The Many Uses of Rasch One-Parameter Analysis in Language Education

Matthew T. Apple

The Rasch model (Rasch, 1960) is a probabilistic statistical modeling analysis approach that transforms raw scores from test or questionnaire items into a logarithmic linear scale. This paper outlines some advantages of Rasch analysis over traditional true-score or classical test theory based analyses such as percentages and mean scores. Three examples of Rasch analysis are described: an in-class multiple-choice test, a hypothetical English placement exam, and a Likert-scale questionnaire for foreign language speaking anxiety. Extensions of the Rasch model, such as two- and three-parameter models for measuring oral and written performance rater and task bias, are also briefly explained.

1. Introduction

The Rasch model, also known as the one-parameter logistics model (1PL), measures the relationship between item difficulty and person ability as the ratio of success/failure for passing an item and expresses the difference in logarithm form (Embretson & Reise, 2000; Rasch, 1960). Simply put, the Rasch model can determine which items in a test or measurement instrument are more difficult than others and which persons in the sample population are more capable than others on the construct being measured. The purpose of this paper is to briefly describe advantages of the Rasch model over traditional methods of test analysis, as well as to provide a few examples of Rasch analysis.

1.1 Why use Rasch measurement analysis?

In language testing and evaluation, item discrimination analysis based on true-score theory, or classical test theory (CTT), has been the primary method for examining test item results. CTT-based item discrimination analysis is frequently used to evaluate the effectiveness of in-class, placement, and standardized test items (Brown, 1996). However, Rasch can determine what CTT analysis cannot, namely measurement error, item difficulty, and person ability, by placing items and persons on the same linear scale.

Every measurement has error, particularly in the social sciences where the words "absolute" and "one hundred percent" can never be used. This is particularly true in social science fields such as language education. Language educators assess attributes and abilities that necessarily derive from and are influenced by human social contexts, frequently resulting in items that can be subject to interpretation and that may differ between sample populations. Creating a measurement instrument, therefore, is not as easy a task as it may seem.

Among objective measurement specialists, a well-known comparison between measurements in social sciences and physical sciences is the example of the meter stick, popularized by Benjamin Wright of the University of Chicago (Bond & Fox, 2007, p. 5). When measuring a person's height, inches or centimeters are used, because no matter whom is being measured or where in the world he or she is being measured, the units of measurement will always measure the same distance. Height is, in this sense, an absolute measurement system on an interval scale. In other words, the difference between 52 and 53 centimeters is exactly the same as the difference between 170 and 171 centimeters regardless of the sample population or the person doing the measuring. Another analogy is measuring temperature using Fahrenheit or Celsius; in each scale the incremental increases between degrees are the same, and although the scales are different, they are calibrated to each other so that we can easily convert between them (e.g. 0℃, 20℃, and 100℃ equal 32°F, 70°F, and 212°F, respectively).

In the fields of social science, however, scales and assessments are typically not analyzed for error or interval accuracy, and rarely are they calibrated with other assessment scales. For instance, language instructors typically grade items on a language vocabulary test by adding them together to produce raw scores or averages. However, there is no way to know for certain that the difference between items is the same regardless of where in the test the items occur. There is no guarantee that the difference between Item 3 and 4 is the same as the difference between Item 50 and 51. By simply tallying up all the points and assuming that each item is the same difficulty as any other, instructors may be giving two students a 75% score even though the students correctly answered items of differing difficulty levels.

The problem is that when constructing exams, language instructors essentially make a guess that certain items are more difficult than others, based on either previous experience with students or intuition about current students' abilities. In both assessing and constructing exams, subjectivity plays too great a role in what is supposed to be an objective measurement of students' abilities. Simply using raw scores based on unique items on constructs that have not been calibrated will not provide an accurate measurement of student ability.

By using a probabilistic model of student responses, Rasch analysis can aid in both construction and assessment of language abilities. Based on Rasch analysis results, educators can mathematically determine which items are more difficult and which are easier to answer. By doing so, we can then determine which items discriminate between high achievers and low achievers, since a high achiever is more likely to answer difficult items correctly than a low achiever.

