Modern English Education, Vol. 11, No. 2, Summer 2010

An Application of the Rasch Model to Assessing an English Achievement Test in a High School Setting*

Jaewoo Shim, Keiko Samimy
Chonbuk National University, The Ohio State University

Shim, Jaewoo & Samimy, Keiko. (2010). An application of the Rasch model to assessing an English achievement test in a high school setting. *Modern Language Education*, *11*(2), 278-294.

This study investigated the acceptability of reading test items developed by non-native Korean teachers of English in a Korean high school by the application of the Rasch model analysis. Data on 17 multiple-choice items of reading achievement for 256 high school students at a high school were collected. The data were submitted to Winstep, a Rasch analysis software, to study general model fit, item difficulty measures, person ability measures, fit indices, reliability, and separation indices. The results of the first Rasch analysis of 17 items indicated 5 items did not fit the Rasch model. The second analysis of the 12 items, after the deletion of 5 misfit items, revealed that the set of test items did not match individual persons′ ability or were poorly targeted with high-and low- ability persons. The content analysis of those 5 misfit items confirmed that the teachers had made serious mistakes in construction. Differential item functioning showed that only 1 item of the 12 test items was biased against gender. The results of the study suggested that Korean teachers of English need more thorough preparation and a higher sense of responsibility for constructing items and assigning grades to their students.

**[reading test items/classroom assessment/Rasch model analysis/
읽기평가/교실수업평가/래쉬분석]**

I. INTRODUCTION

Testing of English reading achievement at a classroom level, whether it be a midterm or final has a significant influence on students since whatever grades they get will be calculated into their school records. Accordingly, summative English scores given to students by their Korean teachers of English as well as those given by other academic

* This paper was supported by research funds of Chonbuk National University in 2009.

content teachers become a factor that helps college administrations decide whether or not accept or reject their college applicants. Thus, achievement tests, from the point of view of students, can be as important as any other high-stakes exams. However, despite the important status of classroom-based achievement tests, few research studies have investigated what types of English achievement tests are constructed by inservice English teachers in Korea. As such, this study was guided by three research questions:

1. How do the teacher-adopted reading test items fit the probabilistic Rasch model?
2. Do the test items present any gender bias?
3. What kinds of problems do the misfit test items indicate?

II. LITERATURE REVIEW

1. Multiple-Choice Tests

Language testing provides various information about what students already know, what they need to learn, what they have failed to learn, who has passed the standards set for the class and who has not, and so on (Brown, 2004). These roles of testing for formative and summative purposes make testing an integrative part of instruction (Brown & Hill, 2007; Merrylees & McDowell, 2007; Moore & Morton, 2007). In particular, to connect classroom learning to testing and assessment of reading skills, Hughes (2003) proposed that reading tests include items without providing time to read the full content of a passage as well as ones that test for such careful reading operations as interpreting topic sentences, distinguishing general statements from examples, inferring the meaning of an unknown word from the context, and so forth. However, according to Grabe (2000), relating reading instruction to testing and assessment is not that straightforward. Grabe (2000) indicated there exist a number of dilemmas associated with reading tests and suggested that test makers ask themselves the following questions in order to see classroom testing as part of instruction:

> 1) Are there any developmental stages of reading? if so, how does assessment change for beginning readers versus intermediate and advanced readers? 2) Are there good reasons to stay within typical bounds of current reading assessment item types? 3) Can we assess reading abilities as they interact with other language abilities, primarily writing? 4) Do we want a straight power test or do we want some measure of reading rate and processing speed as well, in combination or separately? 4) Should some measure of extended reading become part of reading assessment? 5) How can a test measure the extent to which students are becoming strategic readers in the L2?, and 6)

Can we use empirical efforts to establish reading constructs for assessment purposes when the data used are typically based on traditional item types and formats of reading comprehension assessment?

Of course, these questions above are suggestive of the difficulties faced by English teachers in constructing English test items and require teachers to reflect on the implications of tests.

Although neither an innovative way of testing reading skills nor an approach to solving those dilemmas posed by Grabe (2000), multiple-choice reading tests are the most frequent method for testing students′ English in Korean public high schools; this is evidenced by a host of webpages that upload high school English tests administered by local offices of education in Korea. Thus, it seems that the reexamination of multiple choice reading tests is in order so that we can have more understanding of the issues and problems inherent in multiple-choice reading tests.

