that chose different distractors on the multiple-choice question items.

2-3. Advantage of Rasch Analysis

The present study utilized the Rasch measurement model which provides a graphical representation of test takers' abilities and item difficulties on the same scale (Green, 2013). This statistical device was chosen because the Rasch model is the one which produces estimates of each test taker's ability and each item's difficulty measured in mathematical units called logits, unlike the statistics for the Classical Test Theory which summarizes the data in sample dependency. The Rasch measurement device demonstrates probability as an interaction between the test takers and the test items, allowing the researcher to confirm the relative difficulty of specific test items on a person item map. It also makes it possible to compare the average ability of test takers who chose different distractors on a specific item. The results from the data analyses using statistical device for the Classical Test Theory and Rasch model may be the same (Molloy, 2009), but Rasch measurement model allows the researcher to interpret the data with relative ease. Specifically, in this study, running a distractor analysis in Rasch measurement is quite easy to conduct and provides information not only about how the individual options are working but also enables researchers to identify which kind of test taker is endorsing which option.

3. Purposes of the Study

The present study was conducted to examine whether the TOEIC test is appropriate when it is used for an alternative to the Center Test. This comparison study between these two different tests was examined by answering the following research questions :

RQ (a) : Is the difficulty of the Center Test and TOEIC test appropriate for third-year high school learners?

RQ (b) : To what extent are the choice distributions of multiple-choice test items on the Center Test and TOEIC test valid?

4. Methods

4-1. Participants and instruments

A total of 77 third-year students of a Japanese high school participated in this study. Their level of English proficiency was assumed to be intermediate. After eliminating the data for two students who did not complete all the questions, the data for 75 participants were analyzed. This experiment was performed in December 2013, when all the participants had reached in peak form for the preparation of entrance examinations.

The tests were created using items from the Center Test held in 2012 (Kyougakusha, 2013), and the TOEIC practice test textbook (TOEIC® Steering Committee, 2011). Due to a time constraint, some of the Center Test question items which are recorded as intermediate levels (40 – 70 % average of correct answers in the past record) were selected. The original listening section of the Center Test was edited to 9 items, and the reading section of the Center Test was edited to 16 items. In total, 25 items were used from the Center Test.⁴ As for TOEIC test, which originally consisted of four listening sections and four reading sections, the intermediate level question items were selected (stated as “intermediate level” in the practice book). The edited listening test selected 2 items from picture test, 2 from question-response, 3 from short conversation, and 3 from short talk. The edited reading test adopted 2 items from incomplete sentences, 3 from text completion, 5 from single passage, and 5 from double passage. In total, 25 items were used from TOEIC test. A full score of each test was 25.

Then two sets of tests, the edited Center Test and the edited TOEIC test, were administered separately, both beginning with the listening section. The Center Test was 25 minutes long, and TOEIC test was 20 minutes long. The participants were not informed of the purpose of the study in order to avoid any influence on their test strategy use.

4-2. Analyses

The test data were empirically scrutinized using MINISTEP 3.81 (Linacre, 2014) for a series of Rasch model measurements.⁵ The following pieces of information were analyzed in this study : (1) the Rasch person reliability index and separation index in order to confirm that analysis was fairly separating test takers into different levels of ability, and the Rasch item reliability index and separation index in order to confirm that the items were likely to perform in the same way on other test populations⁶ ; (2) Person-item maps providing a visual representation of person and item locations on the construct. The locations of test takers’ abilities and question items’ difficulties show the degree of difficulties of two different tests. These two analyses were conducted for Research Question (a). (3) Distractor analyses to obtain information not only about how the individual choice options are working but also identify which ability of test taker is choosing which option. This analysis based on infit and outfit values enables researchers to distinguish functional distractors and mal-functional distractors. This analysis was performed for Research Question (b).

5. Results

Table 1 shows the descriptive statistics of the Center Test and TOEIC test. The correlation coefficient (Pearson product-moment) between the two tests is relatively high ($r = .73$). However, only the correlation cannot indicate that these two tests are similar (Green, 2013). To obtain the information about test takers’ abilities and test item difficulties (RQ (a)), a Rasch measurement was utilized.