Given enough data, Rasch analysis can even help educators spot the "guessers," or students who accidentally guess correctly on difficult items that the model predicts are not likely for low achievers to answer. Further, by creating a model to measure constructs (for example, "English vocabulary ability"), educators can produce tests that can elicit the same results across groups of students from year to year for the purposes of comparison. In fact, it is in the field of testing that Rasch analysis and other item response theory (IRT) models have been used most frequently during the past two decades (McNamara, 1996; Wilson, 2005). The Rasch model is the simplest of the IRT models and the most accessible to language educators. What the model is and how it works will be explained in the following two sections.

1.2 The Rasch measurement model

Named after the Danish mathematician Georg Rasch (Rasch, 1960), the Rasch model states that the probability a person will get an item correct is logistically related to the difference between the person's ability level and the item difficulty. The logit form of the model is

$$\ln [P_{\mathrm{ni}} / 1 - P_{\mathrm{ni}}] = \theta_{\mathrm{n}} - b_{\mathrm{i}}$$

where

P_{n} = person n

i = the item being answered

θ_{n} = the ability level of person n

b_{i} = the ability level of the item

Data is fit to the Rasch model by mathematically transforming raw scores on items into logarithms and then by placing both item responses and persons on the same log-odds scale. In other words, the unrelated percentages of correctly and incorrectly guessed items become transformed into a linear scale similar to the example of the meter stick or thermometer in the physical sciences (Figure 1). The scale extends from -3 to +3, following the normal distribution curve; that is, in a normally distributed sample population, 99.74% will fall

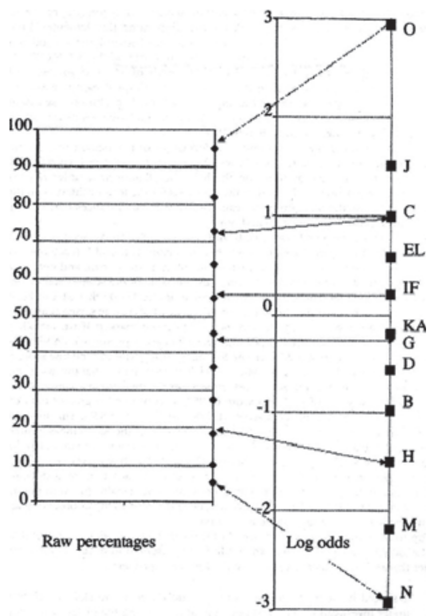

Figure 1. The transformation of raw percentages on items to a "log-odds," or logarithmic linear scale in the Rasch model (Bond & Fox, 2007, p. 25).

within three standard deviations from the mean on a given construct (Elifson, Runyon, & Haber, 1998, p.136).

1.3 Assumptions of the model

The Rasch model does have a few assumptions, however. The first assumption is a normal distribution curve. As normal distribution generally requires large sample populations (i.e., more than 100, preferably 250 to 300), and as small samples tend to be highly skewed, it is unlikely that the average classroom teacher will find a normal distribution within a single class. The assumption of normality could be met, however, by combining results from several similar classrooms. The second assumption the Rasch model makes is that all item responses are unrelated to each other and are ranked in terms of difficulty. Matching item tests, for example, violate this assumption, as one response automatically reduces the possible responses to other items.

The third and most crucial assumption is unidimensionality of construct validity. In other words, the Rasch model assumes that there is only one construct underlying the items we are using to measure person ability. For example, all tems on a vocabulary test should measure vocabulary ability, and not vocabulary plus grammar ability. A test (or set of items) that attempted to measure both constructs at the same time would lead to an inability of the test to accurately measure either vocabulary or grammar. If a student answered an item correctly, the instructor would not know whether the correct answer indicated the presence of grammar knowledge or of vocabulary knowledge.

A final assumption of the Rasch model is that all items fit the underlying construct. In other words, all item responses should directly measure the ability the instructor is attempting to measure within a certain margin of error. The Rasch model uses an analysis technique called "item fit" to determine whether items measure what they are intended to measure. At the same time, the Rasch model also determine the relative difficulty of each item based on how persons (students) in the sample population respond to each item. This in turn allows for "person fit," which can measure how closely each individual student's responses match the model.

The capability of the Rasch model to measure both item difficulty and person ability on the same linear scale is a major advantage of Rasch item analysis over CTT-based item analysis. The following section will briefly describe the statistics used in Rasch item analysis and provide examples of this analysis.