Bachman and Palmer (1996), in an effort to evaluate specific forms of language tests, suggested that teachers consider the usefulness of a test. According to them, test usefulness can serve as a basis for controlling test quality and can be subdivided into construct validity (i.e., whether or not a test measures what it purports to measure), reliability (i.e., the degree of consistency in responding to items), authenticity (i.e., the degree to which a test is related to everyday life situations), interactiveness (i.e., the extent to which testees are allowed to interact with other testees), impact (i.e., beneficial washback effects a test can bring to testees), and practicality (i.e., the efficiency of cost and effort of the test administration and the effectiveness of it). Using these criteria for test usefulness, we may find, as the single advantage of a multiple choice reading test, the efficiency of objective, quick scoring of their performance in which testees demonstrate their understanding of passages by choosing an option. The advantage itself can depend on how well a reading test has been designed. In other words, if items have been poorly designed to include double-barrel items and unclear cues, the interpretation and appropriateness of test results as well as the measure of consistency of the test may be in jeopardy and the whole usefulness of a test will be judged as low.

2. Examples of Studies with the Application of the Rasch Model

In the Rasch model, both dichotomous (i.e., true or false) data and Likert-type responses are transformed into natural logs so that actual performances of persons and difficulties of items are compared to expected probabilities (McNamara, 1996). Thus, the transformation changes dichotomous or ordinal data into probabilistic logit scores, which form a true interval scale. The practical benefits of employing the Rasch model include its abilities to

identify misfit items that behave erratically both in person ability and item difficulty terms and to present graphically the difficulty gap between items for understanding hierarchy among items. In addition, the Rasch model provides features that allow researchers to assess item bias between groups (Bond & Fox, 2007).

Across various disciplines, researchers have been applying Rasch analysis to their studies. To give a few examples of educational studies, Hillocks and Ludlow (1984) examined the hierachical and taxonomic characteristics of items related to reading L1 English literature. Through Rasch analysis, the researchers were able to confirm their hypotheses on reading difficulty. Specially, the item difficulty measure in their Rasch analysis showed the order of difficulty in reading literature could be predicted. According to the results of the study, a failure to deal with the literal level led to another failure in making appropriate inferences. If readers were not able to process basic stated information from the text, then they did not locate key details. This failure to locate key information again resulted in readers′ misunderstanding of stated relationships. Their inability to make many simpler inferences about characters, events, settings, and their relationship made it difficult for readers to understand implied generalizations by an author.

Kulikowich, Mason, and Brown (2008) studied the hierarchical relationships in variables related to L1 English compositions. They asked 72 fifth-and sixth-grade elementary students to write compositions describing real-world problems and how mathematics, science, and social studies information could be used to solve those problems. The researchers′ Rasch analysis of written compositions showed that their subjects found writing an introduction the easiest task, followed by framing mathematics, framing science, framing social studies, and closure. Studies by Hillocks and Ludlow (1984) and Kulikowich, Mason, and Brown (2008) demonstrated that Rasch analysis can be used effectively in research studies aimed to identify hierarchical relationships among constructs or items.

In a pre-test and post-test research design, Lee (1995) tested whether or not reading strategy training for understanding works of literature had any effects on reading scores. The experimental groups in the study were exposed to various classroom activities intended to improve the subjects′ reading strategies. The activities included small group discussions, discussions focusing on student-generated questions, talk about signifying, a focus on the talk in the text, and justification of interpretations through reference both to the text and to the real-world knowledge. Their Rasch analysis results indicated that the subjects gained statistically significant gains in the categories of complex implied relationships (2.5 logits), stated relationships (1.6 logits), and simple implied relationships (1.1 logits).

The study by Waugh (1999) can be another example of the application of Rasch analysis. The research ran a Rasch analysis of the 40 attitude items (prior to studying) and the 40

behavioral items (during studying) answered by his subjects. The initial analysis prompted the researcher to discard 12 attitude items and 12 behavioral items. The 56 newly created interval level scale formed a good scale with satisfactory psychometric properties.

As the above examples of studies based on the Rasch analysis indicate, the Rasch analysis can be applied to understanding hierarchical relationships in variables, testing effects of treatment in pre-test/post-test as well as quasi-experimental, experimental research, and identifying Likert-type scale items or binary (i.e., true or false) test items that do not fit the Rasch model or that do not discriminate high ability persons from low ability persons.

III. METHOD

1. Participants

The data on midterm English were collected from a high school in Jeonju, Korea. The 256 subjects in this study were senior co-ed high school students after 4 students were excluded from the study because they either scored perfectly or answered all the items incorrectly. There were 85 boys and 171 girls in all. At the time of the current study, the subjects had been studying English for a minimum of 10 years since their 3rd year elementary school year. They had been taking English classes focused on developing their reading skills 3 times a week.