Table 1. Descriptive Statistics for Center Test and TOEIC test

	Mean	S.D.	Skew	Kurt	Min	Max	Total
Center	14.30	4.02	-0.08	-0.87	6	22	75
TOEIC	11.82	3.59	0.69	-0.16	6	21	75

Note : Skew = Skewness ; Kurt = Kurtosis ; Min = Minimum ; Max = Maximum (full score = 25)

Table 2. Rasch indices of person-item separation and reliability for Center Test and TOEIC test

	person separation	person reliability	item separation	item reliability
Center Test	1.40	.76	2.97	.90
TOEIC test	1.10	.55	2.10	.82

First, Rasch person and item fit analyses for each test were conducted to determine whether all test takers' answers and all items fit the statistical model's expectations (Table 2). The Rasch person reliability index for the Center Test was estimated at .76, with a Rasch person separation index of 1.40, indicating that the test takers were fairly separated by the items. Rasch item reliability for the Center Test was .90, demonstrating a strong consistency in the items. The item separation index was 2.97, indicating that the examination contains some difficult and easy items.

On the other hand, the Rasch person reliability index for TOEIC test was estimated at .55, with a Rasch person separation index of 1.10, indicating that the test takers were not distinctly separated by the items. Rasch item reliability for TOEIC test was .82, with an item separation index of 2.10. Although both the item separation indices were within positive values of .75 (Linacre, 2014), the item separation of TOEIC test was smaller than that of the Center Test.

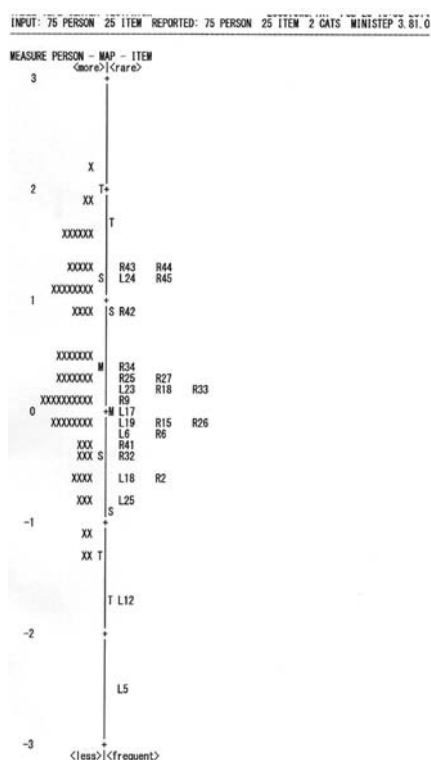
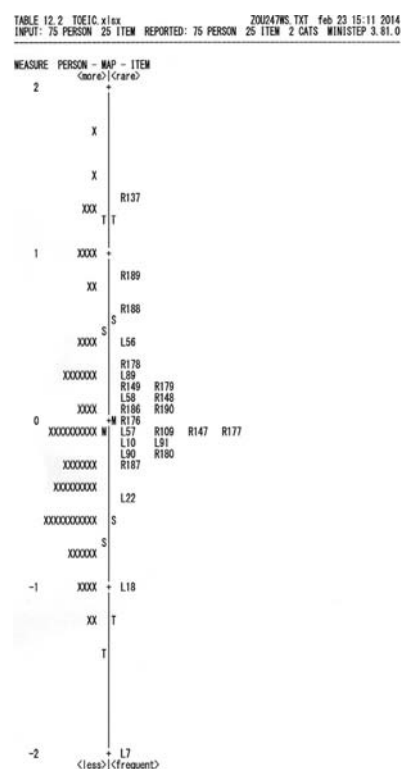
**Figure 1. Person-Item map for Center Test****Figure 2. Person-Item map for TOEIC Test**

Figure 1 and 2 show the item maps for the Center Test and TOEIC test respectively. The left side of the map represents test takers' abilities, and item difficulties are placed on the right side. Each X represents one test taker. The more able test takers are placed towards the top and the less able towards the bottom. The more difficult items appear towards the top and the easier items towards the bottom. The M marker indicates the mean difficulty. The letter S and T on either side of the dividing line refer to the standard deviation ($S = \pm 1$ S.D. and $T = \pm 2$ S.D.).