2. Item and Person Fit

Rasch analysis measures both item responses and individual student responses (termed "person responses" in the model) by placing both on the same logarithmical linear scale. Rasch is a probabilistic model: it assumes that a student who correctly guesses a difficult item is more likely to guess an equally or more difficult item, whereas a student who incorrectly guesses a difficult item is unlikely to guess an equally or more difficult item.

Rasch analysis produces several measurement indices, but the most useful for classroom teachers are item/person fit and the item/person map. Item/person fit statistics can show the result of carelessness, response set answering, or item bias (Wolfe & Smith, 2007, p. 211), while item/person maps display both items and persons on the same logit scale (Wilson, 2005, p. 96).

2.1 Item/person fit

Table 1 displays item/person fit measures from an in-class vocabulary quiz given to all third year students ($N = 203$) at Nara National College of Technology in the spring 2008 semester. For both types of items (labeled "Vocab*N*" and "Prag*N*"), students were asked to circle one of three possible answers. "Vocab"-type items presented one English word with three possible Japanese equivalents (i.e., translation items). "Prag"-type items presented a sample English sentence with a missing word and three possible English answers (i.e., fill in the gap items). Both types of items were assumed to measure the same theoretical construct of "English vocabulary ability."

After checking the assumption of normality, the data were analyzed using WINSTEPS software (Linacre, 2006). The results include the item difficulty level ("measure"), the standard error of each item ("S.E."), and the Infit and Outfit standardized *z*-scores ("ZSTD") as listed in Table 1. The item measures indicate the item difficulty as measured in logits, with zero as the absolute mean difficulty level. Positive numbers indicate difficult items and negative numbers indicate easy items. For the sample population in this analysis, the translation-only items ("Vocab") generally fall towards the "easy" polar end of the scale, while fill in the gap sentence items ("Prag") are generally more difficult.

In addition, the easier items generally have more error associated with them; Vocab1, the easiest item with a measure of -3.81, also has the highest error at 1.01. Since there were only there possible responses to each item, this error seems quite high. Also, the extremely low measure score falls outside the normal ±3 range of normal distribution. Both these statistics seem to indicate that the item is not capable of adequately measuring the vocabulary abilities of participants in this sample. This information cannot be obtained from only raw score data.

Table 1.
Sample Item/person Fit Measures from an In-class Vocabulary Test ($N = 203$)

Item	*Measure*	*S.E.*	*Infit ZSTD*	*Outfit ZSTD*
Vocab5	2.36	.16	-0.4	0.4
Prag3	2.36	.16	1.0	0.9
Prag1	1.78	.16	**2.2**	1.7
Prag6	1.71	.16	0.2	0.2
Prag2	1.13	.17	0.6	0.3
Prag4	0.93	.17	-0.2	-0.8
Vocab3	0.84	.17	**3.0**	1.8
Vocab7	0.68	.18	-0.9	-0.9
Vocab4	0.65	.18	-0.4	-0.9
Prag5	0.62	.18	-1.3	-1.3
Vocab9	-0.25	.22	-0.6	-1.1
Vocab8	-0.65	.25	-0.1	-0.9
Vocab11	-0.65	.25	-0.5	-0.9
Vocab10	-1.51	.35	0.0	1.2
Vocab12	-1.64	.37	0.2	0.1
Vocab2	-2.15	.46	-0.1	-1.2
Vocab6	-2.38	.51	0.3	0.8
Vocab1	-3.81	1.01	0.4	0.4

The infit statistics as indicated by "ZSTD" can also tell educators which items are working to measure person ability. Most IRT statisticians agree that a fit statistic of ZSTD < ±2.0 indicates good fit to the model and further controls for Type I Error (Bond & Fox, 2007; Smith, 1996; Wolfe & Smith, 2007). Items that "under fit" at less than -2.0 indicate patterns that are erratic, typically caused by low ability persons guessing on difficult items. Items that "over fit" at greater than 2.0 indicate patterns close to the hypothetical Guttman-style "perfect" response pattern of only high-ability students correctly answering all highly difficult items (Guttman, 1944). In other words, an overfitting item is "too good to be true," and may be an indication of items that are not independent from other items. For example, distractors to one item may give students clues about the answers to a separate item.