2. Procedures

As a part of their curriculum, the subjects took their English midterm on September 20th. The test items had been developed by a team of teachers who were teaching reading skills. The original test included 17 multiple choice items and 17 short answer questions. However, for the present study, only the 17 multiple choice items marked dichotomously were selected. All of the items were intended to measure skills related to reading such as 'understanding the flow of paragraphs', 'finding main ideas of passages', 'filling in words based on the understanding of paragraphs' and so on.

3. Data Analysis

SPSS version 17 was used for the entry, storage, and retrieval of midterm data. In addition, Winstep, a software for the Rasch model analysis, was used for various data analyses in the present study. In testing and assessment literature, data analyses based on Winstep are frequently reported because of the program's various functions in analyzing dichotomous and polytomous data in a true interval scale.

IV. RESULT

1. The Initial Data Analyses for Identifying Misfit Items

In answering research question 1 (i.e., how do the test items behave in the Rasch model analysis?), several statistics were analyzed. The initial analysis results on item fit are reported in Table 1. Item difficulty estimates ranged from -1.30 logit to 1.17 logit with the mean located at 0 logit. The standard deviation for the item difficulty estimates was .78, indicating that most items were between +1 and -1 logits. The infit MNSQ statistics ranged from .82 to 1.31. However, infit ZSTD values for items 8, 10, 12, 15, and 17 were outside of the commonly accepted range of -2 to +2. Summary output on items was as follows: Model RMSE .15, Adjusted SD.76, Separation 5.03, and item reliability .96. The Chi-Square of 4638.40 was statistically significant at alpha 0.001 level with 4074 degree of freedom. Summary output on persons showed Model RMSE .62, Adjusted SD 1.04, separation 1.72, and reliability .75. The overall results of initial analysis showed that items of 8, 10, 12, 15, and 17 did not fit the Rasch model. This suggested that these items diverged from the expected difficulty pattern within the data. Accordingly, following initial analysis, these items were removed completely so that the interpretations of output were more meaningful. As for the removed items, content analysis of them was conducted to study why the items failed to conform to expectations. The content analysis is reported in a later section.

Table 1

Fit Indices

Entry number	Measure	Model S.E	Infit MNSQ	ZSTD	Outfit MNSQ	ZSTD	PT-Measure
1	-.17	.14	.96	-.7	.93	-.8	.49
2	1.15	.16	1.07	.8	1.05	.4	.42
3	.01	.14	.98	-.4	.99	-.1	.48
4	-1.17	.15	1.05	.8	1.02	.2	.37
5	.49	.15	1.03	.5	1.10	1.0	.44
6	-.19	.14	.90	-1.9	.84	-1.8	.54
7	-.89	.14	.96	-.6	1.22	1.7	.42
8	.60	.15	.86	-2.1	.76	-2.5	.59
9	-.17	.14	.96	-.8	.94	-.7	.49
10	-.15	.14	1.15	2.7	1.27	2.8	.34
11	-1.30	.15	.98	-.2	1.15	1.0	.38

12	.77	.15	.82	-2.5	.83	-1.5	.60
13	1.17	.16	1.13	1.4	1.20	1.4	.37
14	.59	.15	.91	-1.4	.91	-.9	.54
15	.59	.15	1.31	4.1	1.47	3.9	.24
16	-1.30	.15	.98	-.3	.94	-.3	.41
17	-.06	.14	.87	-2.4	.79	-2.5	.57
Mean	.00	.15	1.00	-.2	1.02	.1	
S.D	.78	.01	.12	1.7	.19	1.7	

2. The Reanalysis of the Data after Removing Misfit Items

To investigate how other items, other than ones disqualified by initial analysis, behave in the Rasch model, 12 remaining items were submitted for the reanalysis. In particular, the interplay of person ability and item difficulty, formation of hierarchy among items, and item bias through DIF analysis were examined. Table 2 shows the results of the reanalysis of the items. As infit ZSTD value in TABLE 2 shows, item 6 overall did not fit in the second analysis, indicating that there may be still another dimension of a small amount that this Rasch model did not account for. However, with the purpose of finding general fit of the model in mind and the consideration that item 6 was inside the acceptable infit MNSQ range, we decided to keep item 6 in the interpretation of the model. According to the results of the reanalysis, item difficulty estimates ranged from -1.19 to +1.37 with the mean of 0 and the standard deviation of .88. The reliability of the item difficulty estimates and separation was .97 and 5.62 respectively, suggesting that some items were more difficult than others and the order of item difficulties would be consistent across another group of persons of equal ability if the same items were tested again. Log-likelihood Chi-square was 3160.83 with DF of 2747, which was statistically significant at alpha .001 level. Model RMSE was .15. Person ability estimates showed the mean of -.05, the standard deviation of 1.21, the range of -2.70 to 2.70, model reliability of .65, and separation of 1.36.