In Figure 1, the mean of the test takers' abilities is above the mean difficulty, demonstrating that the item questions of the Center Test were not so difficult and reasonably appropriate for the test takers. On the other hand, Figure 2 shows that the mean difficulty of the items is above the average of test takers' abilities. Moreover, about half of the items (items cluster between R137 and R190) have only 26 persons (about one-third of the test takers) in the ability range. This means that the item questions of TOEIC test were more difficult for the

Table 3. Item statistics for Center Test

TABLE CENTER TEST INPUT : 75 PERSON 25 ITEM 2 CATS MINISTEP 3.81.0					
ENTRY NUMBER	MEASURE	MODEL S.E.	INFIT MNSQ	OUTFIT MNSQ	PT- MEASURE CORR.
R43	1.31	.27	1.00	.99	.35
R44	1.31	.27	.88	.81	.49
R45	1.24	.26	.97	.88	.40
L24	1.17	.26	1.07	1.06	.27
R42	.91	.25	1.06	1.11	.28
R25	.41	.25	1.14	1.18	.19
R34	.35	.25	.92	.90	.46
R27	.35	.25	1.09	1.11	.25
R33	.23	.25	.92	.88	.46
L23	.16	.25	1.04	1.03	.32
R18	.16	.25	1.03	1.04	.32
R9	.10	.25	.92	.93	.45
L17	.04	.25	1.12	1.14	.22
R15	-.09	.25	1.03	1.05	.32
L19	-.15	.25	1.08	1.16	.25
R26	-.15	.25	1.01	.96	.35
L6	-.21	.26	1.13	1.15	.20
R6	-.21	.26	1.02	1.00	.33
R41	-.35	.26	.88	.88	.51
R32	-.41	.26	.78	.71	.60
L18	-.55	.27	.98	.87	.39
R2	-.55	.27	.99	.91	.36
L25	-.85	.28	.91	1.04	.38
L12	-1.71	.35	1.01	1.02	.23
L5	-2.53	.47	.99	1.18	.15

Table 4. Item statistics for TOEIC test

TABLE TOEIC TEST INPUT : 75 PERSON 25 ITEM 2 CATS MINISTEP 3.81.0					
ENTRY NUMBER	MEASURE	MODEL S.E.	INFIT MNSQ	OUTFIT MNSQ	PT- MEASURE CORR.
R137	1.32	.29	.77	.72	.62
R189	.86	.27	.97	.97	.34
R188	.65	.26	1.07	1.03	.22
L56	.46	.25	1.19	1.24	.07
R178	.33	.25	.98	.97	.34
L89	.27	.25	1.02	1.02	.28
R149	.21	.25	1.06	1.07	.22
R179	.21	.25	1.04	1.04	.26
L58	.15	.24	.93	.92	.41
R148	.15	.24	1.01	1.01	.29
R190	.09	.24	.92	.89	.43
R186	.03	.24	1.10	1.09	.17
R176	-.03	.24	1.00	.98	.31
R147	-.08	.24	.99	.99	.31
R109	-.08	.24	1.04	1.04	.24
L57	-.08	.24	1.16	1.21	.10
R177	-.08	.24	.98	.99	.33
L10	-.14	.24	.93	.92	.40
L91	-.14	.24	.97	.93	.36
L90	-.20	.24	.97	.99	.34
R180	-.20	.24	.94	.93	.39
R187	-.26	.24	1.08	1.08	.18
L22	-.44	.24	1.08	1.10	.16
L18	-1.00	.26	.98	.97	.28
L7	-2.00	.33	.94	.82	.30

test takers than the Center Test. Additionally, TOEIC test does not differentiate between test takers' abilities, as it shows the low person separation index ($r = .55$).