In Table 1, two items overfit the model (highlighted in boldface). Before attempting to use items from this test again with a future sample population, both items should be re-examined for wording and independent relation to other items in the test.

2.2　The item/person map

The items and persons in Rasch analysis are placed on to a linear scale. Both items and person responses are expressed in logits along the same linear scale construct map, called a "Wright map" or item/person map (Wilson, 2005, p. 90). Persons are usually located on the left side and are represented by symbols, while items are located on the right side. In Figure 2, the items from Table 1 are displayed. At the right, the most difficult items fall near the top, and the least difficult items fall near the bottom. On the right, higher ability students fall near the top and lower ability students lower down the scale.

Because this sample has 203 students, a distribution similar to the theoretical normal distribution curve appears on the left. However, because the items in the analysis come from an in-class test (hence, a criterion-referenced test rather than a norm-referenced test), there is a strong floor effect for the items. In other words, although the fit statistics indicated that this is a good test of classroom-based learning (i.e., it shows that students have learned the class materials), this quiz cannot adequately measure student overall vocabulary abilities because the majority of the items are below the students' actual vocabulary ability.

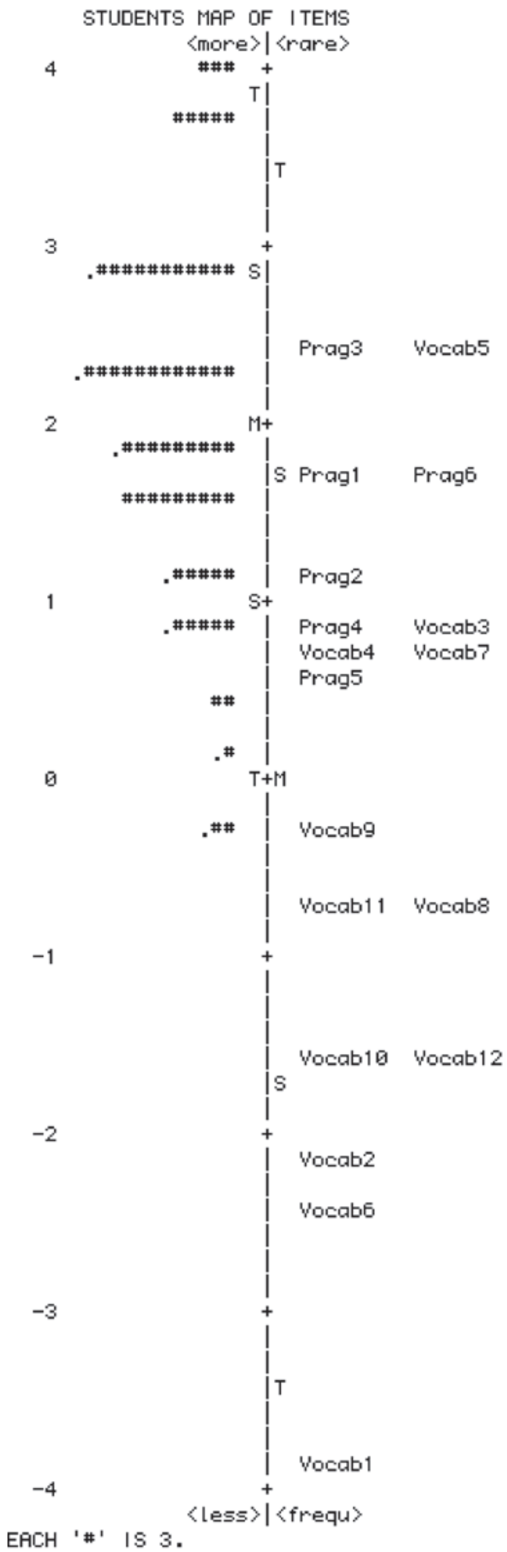

Figure 2. Item/person map for an in-class vocabulary test. (N = 203) Students appear on the left side and items appear on the right side. More capable students and more difficult items are near the top of the scale, while less capable students and less difficult items are at the bottom.

3. Other Uses of Rasch Analysis

3.1 Entrance and placement exams

Norm-referenced exams such as standardized entrance or placement exams would ideally produce item/person maps with a distribution curve approaching normality without floor or ceiling effects. An "ideal" high-stakes exam such as an entrance exam would have multiple items of the same difficulty level around the cut-off point for admissions. Rasch analysis of entrance exams can help determine where cut-off points are and which items to reduce and add in order to increase reliability and validity. Analysis using similar items from different entrance exams can also help ensure that tests retain the same comparable from year to year.