Table 2

Infit and Outfit Measures

Entry number	Measure	Model S.E	Infit MNSQ	ZSTD	Outfit MNSQ	ZSTD	PT-Measure
1	-.03	.14	.99	-.2	1.03	.4	.48
2	1.35	.17	1.09	1.0	1.10	.7	.45
3	.16	.14	1.00	.0	1.02	.2	.49

4	-1.05	.15	1.05	.8	1.09	.7	.39
5	.66	.15	1.06	.9	1.22	2.0	.45
6	-.05	.14	.87	-2.3	.81	-2.2	.57
7	-.77	.15	.93	-1.3	1.10	.9	.47
deleted							
9	-.03	.14	.94	-1.0	.92	-.9	.52
deleted							
11	-1.19	.15	.98	-.3	1.06	.4	.41
deleted							
13	1.37	.17	1.18	1.8	1.25	1.5	.39
14	.76	.15	.90	-1.4	.90	-.9	.56
deleted							
16	-1.19	.15	.96	-.5	.87	-.8	.43
deleted							
Mean	.00	.15	1.00	-.2	1.03	.2	
S.D	.88	.01	.08	1.1	.13	1.7	

The bar chart in Figure 1 confirms graphically logits zone of persons and items. In the bar chart, the items range from approximately -1 to +1 logit zone, while the distribution of persons along the ability scale is twice as wide as the items. This indicates that the test was targeted mostly for the average group or subjects. The marginal person reliability and separation index would certainly increase in a future test, given items that match high- and low- ability groups are inserted. In that case, teachers would be more confident about the replicability of person ability measures across other tests.

The 12 items of the reanalysis formed a hierarchy of item difficulties. The most difficult items turned out to be 13, 2, 14 and 5. Those items seemed to aimed at testing the abilities of the testees on relating idea units in a passage, mostly requiring the testees of top-down processing of texts. For example, to answer item 13 correctly, the testees needed previous understanding of what credit cards can do and how banks charge their customers for loaning money. The second most difficult items included items 3,1, 6, and 9 that tapped their abilities to analyze grammar. For example, the testees had to know advanced language structures (i.e., subject-verb inversion, present participle, and relative pronoun 'what') to answer item 9 correctly. The easiest set of items included 7, 4, 11, and 16 and they had to do with identifying types of discourse and main ideas of passages. For example, item 7 prompted the subjects to find the topic sentence for the passage.

