A Rasch Analysis of a Nationwide English Placement Test in Korea

Munhong Choe
(University of Hawaii)


This study analyzes a nationwide English placement test in the Rasch framework. Data were obtained from 297 first-year junior high school students. The assessed person (.75) and item reliability (.95) indicate that the test was fairly consistent and reproducible with other samples of examinees. About 60% of total variation was explained under the assumption of unidimensionality. The person separation index suggests that two and a half ability levels can be differentiated by means of the test. Overall, the ability distribution of students was higher than the difficulty distribution of items. The fit statistics identified a few misfit items, but their impact on the utility of the test appears nonsignificant.

I. INTRODUCTION

In this study, I carry out an analysis of an English placement test in Korea on the basis of the Rasch model. The test was administered by Korea Institute for Curriculum and Evaluation to first-year junior high school students (corresponding to 7th graders in K-12) nationwide in March 2008, with the express purpose of classifying schools by academic levels. It was a paper-based test with 25 items consisting mostly of comprehension

---

1 I refer to this test as a placement test with no connotation or further justification for the term over other alternate or overarching terms such as proficiency. It is my opinion that because the test, which was composed of 25 written multiple-choice items, may well be thought of as having tested students’ competence in a specific domain of knowledge rather than their overall performance, and besides its main purpose and use were oriented to identify differences between schools, it would be more informative to call it a placement test instead of a proficiency test, though admittedly the latter could be less controversial in contemporary discourse.
questions. Data were collected from 297 students in an urban area and entered into one-parameter Rasch analysis by using a software packet, Winsteps. The present analysis is a preliminary step. Discussion will center on test results and inferences. Subsequent studies may concern further validation or refinement of the original test. The goal is to exemplify a use of modern item response theory (IRT) for a school-level analysis and development of English tests. I will touch upon such questions as what other aspects of measures distinct from classical theory a Rasch approach deals with, what interpretations its item difficulty and person discrimination indices bring forth, what can be inferred about individual items from its model fit statistics, etc. The primary question comes down to whether this kind of analysis is useful enough as a tool of the trade in English testing.  

II. LITERATURE REVIEW

In this section, I will review some of the previous studies on placement tests, and then offer a brief explanation about Rasch models along with their applications to language testing.

Tests have different goals, functions, and decisions made on the basis of the results. The purpose of placement tests, as the name implies, is to assess students’ level of language ability in order to place them to an appropriate class that meets their needs. The placement of students in a reliable and efficient way has been a big challenge faced by many schools and institutions.

Placement tests have received relatively little attention in language testing research. Most previous studies have been concerned with (re-)evaluating and validating tests as an instrument for classifying students in an efficient way. Researchers (e.g., Alderson et al., 1995; Bachman, 1990; Bachman & Palmer, 1996; Brown, 1996; Heaton, 1990, among many) agree on two fundamental roles of placement tests: (a) to group students who are homogeneous in their language ability, (b) to ascertain students’ mastery of the prerequisite subjects. Although schools aim to teach English under the paradigm of communicative language teaching, few use tests that actually reflect the curricular foundation. They are often based on literacy skills, relying for the most part on structural features such as

---

2 As a reviewer correctly noted, this study lacks an independent motive other than practicing another tool for analysis of an end product. There could be no objections to the claim that questions precede methods. However, it is also true that available methods determine valid questions in the context of science where mathematical idealization and quantification are the fundamental grounds of reasoning. The attempt made here is rather at exploring a method and its utility than at formulating questions and finding answers to them. Premature at this stage, it is hoped to be a meaningful step towards the latter.
grammar, vocabulary, translation, and deterministic, as opposed to negotiative, comprehension, which have been dominant long before foreign language teaching emerged as an academic discipline. Clearly, this is in part because it requires a greater amount of time and effort for an individual school to develop and administer communication-oriented tests.

In a similar vein, Brown (1989, 1992, 1996, and subsequent work) emphasized curricular implications in placement test development. First, placement tests consider students’ initial state as well as practicality issues such as time, cost, facilities, number of test-takers and proctors, etc. Second, they intend to measure students’ prior knowledge that pertains to the goals of the program. For example, if the program is designed to develop students’ communicative skills, its placement test is supposed to reveal them in the first place. Third, it is often stressed, yet neglected as well that tests need to be piloted, analyzed, and revised continuously in order to measure students’ latent abilities as accurately as possible.

This quite intuitive remark demands a good placement test to satisfy two criteria: (a) whether it makes plain the dimensions of what it plans to measure, and (b) whether it draws reasonable inferences about the mastery of knowledge or skills that are taught in the program. Besides, students placed in a group are assumed to benefit from studying the same material. It is thus important to ensure in advance that items on the test can be ranked in terms of difficulty, and that this ranking applies to all learners in the program. This gives a reason to posit that both difficulty of test items and ability of test-takers are located on an identical scale, whereby students can be ranked consistently with respect to the range of skills, knowledge, or abilities being tested (See Henning (1987), for a detailed argument).