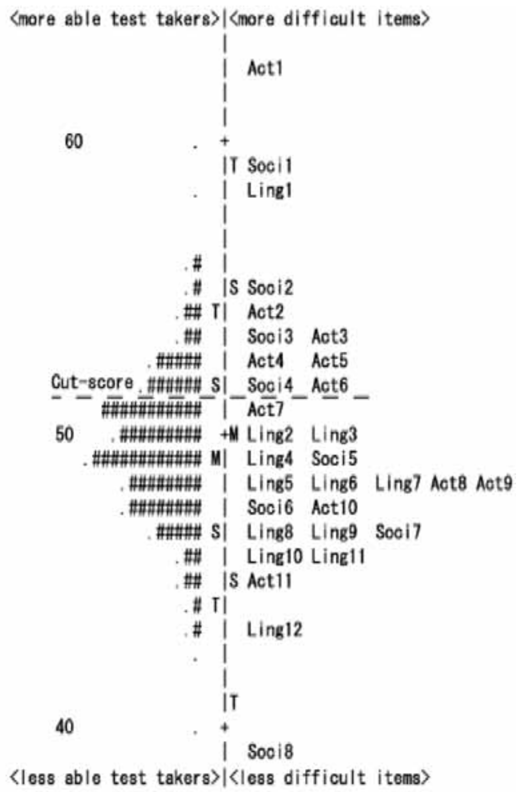

Figure 3. An item/person map from a hypothetical entrance exam, with cut lines for placement (Weaver, 2005).

Placement exams are not dissimilar to entrance exams, however, and some argue that results from entrance exams could also be used to place students into courses based on varying proficiency levels (Weaver, 2005). Figure 3 shows results from a "hypothetical entrance exam," which includes a theoretical cut-off point for placement purposes after entering the institution. Programs with more than two levels, of course, could have two or three separate cut-off points, such as the hypothetical model in Figure 4. At each cut-off point, multiple items of the same approximate difficulty as determined by item/person measure fit statistics would help assist in the placement process (Sick, 2008).

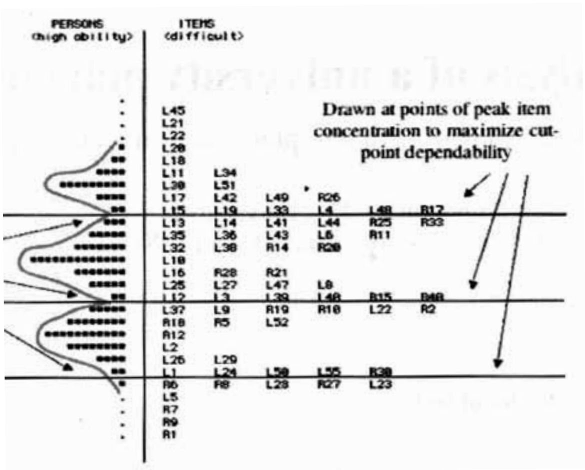

Figure 4. An item/person map for a hypothetical placement exam, showing ideal cut-off points for three ability levels of entering students (based on Sick, 2008).

3.2 Likert-scale questionnaires

Rasch analysis can additional inform questionnaire design for teachers interested in non-linguistic aspects of the language classroom, such as anxiety, motivation, and learner beliefs. In the case of Likert-scale items, the items can be treated as ranging in difficulty from "easy to endorse" to "difficult to endorse" based on the probabilistic model created from participant responses.

However, whereas CTT-based questionnaires rely on multiple items of roughly the same difficulty level that will correlate highly, thus producing a high Cronbach's alpha estimate, Rasch measurement statistics require that items represent varying difficulty levels in order to properly rank-order participants. A typical CTT-based questionnaire such as the Attitude/Motivation Test Battery (AMTB) contain an equal number of positively- and negatively-worded items in order to raise the internal reliability alpha estimate; however, objective measurement studies have consistently demonstrated that negatively keyed items not only do not measure the same construct as positively-keyed items (e.g., Chang & Wright, 2001; Smith, 1996; Wright, 1996). Also, it has been known for some time that high Cronbach alpha reliability estimates have no relation to construct validity or dimensionality (Embretson & Reise, 2000; Green, Lissitz, & Mulait, 1977).