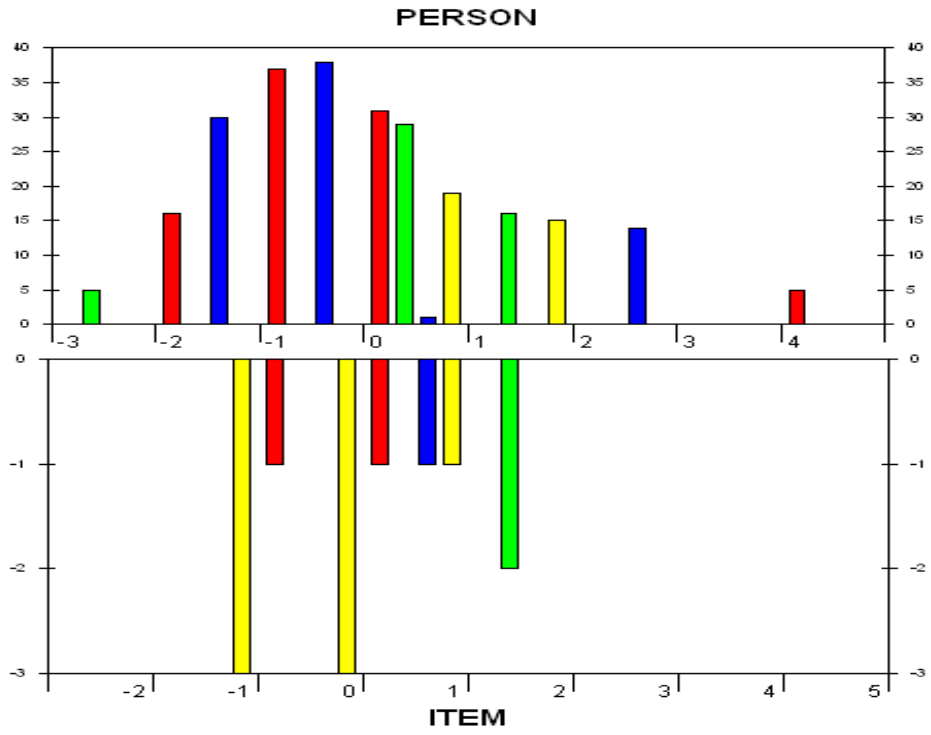

Figure 1. *Persons and items bar chart.*

3. DIF (Differential Item Functioning) Analysis

In answering research question 2(i.e., do items present any gender bias?), DIF analysis was conducted to see if there existed any differences in the way boys and girls responded to the items. All the items did not show any sign of test bias except for item 16. For item 16, girls' logit value was -1.45 with sd of .10, while boys' logit value was -.78 with sd of .24, indicating that girls found the item much easier. The difference was a t-value of 2.15 (p= 0.05) and Mantel Hanzl significance of 0.05 level. Overall, the set of test items indicated no test bias against gender, which can be interpreted to suggest that the test was fair to both groups of gender. In Figure 2, 1 represents male subjects, while 2 represents female subjects.

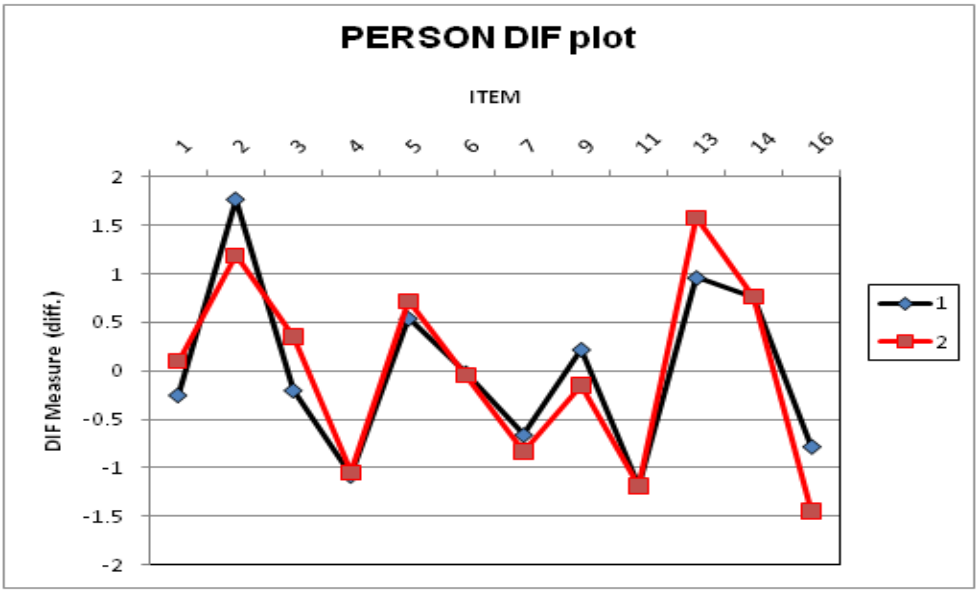

Figure 2. *Person DIF plot*

4. Content Analysis of Items and Reading Texts

Following item fit statistics that suggested misfit items of 8, 10, 12, 15, and 17, content analysis of those misfit items were conducted. The content analysis showed that there were at least a couple of correct answers to each question and/or paragraphs were illogical, insensitive to cultural practices, or causing some confusion among testees. For this item content analysis, each item is presented first followed by analysis.

Question 8. Choose a correct word for each (A), (B), and (C).

Some people believe that lying covers not only what you say, but also what you choose not to say. If you are trying to sell a car that burns a lot of oil, but the buyers don't ask about that particular **(A)[feature/ fracture]**, is it a lie not to tell them? In the United States, a favorite place to **(B)[release/ withhold]** the truth is on people's income tax returns. The government considers this an unquestionable lie, and if caught, these people are severely punished. If **(C)[omission/ addition]** can be lying, history books are great liars. Until recently, most U.S. history textbooks painted Christopher Columbus purely as a hero, the man who "discovered America," and had nothing to say about his darker side.