This reasoning behind IRT allows researchers to assess test items in a more systematic way. For example, Wall, Clapham, and Alderson (1994) examined internal validity and reliability of an institutional placement test that consisted of grammar, writing, reading, and listening tests. All the subtests’ reliability estimates were above .75, whereas the concurrent validity obtained from the correlation between subtest scores and students’ self-assessments turned out to be rather low. Another example is Fulcher (1997) who analyzed data from a placement test that consisted of two essay questions, ten items for grammar, and eight items for reading comprehension. He deduced that a degree of reliability would increase over .80 if the test length were doubled. Kyong-Hyon Pyo (2001) assessed construct validity of essay questions in a university ESL placement test. Using multi-trait multi-method and structural equation modeling, she tested the unitary trait hypothesis of English proficiency by estimating the effects of students’ listening and reading comprehension scores on their essay writing performance. The effect size was alike on both pass and non-pass groups while the two were not homogeneous in their three language abilities. These studies illustrate some practical uses of IRT and advanced
statistical methods in the process of (in-)validating placement tests.

The Rasch model (Rasch, henceforth) was proposed by Danish mathematician Georg Rasch (see Rasch (1980), for its theoretical groundwork). He first started with dichotomous data. Since then, researchers have elaborated the model by incorporating measurement using rating scales (e.g., Likert scales) and partial credit responses. Other facets such as raters and subtasks can now be taken into account in addition to person and item.

Rasch is logit-linear. It represents item difficulty and person ability as probabilities (i.e., odds) that are converted into logarithms. The log units, namely logits, are conceived as an infinite interval scale. It tells us not only whether one item is more difficult than the other but also how difficult it is on a quantified scale. On the logit scale, the average of item difficulty is set at zero so that an item is difficult if it is above the mean (i.e., greater than zero) and it is easy below the mean (i.e., smaller than zero). The logit scale measures the relation between item difficulty and person ability. If a person and an item are at the same location on the scale, the person has a 50% chance of success on the item. In short, Rasch parameterizes person ability and item difficulty separately and assumes that a functional relation between them determines the probability of one's getting an item correct.

It has been argued that Rasch has advantages over classical theory. For example, classical theory is dependent on students who take the test. If the same items were administered to a different group, the estimated item difficulty might vary by groups. It is thus untenable to apply the same items to different groups in other contexts. That is, classical theory is not competent enough to make generalizations from a sample to a new set of persons and items. On the other hand, Rasch, which is essentially probabilistic, is capable of estimating the ability of a person in relation to the universe of parallel test items and the difficulty of an item for the population of prospective test takers.

Moreover, since item difficulty in classical theory is estimated in terms of the proportion of correct answers by the whole group of test takers, it does not provide information about how an individual test taker performs (or is likely to perform) on a specific item. Difficult items for a group may be easier for another group with higher ability. As a result, item difficulty and person ability are not coordinated. Rasch provides a solution for this problem by rating the two on the same scale.

In spite of these advantages over classical theory, Rasch is not without controversy. Above all, Rasch presupposes a single dimension of person ability and item difficulty. As Inn-Chull Choi and Bachman (1992) mentioned, due to the axiomatic assumption of unidimensionality, it has been called into doubt whether using Rasch in language testing is appropriate.

Davies et al. (1999) construed unidimensionality as a single measurement trait or pattern that is sufficient to account for examinees' test performance. According to Henning et al.
(1985), unidimensionality is merely a statistical assumption, so it does not imply that a test must perfectly measure one and only one trait, nor does it endorse the unitary trait hypothesis. McNamara (1996) distinguished two models of unidimensionality: (a) measurement models that refer to a single underlying measurement dimension, and (b) construct models that consist of various skills underlying test performance. The assumption is met when it is possible to sum up scores in different parts of the test in a meaningful way and when the examinees can be measured using the same terms.

Bond and Fox (2007) illustrated the notion of unidimensionality with an example. Length, height, weight, and volume are characteristics of a rectangular solid, but measuring the rectangular as an object and obtaining a meaningful estimation, the focus cannot but be on one characteristic at a time. Unidimensionality is rather an analytic approach to a complex phenomenon by way of measuring one characteristic at a time. Beyond the ability that is measured, it is entirely possible that other factors (e.g., native language and cultural background, world knowledge, nonlinguistic strategies, anxiety, etc) intervene in testing procedures, but if the factor that is intended to measure influences test takers’ performance, the assumption of unidimensionality can be at least partially explained.

Researchers (e.g., Brown & Hudson, 2002; Inn-Chull Choi & Bachman, 1992; McNamara, 1991) take different approaches to investigating dimensionality of a test. Factor analysis (and other variants of canonical correlations) is generally credited as a hands-on tool for checking how many dimensions a test measures. If one factor appears to be more dominant than the others, the test complies with the assumption of unidimensionality.