Rasch analysis of Likert-scale data is important because it treats the data as if the items were a test of person ability. In other words, the Rasch model assumes that some items are easier to answer than others. This kind of questionnaire item difficulty is termed "item endorsability," and can be measured by using Rasch analysis. Further, Rasch analysis of person measures can determine the existence of overfitting responses, which may indicate "response set" answering patterns (i.e., students filling in blanks in a deliberate pattern to avoid answering the questionnaire) or "artifact responses" (i.e., students answering what they think the teacher wishes to hear).

Figure 5 shows an item/person map from a six-point Likert-scale questionnaire study of 172 Japanese university first year students' foreign language speaking self-competence (Apple, 2009). In the figure, students and items at the top represent greater speaking self-competence, and persons and items at the bottom represent less speaking self-competence. The most difficult item for students to endorse was Item 3, "I can make a hotel reservation over the telephone" ($M = 1.67$), while the easiest item to endorse was Item 4, "I can start a group discussion in class if I prepare ahead of time" ($M = 4.05$).

In the case of Item 3, the content of the item clearly falls well out of the realm of experience for all but three or four participants; although many EFL conversation textbooks include hotel reservation lessons, students may have a difficult time even imagining themselves doing such an activity, let alone being able to do it with a certain degree of competence. Of course, we can see from the means score that the item is difficult; however, the Rasch analysis gives us even more telling information.

The item/person map shows a sizeable gap between Item 3 and the next most difficult items (Item 7, "I can easily join in a conversation in a group of native speakers," and Item 14, "I can give street directions to a foreigner."). This indicates that not only is the item too difficult for most participants, but it also is not helpful for determining which participants have self-competence and which do not. Similarly, there is a gap between Item 4 and the next easiest Item 2 ("I can use basic greetings during pair work with a classmate."). More items should be written to fill in the gap and thus more effectively discriminate students at

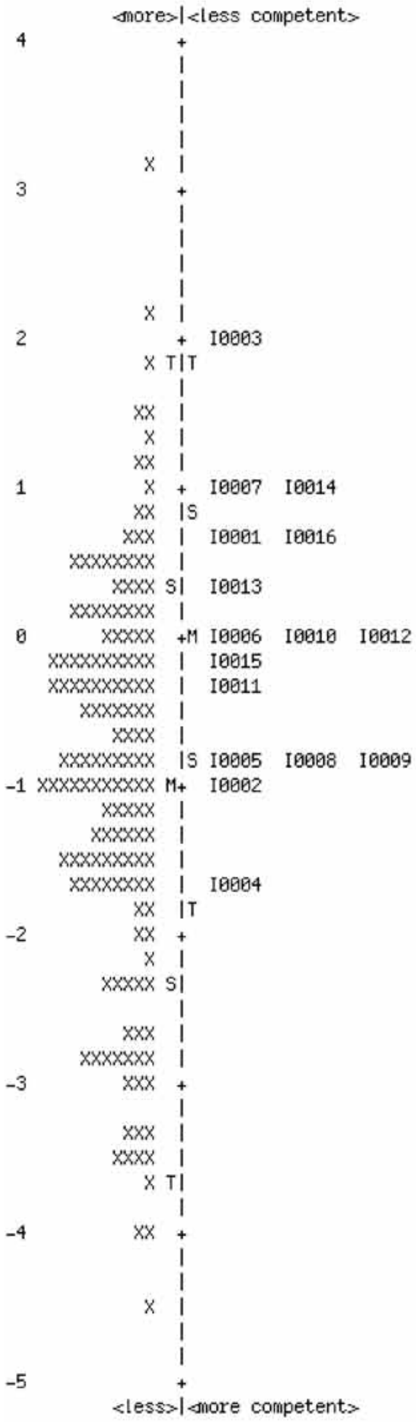

Figure 5. Item/person map of 172 Japanese university participant responses to foreign language classroom speaking competence items (Apple, 2009).

this level of self-competence.