(A) (B) (C)
① feature ---- withhold ---- omission
② feature ---- release ---- omission
③ feature ---- withhold ---- addition
④ fracture ---- withhold ---- addition
⑤ fracture ---- release ---- omission

Suggested correct answer key: 1

In question 8, the intended correct answer was choice 1. However, some sentences in the paragraph are not clear, causing some confusion. For example, "In the United States, a favorite place to withhold the truth is on people's income tax returns." This sentence seems to give the reader a wrong impression of Americans as if they liked to withhold the truth when it comes to filing tax returns and risk being caught. This kind of wrong sociocultural information in the reading texts may have students cause grave social blunders as these students are engaged in a conversation with Americans with regard to filing tax returns. One may argue that any American, just like any Korean, would be interested in paying less tax money, while not violating their tax law. Yet, pinning down their feeling in reference to filing tax returns as their 'favorite' lie would be an unfounded, strong accusation. And a grammatical error was found in line 5. The error might have prompted the testees to wonder how 'omission' or 'addition' can be lying? The correct English would be 'If omission or addition of information can be considered as kinds of lying'.

Question 10. Choose an expression that best describes the following paragraph.

Rex McPherson was a third-generation citrus grower in central Florida. In the early 1980's, he lost 85% of his citrus stock due to severe freezes two consecutive years. This loss forced him to re-think his whole citrus-growing concept. Rex realized that the trees planted by his grandfather in the 1930's and 1940's had been placed fairly far apart because land at that time was cheap. Land values have skyrocketed since then, and he realized that if he wanted to stay in the citrus business he had better re-think his concept. He decided to use new hybrids and irrigation techniques in order to plant the trees close together. As a result, not only has it helped to inhibit freezing, but provided Rex with the impetus to achieve growth and prosperity.

① 고정 관념을 버려야 발전할 수 있다. (Translation: one may make progress only after evading preconceptions of things or people)

② 위기는 새로운 기회가 될 수 있다. (Translation: a crisis brings an opportunity)
③ 옛 것에서 배우려는 자세는 미래 발전에 도움이 된다. (Translation: an attitude to learn from the past helps one make progress)
④ 이미 일어난 일은 돌이킬 수 없다. (Translation: one thing that has already occurred cannot be reversed)
⑤ 세상사는 마음먹기에 달려있다. (Translation: everything depends on how you think)

Suggested correct answer key: 2

In question 10, choices written in Korean seemed to be the cause of confusion. Choice 1, 3, and 5 may be as good as 2. For example, choice 1 makes a good sense as much as choice 2 as Rex McPherson made a progress by adapting to new hybrids and irrigation techniques. In the same sense, choice 5 can be accepted as correct a key as choice 2 for Rex McPherson did not want to remain frustrated over frozen citrus and thought of the way to improve his farming. In line 8, 'it' is grammatically wrong. Instead, 'it' should have been a plural reference pronoun or 'multiple techniques'.

Question 12. Choose the best clause for the blank that would make the whole paragraph meaningful.

Once there was a little boy who became angry so easily and called other names. His father gave him a bag of nails and told him that every time he lost his temper, he must hammer a nail into the back of the fence. The first day the boy had driven 37 nails into the fence. Over the next few weeks, the number of nails hammered daily gradually dwindled down. Finally the day came when the boy didn't lose his temper at all. He told his father about it and the father suggested that the boy now pull out one nail whenever he could hold his temper. The days passed and the young boy could tell his father that all the nails were gone. The father took his son by the hand and led him to the fence. He said, "You have done well, but look at the holes in the fence. ( ), they leave a scar just like this one."

① If you do a lot of mistakes to reach your goal
② If you drive as many nails as you can
③ When you say things in rage
④ When you pull out some nails from the fence
⑤ When you do bad behaviors in joy

Suggested correct answer key: 3

As for Question 12, one may argue that choice 2 is as good as 3 since 'driving as many nails as you can' means 'saying things in rage' in a metaphorical sense. Some Americans would choose to say choice 2 over choice 3, given the context of the story. Two grammatical errors were found in the passage. In line 1, 'and called other names' may have been intended to say 'and called other people names'. In line 9, 'they leave a scar just like this one' should be corrected as 'they leave scars just like this one'.

Question 15. Choose a wrong usage of grammar.