Although Rasch has been employed for over two decades in language testing, the unique diagnostic potential afforded by Rasch at the individual item level has not been given its proper due (Clark, 2007). Exploiting the properties of Rasch does not simply mean using a Rasch approach to analyzing test results. Rasch renders the requirements of measurements in testable form from a practical perspective. When it holds, Rasch certifies the measurement. If it does not hold, the misfit between the model and the actuality of the data calls for an investigation.

In view of Rasch, it is not the model that should be questioned, but rather the data themselves. In constructing data to fit the model, it is likely for the model to disclose anomalies in the data – anomalies that must be understood and for which the model provides clues as to where to look (Andrich, 2004). So the multi-faceted nature of language does not preclude it from being measured by tests which require the assumption of unidimensionality. Rasch is not a theory that argues for the unidimensionality of what to be measured, but rather it is an approach to a complexity on the basis of simple mathematical principles (cf. Galilean approach).
III. METHOD

1. Purpose

As aforementioned, this study applies Rasch to the analysis of an English placement test in an EFL context both at the whole test and individual item levels. The Rasch-based statistical results and graphic figures are presented and discussed in some detail in terms of overall model fit and distribution estimates. The second part of discussion brings up individual items that were identified as misfits and suggests a few modifications to such items.

2. Participants

The test under analysis was administered nationwide in March, 2008. Since tests of the kind had never been conducted for the past ten years, it caused social concern for a period of time. The purpose of the test was to categorize students and schools by several levels and use the information for differential government support and teacher assignment in the future. The test results were obtained from 297 (181 male, 116 female) first-year students in a (nonspecific) middle school.

IV. RESULTS AND DISCUSSION

1. Person-Item Reliability and Separation

Every statistical analysis has its particular underlying assumptions to be fulfilled in order for the results to make meaningful interpretations. It is thus important, albeit often overlooked, to ascertain that data meet the requirements for the analytic procedure in use. Rasch works on the assumption that the test is not speeded. Incorrect responses indicate lack of ability rather than lack of time. For the test at issue, students were given 45 minutes to complete 25 multiple-choice questions that uniformly had one correct answer out of five choices. Under normal conditions, each item would not take more than one minute. There was no single blank response on the test out of a total of 7425 responses (297 students × 25 items), which confirms that the test was not speeded.³

Reliability is calculated as a proportion of true variance to total variance in the test scores. The resulting reliability estimate is an index of how consistent the scores from the test were. In Rasch analysis, reliability is assessed with respect to items as well as persons. The following table shows summary statistics on person measures. The person reliability is analogous to Cronbach’s alpha in classical theory. As seen in Table 1, it showed a fair
A Rasch Analysis of a Nationwide English Placement Test in Korea

degree of reliability (.73) given the relatively small number of items on the test. This index indicates the replicability of person ordering we could expect if this sample of persons were given another parallel set of items measuring the same construct.

TABLE 1
Summary of 297 Measured Persons

<table>
<thead>
<tr>
<th>Raw Score</th>
<th>Count</th>
<th>Measure</th>
<th>Model Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>19.6</td>
<td>.25</td>
<td>2.23</td>
</tr>
<tr>
<td>S.D.</td>
<td>6.4</td>
<td>.0</td>
<td>2.04</td>
</tr>
<tr>
<td>Max.</td>
<td>25.0</td>
<td>.25</td>
<td>4.83</td>
</tr>
<tr>
<td>Min.</td>
<td>0</td>
<td>.25</td>
<td>-4.72</td>
</tr>
<tr>
<td>Real RMSE</td>
<td>1.06</td>
<td>ADJ.SD</td>
<td>1.74</td>
</tr>
<tr>
<td>Model RMSE</td>
<td>1.05</td>
<td>ADJ.SD</td>
<td>1.74</td>
</tr>
<tr>
<td>S.E. of Person Mean = .12</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Person Raw Score-to-Measure Correlation = .95
Cronbach Alpha (KR-20) Person Raw Score Reliability = .94

In addition to reliability, a measure of person separation is also provided. The higher the separation value, the easier it becomes to distinguish between persons. It is an estimate of as it were, how reliably one can differentiate persons on the measured variable. The estimate is based on the same concept as Cronbach’s alpha. That is, it is the fraction of observed response variance that is reproducible by the model (in the above table, Separation (1.66) = Adjusted Standard Deviation (1.74) / Rasch Measured Standard Error (1.05)). The separation measure can be used to calculate STRATA, which indicates the number of distinct ability levels separated by three errors of measurement (Shumacker, 2004). The formula is STRATA = (4G_p + 1)/3 in which G_p is person separation index. This results in a value of 2.55, suggesting that two and a half statistically distinct groups can be identified in the data. It is likely that the separation value will increase if the number of items increases and gives more information about each test taker’s ability. Provided the formula above, a separation value of G_p=2.0 or greater will be needed to identify at least three STRATA. Applying the Spearman-Brown prophecy formula for separation (Linacre, 2000), a test length of approximately 35 similar items would have resulted in separation

---

3 Another important assumption is that the residuals of each item estimate show a normal distribution, independently of (i.e., non-covariate with) other items. This assumption is fundamental to any analysis using maximum likelihood estimation, and it is the reason that probabilistic models like Rasch require quite a big sample size, which ought not to be overlooked by multiple facets analyses (in principle, one more facet requires that amount of additional data).