Next, to examine the extent to which the choice distributions of multiple-choice test items on Center Test and TOEIC test are valid (RQ (b)), the distractor analyses were performed. These analyses were carried out by identifying misfit data in the item infit and outfit mean-square (MNSQ) (Green, 2013). According to Green (2013), the items that misfit the model may have problematic distractors. Misfit is defined as less than .80 MNSQ, or greater than 1.20 MNSQ⁷, based on the recommended criteria of Bond & Fox (2007). In other words, the misfit items out of the infit and outfit value range suggest that the performance of these items is unpredictable, and test takers were answering these items incorrectly when they were not expected to and vice versa. Table 3 and Table 4 show the infit MNSQ values and outfit MNSQ values of all items in both the Center Test and TOEIC test respectively.

In Center Test, all of the items have infit MNSQ values between .80 and 1.20, with the exception of item R32 which has an infit MNSQ value of .78 and outfit of .71. This item also has a relatively high PT-Measure Corr,⁸ suggesting that this is performing in a way the model did not expect. In other words, this question item is considered to be less productive for measurement, not discriminating between test takers' abilities.

Compared with the Center Test which has only one misfit item, TOEIC test has several misfit items in outfit MNSQ (highlighted in item R137, L56 and L57). As with the misfit item in the Center Test, these three items also have figures below or above the infit and outfit value ranges. This demonstrates that TOEIC test includes more inappropriate distractors in the items for third-year high school learners than the Center Test.

Table 5 shows three distractor frequencies that were selected from the Center Test and TOEIC test in order to compare functional distractors with nonfunctional distractors.

Table 5. Distractor frequencies of item R27 in Center Test, L56 in TOEIC and L57 in TOEIC

ENTRY NUMBER	DATA CODE	SCORE VALUE	DATA COUNT %	AVERAGE ABILITY
R27 (Center)	c	0	11	-.19
	a	0	16	.13
	b	0	22	.23
	d	1	51	.59
L56 (TOEIC)	c	0	8	-.26
	b	0	22	.05
	a	0	36	.08
	d	1	34	.04
L57 (TOEIC)	b	0	11	-.11
	d	0	14	-.20
	a	0	26	.11
	c	1	49	.02

DATA CODE represents the multiple-choice options; in SCORE VALUE, 0 represents an incorrect answer and 1 represents a correct answer. DATA COUNT shows what percentage of the test takers endorsed each of the choice options. AVERAGE ABILITY tells us the average ability of those test takers who endorsed the choice options.

As an example of a good distractors, item R27 in the Center Test (see Appendix A) was examined. This item

statistically appears to be working satisfactorily. In general, the choice distribution of the four-option multiple-choice item is well-balanced with 50% correct answers and 25%, 15%, and 10% incorrect answers respectively (Coombe, et al, 2012). On the contrary, in item L56 in TOEIC test (see Appendix B), a large number of test takers selected the wrong choice (a) rather than the correct choice (d). Moreover, as can be seen in the section of AVERAGE ABILITY, the abilities of test takers who chose distractor (a) and (b) were higher than those who chose the correct answer (d). Item 57 also includes mal-functional distractors which attract the test takers with higher ability while those with lower ability choose the correct answer (c). These examples suggest that the answer was not clear and the item included some malfunctioning distractors.

6. Discussion

This study examined two research questions: (a) the difficulty of the Center Test and TOEIC test for third-year high school learners, and (b) the extent to which the choice distributions of multiple-choice test items on the Center Test and TOEIC test are valid. The result for RQ (a) shows that the Center Test is reasonably appropriate for differentiating between test takers' abilities, but TOEIC test is more difficult for third-year high school learners than the Center Test, and it does not create distinct ability groups. This result implies that the content of TOEIC test is beyond areas that high school learners have studied at school. Since TOEIC test focuses on English used in industry and commerce (Chapman, 2006), the content is unfamiliar to high school learners. In other words, TOEIC test, which does not reflect the text diversity of real high school life, should be categorized as a different type of NRT from the Center Test. Therefore, it can be said that the contents of the passages in TOEIC test do not necessarily match those in the Center Test.