Suppose a friend approaches you after class and remarks that your party last week was terrific. This remark causes you to remember ① meeting a very attractive person at your party, which in turn reminds you ② to ask this person for a date. This whole thought process reflects the concept of the association of ideas. Two events ③ become associated with each other, and thus thinking of one event automatically leads to ④ (recalling others). Aristotle proposed that in order for an association to develop, the two events must ⑤ be temporally paired, that is, occur together, and either similar to or opposite each other.

Suggested correct answer key: 4

Question 15 was intended to assess grammar skills in reading. However, the intended readership does not seem to fit with Korean high school students, who are less likely to have conversations introduced in the text. This example seems to go against modest cultures of Korean high school students. This item of 15 might have failed to fit the data because the testees knew collocation of 'remind' and the preposition 'of' better than the less frequent collocation of 'remind' followed by object and 'to infinitive', referring to the future event or action. In addition, some testees might have been confused as to whether or not 'leads to' should be followed by root form or ing-form. In this case, however, when 'lead' means 'to induce or cause, it is typical that 'lead' is followed by object and to infinitive as in 'Subsequent incidents led us to believe his arguments.' In particular, in this pattern, 'lead' most of the time requires a noun or pronoun before a to-infinitive, which is also known as the accusative and infinitive construction. Accordingly, the deletion of the accusative case was undesirable. Yet, in colloquial American English, native speakers of English use 'lead to –ing form' all the time, making the suggested correct answer 4 wrong. Accordingly, it can be argued that there is no correct answer to question 15.

Question 17. Choose the best clause for the blank.

One woman was especially attentive while a noted educator was giving a lecture on the importance of training children. After the talk, she came up to the podium and asked him, "How early can I begin the education of my child?" "When will your child be born?" he queried. "Born?" she exclaimed, "Why, he is already five years old!" "My goodness," the expert cried, "don't stand there wasted the best five years!" He is quite right! Difficult as it is for us to realize, a tiny baby in the crib is even then beginning to take on the outlook and character that will shape his or her life. We should teach our children _____ .

① when they are ready to receive external input.
② when they are old enough to understand what they see.
③ when they can handle some instruments.
④ when they get some physical maturation.
⑤ when they are very young.

Suggested Correct answer key: 5

In Question 17, the whole paragraph is confusing. In the first sentence, the author indicated that a noted educator was providing a session on the importance of training children right after they are born. However, the testees are hinted that the noted educator is somebody who knows all. How can be the noted educator be such a dogmatic person? How could one say something like that directly without considering the other person's face or feelings? This kind of pragmatically wrong conversation can give testees a wrong impression that they should be direct in conversing in English without considering sociopragmatic aspects of interaction. In addition, the sentence "don't stand there(, you) wasted the best five years!" has another grammatical mistake, perhaps adding more confusion to the testees, who tend to trust the authority of written text in a test material. Another flaw in Question 17 is that the correct answer may be multiple. 'A tiny baby in crib' can be ready to receive external input, may be old enough to understand what they see, and may have already achieved some physical maturation to see and understand what is going on around him or her.

V. DISCUSSION

Although the Rasch model analyses of various proficiency tests have been done continuously in order to study how items work with testees as well as to bank test items, classroom-based achievement tests received little scrutiny. The present study of teacher-

constructed items found that classroom-based achievement tests may be problematic for the following reasons: multiple misfit items due to unclear answer choices, off-the-target items that were either too easy or too difficult, and lack of various types of items to cover sub skills of reading.

The most crucial problem of the teacher-developed test in the present study was the existence of misfit items. The close examination of the misfit items revealed that the teachers made mistakes in composing multiple choices that had multiple answers to the questions along with some grammatical errors in the passages. This may have done more damage than the English teachers can imagine. In other words, the interpretation made with regard to the achievement test performance based on all of the items including the 5 aberrant ones is neither valid nor reliable. So, the result of misfit analysis suggests that English teachers should bear more responsibility for constructing items for school-wide achievement tests, perhaps by having a teacher conference to go over test items before printing, if necessary, together with proficient native or near-native speakers of English.

It is obvious that many subjects in this study were off the target of the items that were intended to measure their achievement levels. 49 subjects, about 20% of the total number of subjects, were challenged by either too easy or too difficult items, an indication that the teachers may have provided their students with items without knowing clearly how challenging or easy each item was. Thus, it seems necessary that teachers hold meetings to critique and analyze item difficulties after marking responses and conduct focused group interviews with students in order to recognize which items have been challenging for testees and which have been too easy.