4 Sep_M \approx Sep_K \times \sqrt{M/K}, Where Sep_M is the desired separation, Sep_K is current separation, and M and K are the number of items on the new and current test respectively.
reliability sufficient to identify three ability levels.

Next, we have a greater interest in the test items. Table 2 displays summary statistics on items.

<table>
<thead>
<tr>
<th>TABLE 2 Summary of 25 Measured Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw Score</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td>Mean 1.726</td>
</tr>
<tr>
<td>S.D. 25.7</td>
</tr>
<tr>
<td>Max. 212.0</td>
</tr>
<tr>
<td>Min. 108.0</td>
</tr>
<tr>
<td>Real .20</td>
</tr>
<tr>
<td>RMSE Model .19</td>
</tr>
<tr>
<td>RMSE Reliability</td>
</tr>
</tbody>
</table>

S.E. of Person Mean = .18

Item Raw Score-to-Measure Correlation = -.99
5900 Data Points. Approximate Log-Likelihood Chi-Square: 4365.71

The separation and reliability values indicate that the difficulty order of items on the test is consistent and reproducible with another sample of examinees. Winsteps also provides an estimate of item separation. One-parameter Rasch assumes a uniform item discrimination value 1.00 for all items. The separation value was 4.22 with reliability .95. This suggests that the items can be grouped into four levels of difficulty. Meanwhile, outfit is based on the sum of squared residuals for every item per each person, and the sum is divided by the number of items, hence mean squares. It is more sensitive to responses to items with a difficulty far from the person, and vice versa. For example, outfit reports overfit for imputed responses and underfit for lucky guesses or careless mistakes. On the other hand, infit uses a weighted scale. Squared residual for each item is weighted by its variance and then summed. Dividing that total by the sum of the variances leaves the differential effects of the weightings (Bond & Fox, 2007). In general, infit and outfit mean square (MNSQ) values greater than 1.3 and the standardized fit statistics (ZSTD) values greater than +2.0 or less than -2.0 are considered misfitting. These items have less compatibility with the model than expected. I return to fit statistics on each item shortly.

2. Person-Item Distribution

Figure 1 shows a Rasch person-item map, illustrating the connection between person ability and item difficulty on the same scale.
297 Persons, 25 Items, Each #: 5

FIGURE 1
Person-Item Map

The ability estimates of the 297 measured students are shown on the left side and the
item difficulty locations are shown on the right. Both increase as moving towards the top. It turned out that the majority of ability estimates fall above the range of the test items. Overall, the ability distribution is higher than the difficulty distribution. The person ability mean is above +2 logit. This implies that the items were very easy for most students. About three fourths of students lie above the ability estimates that the items can measure. Consider Figure 2.

**FIGURE 2**
Expected Score: Mean
Expected score: Mean (Rasch-score-point threshold, ":" indicates Rasch-half-point threshold)

The difficulty metric is shown horizontally along the X-axis. Items are listed along the right side with the most difficult at the top. The line of colons indicates the point along the scale at which the probability of success on the given item is 50%. The maximum of coverage and efficiency will form a perfect diagonal across the range of the scale.
gaps and stacks indicate areas in which there is no corresponding item to the difficulty level, or conversely, there are more than one corresponding items that are superfluous. For example, gaps exist between Items 11 and 1; 15 and 4; 19 and 9. The presence of item stacks, namely more than one item of the same difficulty (e.g., Items 21, 12, 18, and 8, 3, 16), is readily observed.

As discussed earlier, the fundamental assumption underlying Rasch is the unidimensionality of item difficulty and person ability. It is therefore in due course to verify that the resulting model meets the condition. There is no single, uncontroversial test for unidimensionality. In Winsteps, variance components analysis of residuals is provided as a measuring tool for the dimensionality of data. It investigates whether the measured dimension accounts for more variance than any other potential dimensions which in principle must not be significant. If two or more substantial dimensions are found, the results should be taken to be multidimensional and reported as such.

Table 3 summarizes the variance estimates of measured and non-measured contrasts.