The second finding was that while most of the items in the Center Test had appropriate choice distributions in the multiple-choice options, TOEIC test had several nonfunctional distractors. This result does not mean that TOEIC test in itself is an impractical test, but means that TOEIC test is inappropriate for use with high school learners as a criterion for university entrance admissions. This is because TOEIC test has difficulty in successfully differentiating between the abilities of high school test takers. For example, item R137 of TOEIC test used in the study is an example of malfunctioning distractors. (Figure 3).

137. According to the proposal, a large block of rooms in the east wing of the new building will be ----- for storage. (a) designated (b) detained (c) reciprocated (d) signified
--

Figure 3. Question Item R137 in TOEIC test

In this question, most of the test takers chose incorrect answers, whereas only a few test takers with higher ability chose the correct answer. This type of question items is unproductive for construction of measurement as it is too predictable for having most test takers answer incorrectly. This lack of discriminating power comes from the fact that most of the test takers are confused with the unfamiliar words in the multiple-choice options. These words do not appear in the textbooks stipulated by the Course of Study.⁹

Another example is in reference to communicative situation. Figure 4 (also see Appendix B) shows examples of listening question with inappropriate choice distributions. The situation of the short conversation for these question items was in a company office where two people were talking about the deadline of a financial report in the company.

The high school learners who had not experienced this situation could not have guessed the social background. The distractors also included the names of official documents in the company which the high school learners had

56. What is the man preparing?
 (a) A sales plan (b) A research proposal
 (c) A client invoice (d) A financial summary
57. What does the woman say she will do?
 (a) Confirm an appointment (b) Prepare a presentation
 (c) Deliver a document (d) Speak with the president

Figure 4. Question Item L56 & L57 in TOEIC test

never heard of. This type of question items requires the test takers to possess sociolinguistic competence¹⁰ (Bachman, 1990) that is beyond their linguistic levels. Social situations which are quite different from high school life must have had great influence on the choice distributions in the multiple-choice constructions. In the case above, the test takers who did not realize the situation might have misunderstood what they had heard. Therefore, it is possible that low-level test takers are simply guessing at correct answers to items containing unknown social situations.

7. Conclusion

The present study has found that the prediction of the Center Test score using TOEIC score alone is less likely to be sufficient. It is true that TOEIC test can be valid for specific types of test takers with specific ranges of ability, such as workers who aim at promotion in their companies and university students who attempt to obtain a job requiring high English proficiency, but TOEIC test does not offer criteria that are directly relevant to the needs of university entrance admissions. It is, therefore, premature to be excessively reliant on TOEIC test as an alternative to entrance exams.

Despite the fact that MEXT encourages high school learners to take external exams such as TOEIC[®] and TOEFL[®] as alternatives to the Center Test, test difficulties and multiple-choice distributions must be carefully confirmed prior to implementing these new criteria. If MEXT promotes the use of TOEIC[®] and if universities require high school learners to gain a certain competence measured in TOEIC[®], the test specifications¹¹ of the Center Test must be expected to make modifications to adjust the differences from TOEIC test, such as vocabularies, contents of the passages, test formats, and even test purposes. Otherwise, TOEIC[®] should be used as another type of university entrance criterion such as the admission office entrance standard or the special framework of admission standard. In any case, what is demanded of examiners and administrators is to consider how TOEIC[®] should be properly used. As Bachman and Palmer (2010) states, examiners and administrators cannot simply give a test and use it to make inferences and decisions without proving adequate justification for the test. When they use a test score, they should make implicit connections among performance on the test, the ability the test takers have and the educational objectives the test has.

Finally, the limitations of this study should also be addressed. Since this study only analyzed a small number of items from the Center Test and TOEIC test, more items must be used if a clear conclusion is to be reached. Furthermore, other external examinations such as TOEFL[®] and STEP should be examined so that both test takers and examiners of NRT can properly interpret the test scores for university entrance examinations. Nonetheless, the information of the present study will provide a helpful suggestion of the necessity of considering the validation of external language exams when the abilities of high school learners are examined for university entrance admissions.