The "cut-line" for more or less English speaking self-competence falls at the "M" mark for items on the right side. As in the ideal placement exam example above (Figure 4), there are several items gathered around this point. In this study, Items 6, 10, 11, 12, and 15 seemed most effective for determining self-competence. Of the five items, three of them concerned standing in front of the class, indicating the kind of activities that would allow teachers to discern whether students felt themselves competent at speaking English.

On the left side of the item/person map is the person ability levels. Since the mean person ability is roughly one standard deviation (S) lower than the item difficulty mean (M), it appears that participants in this sample found most of the items too difficult to endorse. If the data had been examined using only CTT-based analysis such as mean and SD scores, the need for further analysis would not have come to light. However, after Rasch analysis, the results indicate that to verify the speaking self-competence construct, a larger sample with participants of greater ability levels and slightly revised items are necessary.

4. Extensions of the Rasch model

In addition to analyzing dichotomous (true-false) and polytemous (multiple-choice) data, the Rasch model can be extended into the "multiple rater" model (also known as the two parameter logistics model or 2PL) and the "multifaceted model" (also known as the three parameter logistics model or 3PL). These more complicated models take into account rater severity or leniency and task bias. Raters and knowledge of varying tasks can impact on student ability for performance-based assessments such as oral speeches and essay writing. For example, one rater could consistently give lower grades on student essays than another rater, or one rater could consistently give lower grades to essays at the start of the rating process, but give higher grades to essays near the end. Adjusting grades according to rater bias is a key to changing what is essentially a subjective grading process into an objective grading process that is consistent between raters and between groups of students over time.

Since the newly revised TOEFL includes both speaking and writing components, analysis of rater bias has become a crucial element of language testing (Schaefer, 2008). Traditional interrater reliability coefficients have sought to account for individual rater differences through simple correlations or averaging rater grades. However, interrater reliability does not account for inconsistency within a single rater's grades, nor does it account for possible changes in grades from one rater over time. One way to adjust for rater bias is to use the many-faceted

Rasch model, which looks like this

$$\ln\left[P_{nijk} / 1 - P_{nijk}\right] = \theta_n - b_i - \alpha_j - \tau_k$$

where

P_n = person n

i = the item being answered

θ_n = the ability level of person n

b_i = the ability level of the item

α_j = the severity of the rater

τ_k = the difficulty of receiving a rating k relative to a rating of k-1

In the many-faceted Rasch (MFR) model, rater tendencies can be adjusted to calibrate person ability across rater and task categories. Figure 6 presents results from Grade 3 primary school students who were assessed by the Michigan Educational Assessment Program (Engelhard, 2008). The data were analyzed using the MFR model according to items, raters, rounds, and four levels of performance standards. The item map shows not only item difficulty level, but it also displays rater severity as well as round severity and overall evaluation according to the grades "A" (apprentice), "B" (basic), "M" (met), and "E" (exceeded). Some of the 21 raters typically gave more severe or more lenient assessments than others. Assessments given in the first round tended to be more severe than in the second and third round; raters tended to be slightly more demanding when giving a grade of "M," but they also tended to give "B" evaluations more easily.

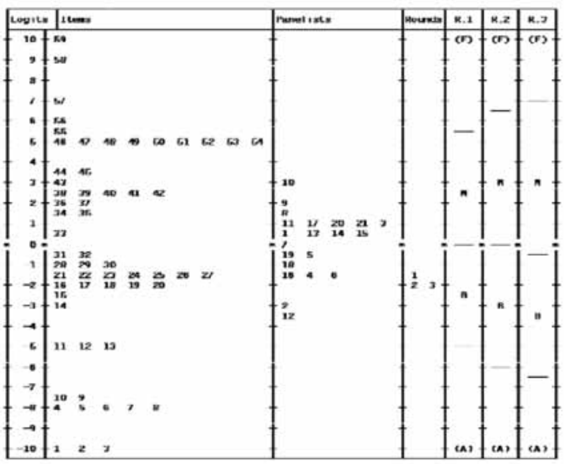

Figure 6. Item/person map for the many-faceted Rasch model (Engelhard, 2008, p. 1156)

For foreign language education specialists, the MFR model could be used to analyze ratings for oral or written performance. For example, in order to assess performance on an English speech or presentation, raters give points on several categories (items) on a predetermined rubric (i.e., English pronunciation, eye contact, gestures, etc.). The resulting rater grade sheets would then be input into the MFR model and analyzed to adjust for rater severity before giving an overall grade. Students may also perform better depending on the topic of the assignment and depending on how many times the task is performed. MFR also accounts for task difficulty and task repetition, giving a clearer picture of "true" performance, rather than relying solely on subjective rater scores.