<table>
<thead>
<tr>
<th></th>
<th>Empirical</th>
<th>Modeled</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Variance in Observation</td>
<td>65.2</td>
<td>100%</td>
</tr>
<tr>
<td>Variance Explained by Measures</td>
<td>40.2</td>
<td>61.7%</td>
</tr>
<tr>
<td>Unexplained Variance</td>
<td>25.0</td>
<td>38.3%</td>
</tr>
<tr>
<td>Unexplained Variance in 1st Contrast</td>
<td>1.9</td>
<td>2.9%</td>
</tr>
<tr>
<td>Unexplained Variance in 2nd Contrast</td>
<td>1.5</td>
<td>2.3%</td>
</tr>
<tr>
<td>Unexplained Variance in 3rd Contrast</td>
<td>1.4</td>
<td>2.1%</td>
</tr>
</tbody>
</table>

The following scree plot visualizes measured variance in comparison with non-measured potential contrasts. More than 60% of total variance is explained under the assumption of unidimensionality, whereas the contributions of the other contrasts are marginal. The percentage of variance explained by the model is far greater than any of the contrasts. This fact confirms that the variance of the data fits the model. The other contrasts are of a limited size and are not a serious threat to unidimensionality.
3. Item Fit Statistics

Fit statistics are to find out items that are misfitting to the expected model, thereby check to see if their deviation is large enough to degrade the function of the measures. Misfit may indicate that the item is poorly constructed or that it is measuring something different. Infit refers to a pattern of responses near a person's ability estimate. It gives relatively more weight to the performance that people who are closer to the item difficulty estimate put in. This does not mean that weighting itself is heavier when person and item values are closer. In contrast, weight becomes bigger when a person-item difference is bigger so that the model predicts more accurately when the person ability value is -1 and the item difficulty value is 5 than when the former is -1 and the latter is 1. The statistical procedures use fractions of distance (variance), and thus, in analysis, the closer the values between person and item, the more sensitive weighting is. On the other hand, outfit is influenced by unexpected responses that are outside the estimated confidential interval. It is for this reason that infit values should be paid more attention than outfit values.

Underfit (i.e., a lack of fit) refers to erratic items that are not predictable by reference to the model. It is detected when the fit statistics are too high to meet the model's cutoffs. Overfit (i.e., overly fit) reflects items that are short of meeting the model's expectations because of too small a variation. It indicates smaller standard errors and inflated reliability. Overfitting items are considered to be either redundant or dependent on other items around them, and are candidates for elimination or revision (McNamara, 1996).

To put it simply, overfit occurs when an item gives little information about person ability.
A Rasch Analysis of a Nationwide English Placement Test in Korea

For example, if all or none of students succeeded, the item would not be much useful. Underfit is the case when students' responses to the item cannot be accounted for by the model. For example, if a highly able student fails while a less able student succeeds at an item, the model has little to inform about the item. Table 4 exhibits fit statistics item by item.