Notes :

- 1) MEXT urged teachers to examine student's English abilities by utilizing external language exams and expand the

utilization of such exams which measure all four skills for university entrance. These include the Eiken, TOEFL and TOEIC.

- 2) MEXT is exploring the possibilities for an idea that high school students who score a 71 or higher on TOEFL (iBT), or a 780 or higher on TOEIC test, are considered able to gain a full score of the Center Test.
- 3) TOEIC test consists of two separately scored subtests, namely “listening comprehension” and “reading comprehension”. The score report provides listening (5 – 495), reading (5 – 495), and total scaled scores, which is derived by adding the listening and reading scaled scores together (10 – 990).
- 4) Reading section of the Center Test includes questions on pronunciation, words & phrases, and matching, in addition to reading comprehension questions.
- 5) Rasch measurement has a scale in which equal distances anywhere on that scale have equal value. This is based on probability theory — the chances of a person answering an item correctly is a function of his/her ability and the item's difficulty.
- 6) Person reliability or separation index, and item reliability or separation index, provide estimates of how much confidence we can have in the test takers' results and in the performance of the items.
- 7) Acceptable values of infit/outfit mean-square are set up as 0.75 – 1.3 by McNamara (1996) and 0.5 – 1.5 by Linacre (2008), but the value range depends on population size. This study's population is small enough to adopt the value range 0.8 – 1.2 suggested by Bond & Fox (2007).
- 8) This statistic provides information about how strongly the responses to the items align with the ability of the persons ; that is, it represents the relationship between the responses to the items and the ability of the test takers. Low correlation indicates that an item is problematic. Too high correlation indicates that the item does not discriminate between test takers' abilities.
- 9) The Course of Study (2009) requires 3,200 words for high school learners including the words that they are supposed to learn at junior high school level. This number of words has increased from 2,200 in the 1998 version of the Course of Study.
- 10) Bachman's (1990) phrase of sociolinguistic competence “the ability to interpret cultural references and figures of speech” highlights the fact that, in order to understand language, it is necessary to appreciate the extended meanings by a specific situations, particular events, places, institutions, or people.
- 11) Test specification is defined as a set of test guidelines for deciding what a test is constructed to measure and what language skills and language content will be measured.

References

- Alderson, C. J. (2000). *Assessing reading*. Cambridge: Cambridge University Press.
- Bachman, L. F. (1990). *Fundamental consideration in language testing*. Oxford: Oxford University Press.
- Bachman, L. F. & Palmer, A. S. (2010). *Language assessment in practice: developing language assessment and justifying their use the real world*. Oxford: Oxford University Press.
- Bond, T. & Fox, C. (2007). *Applying the Rasch model: fundamental measurement in the human sciences*. (2nd ed.). Mahwah, NJ: LEA.
- Brown, D. (2010). *Language assessment: Principles and classroom practices*. (2nd ed.). New York : Pearson Education.
- Chapman, M. (2006). An over-reliance on discrete item testing in the Japanese business context. Paper Presented at the International Conference on English Instruction and Assessment, April 22-23, 2006, National Chung Cheng University, Taiwan.
- Coombe, C., Davidson, P., O'Sullivan, B., & Stoyanoff, S. (2012). *The Cambridge Guide to Second Language Assessment*. Cambridge: Cambridge University Press.
- ETS. (2013). *TOEIC® technical manual*. Princeton, NJ: Educational Testing Service.
- ETS report. (2013). *TOEIC report on test takers worldwide – 2013*. Educational Testing Service.
- Green, A. (2014). *Exploring language assessment*. Abingdon, Oxon: Routledge.
- Green, R. (2013). *Statistical analyses for language testers*. UK: Palgrave MacMillan.
- Guest, M. (2008). A comparative analysis of the Japanese university entrance Senta Shiken based on a 25-year gap.