The present study also found that test items were limited to questions related to discourse genre or discourse structure, grammar, and reading for details. The test items may have been narrowed down to the types of questions used perhaps due to the fact that the length of each passage used in the test was only in a single paragraph. However, it is expected that high school students in Korea are able to read long English texts of multiple paragraphs, a skill that can be directly related to reading authentic materials in the real world. So, rather than limit reading passages to a paragraph length, the teachers might have given the testees longer passages to assess various reading skills including guessing and inferring information.

DIF (Differential Item Functioning) analysis of items for any test bias revealed that overall the test had little bias toward gender. Only item 16 was statistically significant in the DIF analysis. On item 16, girls had higher ability than boys, though it is not clear why girls excelled. However, teachers may continue to monitor potential test bias toward boys or girls to make sure that tests do not show any bias toward any groups and are fair to everybody.

With the application of the Rasch model, this study also was able to indentify hierarchical

order of reading aspects. The testees found items related to discourse structure easier than items requiring them either to tap their world knowledge or grammatical competence. This result suggests that English teachers teach paragraph or essay structures without worrying that students may have difficulties mastering discourse structures as evidenced by their ability to recognize what is required of a paragraph or an essay.

VI. CONCLUSION

English as a foreign language in Korean contexts has a great influence on high school curriculum as it is tested both at classroom level and at national college entrance exams. However, an examination of test items constructed by Korean inservice teachers of English revealed that about 30% of them could lack consistency in measuring reading ability. To make classroom-based English achievement tests valid and reliable, first, Korean inservice teachers of English need to understand implications of bad items bearing on test takers who may be put at a disadvantage because of systematic errors in measuring their English reading ability. Second, Korean teachers of English should be reminded of basic concepts of testing and assessment including issues of validity, reliability, item analysis techniques in teacher seminars. More specifically, practical application of Rasch analysis of items can bring new insights into how items have behaved among test takers. Lastly, Korean inservice teachers of English should conduct frequent focused group interviews with test takers to understand how items have worked for them. Any qualitative data from interviews may reveal some errors that were not obvious at the stage of item construction.

REFERENCES

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bond, T., & Fox, C. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences (2nd ed.)*. New Jersey: Lawrence Erlbaum.

Brown, D. (2004). *Language assessment: principles and classroom practices*. New York: Person Education.

Brown, A., & Hill, K. (2007). Interviewer style and candidate performance in the IELTS oral interview. In L. Taylor & P. Falvey (Eds.), *Studies in language testing (19) IELTS collected papers* (pp. 37-60). Cambridge: Cambridge University Press.

Grabe, W. (2000). Reading research and its implications for reading assessment. In A. Kunnan (Ed.), *Studies in language testing (9): Fairness and validation in language assessment* (pp. 226-262). Cambridge: Cambridge University Press.

Hillocks, G., & Ludlow, L. (1984). A taxonomy of skills in reading and interpreting fiction. *American Educational Research Journal, 21*(1), 7-24.

Hughes, A. (2003). *Testing for language teachers*. Cambridge: Cambridge University Press.

Kulikowich, J., Mason, L., & Brown, S. (2008). Evaluating fifth-and sixth-grade students' expository writing: Task development, scoring, and psychometric issues. *Reading and Writing, 21*, 153-175.

Lee, C. (1995). Signifying as a scaffold for literary interpretation. *Journal of Black Psychology, 21* (4), 357-381.

Merryless, B., & McDowell, C. (2007). A survey of examiner attitudes and behavior in the IELTS oral interview. In L. Taylor & P. Falvey (Eds.), *Studies in language testing (19) IELTS collected papers* (pp. 142-178). Cambridge: Cambridge University Press.

McNamara, T. (1996). *Measuring second language performance*. New York: Addison Wesley Longman.

Moore, T., & Morton, J. (2007). Authenticity in the IELTS academic module writing test: A comparative study of task 2 items and university assignments. In L. Taylor & P. Falvey (Eds), *Studies in language testing (19) IELTS collected papers* (pp. 197-248). Cambridge: Cambridge University Press.

Waugh, R. (1999). Approaches to studying for students in higher education: A Rasch measurement model analysis. *British Journal of Educational Psychology, 69*, 63-79.

Jaewoo Shim
Chonbuk National University
664-14 Duckjin-Dong, 1Ga, Jeonju,
561-756, Korea.
Tel: (063) 270-2728
Email: shimjw@jbnu.ac.kr

Keiko K. Samimy
1945 N. High St.
Columbus, OH 43210-1172, USA.
Tel: (614) 292-7597
Email: samimy.2@osu.edu

Received 28 May 2010
Revised 23 July 2010
Accepted 17 August 2010