**TABLE 4**

<table>
<thead>
<tr>
<th>Item</th>
<th>Count</th>
<th>Total Score</th>
<th>Model S.E.</th>
<th>Infit MNSQ</th>
<th>ZSTD</th>
<th>Outfit MNSQ</th>
<th>ZSTD</th>
<th>PTMFA</th>
<th>EXACT OBS%</th>
<th>MATCH EXP%</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>168/297</td>
<td>2.00</td>
<td>.16</td>
<td>1.41</td>
<td>4.8</td>
<td>2.06</td>
<td>5.2</td>
<td>A.53</td>
<td>69.5</td>
<td>76.5</td>
</tr>
<tr>
<td>6</td>
<td>202/297</td>
<td>1.06</td>
<td>.17</td>
<td>1.43</td>
<td>4.0</td>
<td>1.71</td>
<td>4.2</td>
<td>B.53</td>
<td>72.0</td>
<td>80.3</td>
</tr>
<tr>
<td>17</td>
<td>253/297</td>
<td>-.71</td>
<td>.21</td>
<td>1.13</td>
<td>1.1</td>
<td>1.48</td>
<td>1.4</td>
<td>C.51</td>
<td>83.9</td>
<td>85.7</td>
</tr>
<tr>
<td>12</td>
<td>220/297</td>
<td>.51</td>
<td>.18</td>
<td>1.36</td>
<td>3.2</td>
<td>1.43</td>
<td>2.3</td>
<td>D.54</td>
<td>74.2</td>
<td>82.2</td>
</tr>
<tr>
<td>11</td>
<td>170/297</td>
<td>1.95</td>
<td>.16</td>
<td>1.11</td>
<td>1.3</td>
<td>1.42</td>
<td>2.4</td>
<td>E.64</td>
<td>72.9</td>
<td>76.6</td>
</tr>
<tr>
<td>15</td>
<td>224/297</td>
<td>.38</td>
<td>.18</td>
<td>1.21</td>
<td>2.0</td>
<td>1.39</td>
<td>1.9</td>
<td>F.58</td>
<td>77.1</td>
<td>82.6</td>
</tr>
<tr>
<td>18</td>
<td>221/297</td>
<td>.47</td>
<td>.18</td>
<td>1.32</td>
<td>2.8</td>
<td>1.33</td>
<td>1.8</td>
<td>G.56</td>
<td>75.4</td>
<td>82.3</td>
</tr>
<tr>
<td>23</td>
<td>245/297</td>
<td>-.38</td>
<td>.20</td>
<td>1.09</td>
<td>.8</td>
<td>1.30</td>
<td>1.1</td>
<td>H.55</td>
<td>83.1</td>
<td>84.8</td>
</tr>
<tr>
<td>13</td>
<td>249/297</td>
<td>-.54</td>
<td>.20</td>
<td>1.10</td>
<td>.0</td>
<td>1.05</td>
<td>.3</td>
<td>I.55</td>
<td>82.2</td>
<td>85.2</td>
</tr>
<tr>
<td>1</td>
<td>202/297</td>
<td>1.06</td>
<td>.17</td>
<td>1.05</td>
<td>.6</td>
<td>1.06</td>
<td>.5</td>
<td>J.66</td>
<td>80.5</td>
<td>80.3</td>
</tr>
<tr>
<td>10</td>
<td>258/297</td>
<td>-.93</td>
<td>.21</td>
<td>1.01</td>
<td>.2</td>
<td>.84</td>
<td>-.3</td>
<td>K.53</td>
<td>86.4</td>
<td>86.5</td>
</tr>
<tr>
<td>19</td>
<td>258/297</td>
<td>-.93</td>
<td>.21</td>
<td>1.95</td>
<td>-.4</td>
<td>.78</td>
<td>-.5</td>
<td>L.55</td>
<td>86.4</td>
<td>86.5</td>
</tr>
<tr>
<td>8</td>
<td>247/297</td>
<td>-.46</td>
<td>.20</td>
<td>.95</td>
<td>-.4</td>
<td>.86</td>
<td>-.4</td>
<td>M.60</td>
<td>84.7</td>
<td>85.0</td>
</tr>
<tr>
<td>21</td>
<td>219/297</td>
<td>.54</td>
<td>.18</td>
<td>.86</td>
<td>-.14</td>
<td>.94</td>
<td>-.3</td>
<td>I.69</td>
<td>85.6</td>
<td>82.1</td>
</tr>
<tr>
<td>22</td>
<td>240/297</td>
<td>-.19</td>
<td>.19</td>
<td>.93</td>
<td>-.6</td>
<td>.70</td>
<td>-.13</td>
<td>k.63</td>
<td>85.6</td>
<td>84.2</td>
</tr>
<tr>
<td>14</td>
<td>236/297</td>
<td>-.04</td>
<td>.19</td>
<td>.91</td>
<td>-.8</td>
<td>.72</td>
<td>-.13</td>
<td>j.65</td>
<td>84.3</td>
<td>83.8</td>
</tr>
<tr>
<td>4</td>
<td>233/297</td>
<td>.07</td>
<td>.19</td>
<td>.80</td>
<td>-.2</td>
<td>.65</td>
<td>-.18</td>
<td>i.69</td>
<td>86.0</td>
<td>83.5</td>
</tr>
<tr>
<td>7</td>
<td>251/297</td>
<td>-.62</td>
<td>.20</td>
<td>.76</td>
<td>-.23</td>
<td>.52</td>
<td>-.18</td>
<td>h.64</td>
<td>89.0</td>
<td>85.5</td>
</tr>
<tr>
<td>24</td>
<td>244/297</td>
<td>-.34</td>
<td>.20</td>
<td>.76</td>
<td>-.24</td>
<td>.71</td>
<td>-.11</td>
<td>g.66</td>
<td>88.6</td>
<td>84.6</td>
</tr>
<tr>
<td>3</td>
<td>248/297</td>
<td>-.50</td>
<td>.20</td>
<td>.76</td>
<td>-.24</td>
<td>.42</td>
<td>-.25</td>
<td>f.66</td>
<td>86.9</td>
<td>85.1</td>
</tr>
<tr>
<td>20</td>
<td>210/297</td>
<td>.82</td>
<td>.18</td>
<td>.75</td>
<td>-.27</td>
<td>.67</td>
<td>-.24</td>
<td>e.74</td>
<td>86.9</td>
<td>81.2</td>
</tr>
<tr>
<td>2</td>
<td>242/297</td>
<td>-.26</td>
<td>.20</td>
<td>.72</td>
<td>-.28</td>
<td>.47</td>
<td>-.25</td>
<td>d.69</td>
<td>89.8</td>
<td>84.4</td>
</tr>
<tr>
<td>5</td>
<td>255/297</td>
<td>-.79</td>
<td>.21</td>
<td>.72</td>
<td>-.28</td>
<td>.48</td>
<td>-.18</td>
<td>c.63</td>
<td>93.2</td>
<td>86.0</td>
</tr>
<tr>
<td>16</td>
<td>248/297</td>
<td>-.50</td>
<td>.20</td>
<td>.68</td>
<td>-.33</td>
<td>.67</td>
<td>-.12</td>
<td>b.66</td>
<td>91.9</td>
<td>85.1</td>
</tr>
<tr>
<td>9</td>
<td>272/297</td>
<td>1.66</td>
<td>.25</td>
<td>.66</td>
<td>-.27</td>
<td>.24</td>
<td>-.20</td>
<td>a.55</td>
<td>91.5</td>
<td>90.3</td>
</tr>
</tbody>
</table>