- Japan Association for Language Teaching Journal, 30, 85-104.
- Kato, C. (2011). A comparative study on TOEFL-based college English curricula in two universities: analysis with the scores of Center Test, TOEFL, and "RLG Test". *Journal of the Liberal and General Education Society of Japan*, 33 (1), 104-113.
- Kyougakusha (2013). *National Center Test for University Admissions, (Japanese version)*. Tokyo: Kyougakusha.
- Linacre, J. M. (2014). Ministep® (Version 3.81.0) [Computer Software]. Beaverton, Oregon: Winsteps.com. Retrieved from <http://www.winsteps.com/>
- MEXT. (2010). *English Education Reform Plan corresponding to Globalization*. Retrieved from <http://www.mext.go.jp/english/topics/1343591.htm>
- MEXT. (2011). *The course of study for senior high schools guidelines explanation : English translation*. Retrieved from http://www.mext.go.jp/a_142menu/shotou/new-cs/youryou/eiyaku/_icsFiles/afieldfile/2011/04/11/1298353_9.pdf
- Molloy, J. L. (2009). Measuring the English proficiency of students during a course. *Journal of Business Administration*, 74, 161-171.
- Saida, C., Kobayashi, K., & Noguchi, H. (2009). Using proficiency tests for curricular innovations in university English education. *The Japanese Journal for Research on Testing*, 5 (1), 95-105.
- Sankei News msn (2013). *Gaining TOEFL iBT score of 71 or TOEIC score of 780 is equivalent to gaining a full score of National Center Test*, 2013. Retrieved from <http://sankei.jp.msn.com/life/news/131231/edc13123108530000-n1.htm>
- 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.
- TOEIC® Steering Committee. (2011). *TOEIC test new official preparation book, vol. 4 (Japanese version)*. Tokyo: The Institute for International Business Communication.
- Underwood, P. R. (2010). A comparative analysis of MEXT English reading textbooks and Japan's National Center Test. *RELC Journal*, 41 (2), 165-182.

(Appendix A)

次の問いの英文を読み、下線部の語句の意味を文章から推測し、(27)に入れるのに最も適切なものを、(a) ~ (d)のうちから1つ選べ。

Mr. Matsumoto is an English teacher who believes English classes should start with a joke. He always tries to create funny jokes. Some of his students have complained about his jokes being a waste of time. His colleagues also have advised him not to spend so much time writing jokes. However, Mr. Matsumoto is such an obstinate person that he will not listen to them and continue to spend a lot of time making up jokes for his classes.

In this situation, an obstinate person means a person who is (27).

- (a) flexible about ideas (b) generous to his students
(c) unable to reject criticism (d) unwilling to change his mind

(Appendix B)

Question 56 through 58 refer to the following conversation.

- (Man) Ms. Smith, have we received the sales division's expense report? I'm working on a summary of our second quarter financial results, and that's the only things I'm missing.
- (Woman) Oh, Masako just called to say she's finishing up the expense report now. She said she'd sent it to me within the hour. I'll bring it over to you as soon as it arrives.
- (Man) Great! That'll leave me enough time to finish the results this afternoon. I have a meeting with the company president tomorrow, and I was worried I wouldn't have all the information in time.

56. What is the man preparing?

- (a) A sales plan (b) A research proposal
(c) A client invoice (d) A financial summary

57. What does the woman say she will do?

- (a) Confirm an appointment (b) Prepare a presentation
(c) Deliver a document (d) Speak with the president

58. What does the man plan to do tomorrow?

- (a) Attend a meeting (b) Telephone a colleague
(c) Send a package (d) Request some information

Abstract

The present study examines whether TOEIC scores can be an indicator of the results of the National Center Test for university admission for third-year high school learners. The focus was to compare the difficulties and multiple-choice distributions of these two types of norm-referenced tests. The data of 75 high school learners who took both of the tests were analyzed using the Rasch measurement model. The result indicated that TOEIC test was more difficult for the high school learners than the Center Test. Additionally, there were several misfit items in Rasch analysis for TOEIC test, demonstrating that the performance of TOEIC question items is unpredictable for differentiating between high school learners' abilities. The question of how external standardized tests are to be used for university entrance evaluation still remains a matter of guesswork.