5. Summary and conclusion

Compared to traditional statistical analysis, Rasch analysis has much to offer language educators. In addition to helping educators improve in-class exams for the purposes of classroom evaluation, Rasch analysis can also contribute to the entrance exam decision-making process as well as inform decisions for English proficiency placement. Moreover, Rasch analysis of questionnaire data can also assist in revision and interpretation of non-linguistic, language-related psychological constructs such as language anxiety, motivation, and perceived self-competence.

Rasch analysis was first used extensively in Australia to assess English reading and listening comprehension on citizenship examinations (McNamara, 1991; McNamara, 1996). However, demand for Rasch analysis has increased rapidly in the United States since the onset of the "No Child Left Behind" policies of the Bush administration (Engelhard & Myford, 2003). Rasch analysis can thus play an important role in the improvement of pre-existing tests and assessments on a national level.

On other hand, Rasch analysis can also play a crucial role in informing individual institutions' curricular decisions about classroom and course objectives. If one of the goals of a language curriculum is to encourage the active oral or written production of the target language, assessing productive language use can create powerful washback effects that will encourage students and teachers alike to approach foreign language education from a more communicative perspective (Brown & Hudson, 1998). Rasch analysis of rater assessments of student oral and written English production can increase reliability and validity, turning what appears to be subjective opinions on the surface into objective, verifiable measurements of language performance.

References

Apple, M. (2009). Foreign language speaking anxiety and self-confidence: A comparison of true-score based factor analysis and Rasch analysis. Unpublished manuscript.

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

Brown, J. D. (1996). *Testing in language programs.* Upper Saddle River, N.J.: Prentice Hall Regents.

Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly*, 32(4), 653-675.

Chang, C.-H., & Wright, B. D. (2001). Detecting unexpected variables in the MMPI-2 Social Introversion Scale. *Journal of Applied Measurement*, 2(3), 227-240.

Elifson, K., Runyon, R. P., & Haber, A. (1998). *Fundamentals of social statistics* (3rd Ed.). Boston: McGraw-Hill.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Engelhard, G. (2008). Standard errors for performance standards based on bookmark judgments. *Rasch Measurement Transactions*, 22(1), 1156-1157.

Engelhard, G., & Myford, C. M. (2003). *Monitoring faculty consultant performance in the advanced placement English literature and composition program with a many-faceted Rasch model*. New York: College Entrance Examination Board [College Board Research Report No. 2003-1].

Green, S. B., Lissitz, R. W., & Mulait, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement*, 37, 827-838.

Guttman, L. A. (1944). A basis for scaling qualitative data. *American Sociological Review*, 9, 139-150.

Linacre, J. M. (2006). WINSTEPS Rasch measurement computer program [Computer software]. Chicago.

McNamara, T. F. (1991). Test dimensionality: IRT analysis of an ESP listening test. *Language Testing,* 8(2), 139-159.

McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago.

Schaefer, E. (2008). Rater bias patterns in EFL writing assessment. *Language Testing*, 25(4), 465-493.

Sick, J. (2008, May 11). A Rasch analysis of a university placement test: Using Rasch analysis to inform student placement in a language program. Presentation given at the 7th JALT PanSIG Conference, Doshisha University, Kyoto, Japan.

Smith, R. M. (1996). A comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modeling*, 3(1), 25-40.

Weaver, C. (2005). How entrance examination scores can inform more than just admission decisions. *Proceedings of the 4th Annual JALT Pan-SIG Conference.* Retrieved October 13, 2008, from http://www.jalt.org/pansig/2005/HTML/Weaver.htm

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum.

Wolfe, E. W., & Smith, J., E. V. (2007). Instrument development tools and activities for measure validation using Rasch models, Part II : validation activities. *Journal of Applied Measurement*, 8(2), 204-234.

Wright, B. (1996). Comparing Rasch measurement and factor analysis. *Structural Equation Modeling*, 3(1), 3-24.