Recall that Rasch assumes one single continuum of item difficulty and person ability. For every student, an item is either more difficult or easier than another, unless equal. However, this is not always the case in reality. An item, say, a knowledge that an able student does not know may be known to a less able student. This does not entail that the
item is bad. It just misfits to the unitary model, suggesting that the knowledge in question is not of the sort that lies on the same scale of knowledge measured by other items. In fact, knowledge on a clear implicational scale will best fit Rasch, and one's knowledge of non-native language often eludes being implicational. The point is, whether fit or misfit, Rasch provides information about the item that cannot be obtained otherwise.

In Winsteps, two values are used for assessing model fit – mean squares (MNSQ) and its standardized value (ZSTD). Conventionally, MNSQ values between .7 and 1.3 are considered productive for measurement (e.g., Bond & Fox, 2007). Some of infit and outfit values deviate from the expected range. However, their effects on the general usefulness of the test seem to be nonsignificant. Insofar as ZSTD is concerned, items exceeding the absolute value |2.0| are considered to be misfitting. Point-measure correlation is the correlation between success on an item and the Rasch estimate. A low coefficient indicates that success on the item is only weakly correlated with an increasing ability estimate.

V. DISCUSSION OF MISFITTING ITEMS

According to Table 4, six items turned out to be misfitting in terms of both criteria. While Items 25, 6, 12, 18 are underfitting, 16 and 9 are overfitting to the contrary. What follows is a brief discussion about these six items.

Item 25, which is paradigmatic of grammar-translation, was not only the most difficult but also the most deviant item on the test. On the face of it, it tests students' knowledge about English word order or a formulaic construction "It is time...". It is crucial that the knowledge is supposed to be acquired implicitly with listening and speaking practice not by explicit instructions about grammar. It is unlikely, or otherwise would be inconsistent with the current paradigm of English teaching, that the test developer intended to test 7th grade students' grammatical knowledge such as expletive it, to-infinitive, parts of speech, etc. The problem is not the presence or importance of the targeted knowledge, but the way of testing it, which compels students to decompose the knowledge and notice the decomposability by a direct comparison with its Korean translation.

Furthermore, the way of decomposition is arbitrary for a testing purpose and even misleading. It is not clear why it's was contracted, therefore compromised phon-orthographical knowledge. Insofar as morphosyntax is concerned, it's has no better grounds for being phrased than to go or to go back. The grammatical category of a word is its use, not its lexicographic form. The Korean prompt takes for granted and contributes to the conception that L2 expressions are learned via L1 translation and thus L1 activates L2 encoding.
Item 25

Instruction: Choose a correct word for the slanted box so that you can express the same meaning as the underlined Korean translation.

\[ \text{집으로 돌아가기 시간이다.} \]

\[ \rightarrow \text{It’s} \]

(1) go (2) to (3) back (4) time (5) home

Item 6 was one of the most unpredicted by the model. It showed the biggest difference between observed and expected counts, which implies that students who were expected to succeed actually did not.

Item 6: Who am I?

- I can’t fly.
- I have four legs.
- I have a long nose.

Two issues underlie the fitness of this item to a unidimensional model. First, it assesses students’ acquaintance with negation (i.e., can’t) and content words (i.e., four legs, long nose, etc) at the same time. An equivalent item with one focused construct might have

---

5 The instruction was given in Korean. The author translated it into English.
reduced the variation. Second, while one of the three clue statements is insufficient, two are self-sufficient, which obscures what is being tested. If one knows the meaning of either the second or the third clue, then he/she will be able to choose the correct answer. The incoherent properties of the clues might mislead students to doubt that there is no reason for three when one is enough.

Similarly, Item 12 generated a large gap between actual and expected responses. Students were asked to find an expression appropriate for the given situation. The results indicate that a number of students did not know the meaning of “So long!” which is rarely used in comparison with more common expressions such as “Bye!”, “Good Bye!” , “Take Care!”, “See you!” , etc. If the target construct were students’ ability to use language for a certain purpose rather than their awareness of a specific expression, a more frequent expression should have replaced it. Although it is necessary to examine students’ response patterns before making any reasoned inference, it seems plausible that students might have chosen one among the four wrong choices even if they knew when to use them, yet were not sure of the function of “So long!”. If L2 translations were a medium of learning, they might bring about false generalizations too.

Item 12: Choose an appropriate expression for the blank.

(1) Hello. (2) Pardon? (3) So long! (4) Welcome. (5) Nice to meet you.

Item 18 inquires students’ understanding of English comparative construction. Its standardized infit and outfit values are 2.8 and 1.8, respectively. This indicates that the misfit occurs among students in a similar range of ability, not due to erratic outliers. In other words, students who were expected to answer the question correctly did not react consistently, and the degree of expectancy was heightened by a considerable number of correct answers from students who were estimated to be less able. One reason is attributable to the limitations of the test format that cannot measure this type of question in a desired way. It is likely that some students took advantage of the apparent hint in the list
of choices that is irrelevant to the tested construct. That is, students’ test strategies, irrespective of the target knowledge, influence the chance of their being correct on the item.

Item 18: Choose the correct pair of names for (A) and (B).

![Diagram of names]

(A) is faster than (B)

(1) Jiho Minsu
(2) Jiho Joon
(3) Seho Minsu
(4) Joon Seho
(5) Minsu Seho

On the other hand, Items 16 and 9 turned out to be overfitting to the expected model. This is the opposite case in that observed correct responses are by some reason considerably more than expected counts, or variation across students is too small to provide useful information about students’ ability.

Item 16 examines students’ vocabulary knowledge in discrete form. The fit statistics in Table 4 show that a considerable number of students who were not expected to solve this question found the correct answer in practice. To rephrase, the item difficulty was estimated higher than its actual value by some potentially able students’ wrong responses to a non-chance degree, resulting in a high infit estimate ($ZSTD=-3.3$). It can be further conjectured that such inconsistency was due, at least to some extent, to not in the instruction which is notorious for causing mistakes from young learners.

Another source of the overfit is imputed to its correlation with other items on the test. As seen in the above, Item 16 stacks with Items 3 and 8 at the same difficulty level. Rasch, and almost all parametric statistical procedures alike, assumes the independence of each observation wherein the mean and variance of a set of observations can be licensed as a
representative center and distributional property of the set.

Item 16: Choose the occupation that is not (underline original) mentioned in the passage.

This is my family. My father is a teacher. My mother is a doctor. My brother is a police officer. My sister is a cook.

(1)  (2)  (3)  (4)  (5)

If observations in the set systematically covariate with each other, the mean and variance may poorly represent the sample, leading to misinterpretations and flawed inferences about the population.

Item 9 was the easiest item on the test. 272 out of total 297 students answered this question correctly.

Item 9: What is he doing?

(1) He is swimming.
(2) He is watching TV.
(3) He is making a card.
(4) He is reading a book.
(5) He is playing basketball.
There is no significant difference between observed and expected counts (91.5% vs. 90.3%). Therefore, the overfit of the item appears largely due to its lack of functionality: it failed to induce enough variation from students’ responses. Apart from this, two factors play a part in its overfit status. First, the high outfit value (ZSTD=-2.0) suggests that there exist some outliers who made a careless mistake or solved it with little knowledge pertaining to what is being tested by other items. In connection with this, it is notable that there exists a substantial gap between this item and the other items. Its logit scale was -1.66, the only item lower than -1.0. The next lowest was -0.93. In other words, Item 9 is quite distinct from the rest of items, departing far from the continuum of difficulty. Since the whole model is calibrated in consideration of one’s relation to the others, the item ended up with a higher estimate than it actually was.

VI. CONCLUSION

A comprehensive study of language testing takes steps of diagnosis, revision, and revalidation. This study, which is short of the latter two, exemplified an immediate use of Rasch for analysis of a normal English test in school. Through a Rasch analysis, it was possible to draw a more concrete picture of how the difficulty of each test item matched students’ latent abilities. This merit will aid teachers to judge how a test can be better reformed on theoretical and empirical grounds. Moreover, its separation index helps identify the number of distinct ability and difficulty levels that can be measured by means of the test.

Rasch also provided a wealth of information on the functioning of individual items. When misfitting items were detected, it was possible to pinpoint where in the data they came from, even isolating responses that had not been expected. Variance components analysis of residuals made it possible to assess the degree to which the test results deviated from unidimensionality.

Rasch makes a probabilistic rather than deterministic interpretation of students’ abilities. Since Rasch is based on a binary measure (i.e., either right or wrong), it disregards the effects of choices on students’ responses to multiple-choice items. It does not tell us in what patterns students respond to incorrect choices that may vary by their degree of attractiveness. The choices themselves may contain additional information about the students’ abilities. If an item is designed for such use, classical analysis will follow for a close investigation into students’ responses to each choice. As seen in the foregoing discussion, insofar as individual choice is concerned, Rasch only draws rough and first-order inferences.
REFERENCES


Applicable levels: secondary education

Key words: Rasch analysis, placement test, reliability, item fit, unidimensionality, item difficulty

Munhong Choe
Department of Second Language Studies
University of Hawaii at Manoa
1890 East-West Rd.
Honolulu, HI 96822
U.S.A.
Phone: 808) 398-8283
Email: munhong@hawaii.edu

Received in March, 2010
Reviewed in April, 2010
Revised version received in May, 2010