Shiken: JALT Testing & Evaluation SIG Newsletter, *13* (1) January 2009 (p. 4 - 10)

Rasch Measurement in Language Education Part 3: The family of Rasch Models

James Sick, Ed.D. (J. F. Oberlin University, Tokyo)

In Parts 1 and 2 of this series, I presented an overview of Rasch measurement theory (RMT), discussed its relationship to classical true score theory and item response theory, and discussed the property of invariance, a property that Rasch theorists consider fundamental to all measurement. In this installment, I will elaborate further on the family of Rasch models, including their origin and their application. As usual, this series will be presented in question and answer format and readers are invited to send in questions for future installments.

Q: I often see references to "*the* Rasch Model," but in previous installments you have implied that there is more than one model. Could you elaborate on your statement that RMT refers to a *family* of statistical models?

A: Georg Rasch's original work (Rasch, 1960) was done with dichotomously scored tests. Specifically, a set of IQ and aptitude tests that he had been asked to equate by the Danish Department of Defense. In the Rasch dichotomous model, person ability and item difficulty are viewed as population parameters that can be estimated from the responses of an adequate sample of test items and test takers. Items are scored as either "correct" or "incorrect." The total number of items answered correctly is used to estimate each person's ability on the underlying construct. The total number of correct responses to each item, the *item* raw score, is used to estimate its difficulty. In addition, the magnitudes of the raw scores are used to estimate the distance *between* person and item rankings. This last point differentiates Rasch from classical test theory, where ascending raw scores are assumed to delineate *equal* distances in ability or difficulty.

It bears repeating that the Rasch ability and difficulty estimates are inferential, as opposed to descriptive. The person and item measures constructed by a Rasch analysis allow us to speculate about how likely a person of *B* ability is to succeed at an item of *D* difficulty. And by extension, infer how much of a given ability a candidate possesses by identifying the point at which items have become so difficult that the candidate is as likely to fail them as to succeed.

The Rasch-Andrich Rating Scale Model

Rasch's original conceptualization was extended considerably by David Andrich (1978) when he proposed that responses to Likert-style questionnaire items could be ordered and used in a similar way to infer how much of an attitude or psychological attribute a questionnaire respondent possessed. In the Rasch-Andrich rating scale model, item difficulty is re-conceptualized as the resistance to endorsing a rating scale response category.

To illustrate, imagine that we wish to measure an hypothesized construct that we have labeled "willingness to communicate in English" (WTC). We have assembled a set of Likert-style questionnaire items that we believe tap into this construct. Our first item is "How willing would you be to tell someone the time in English?" Possible responses are 1) almost certainly not willing, 2) probably not willing, 3) probably willing, and 4) almost certainly willing.

How much WTC would an individual need to honestly endorse Step 3, probably willing, for this item? Probably not that much, considering that the task does not require a great deal of skill or entail much risk of embarrassment. A more challenging item, however, such as "make a welcoming speech to a group of foreign students visiting our school" would require much more willingness to communicate in English before an individual could honestly endorse one of the higher scale steps. When we conceptualize the construct in terms of how easy or difficult it is to endorse a particular scale step of a particular item, we establish links amongst the questionnaire items. For example, we might discover that endorsing Step 2 for the harder item requires (or implies) about the same degree of WTC as endorsing Step 4 for the easier item.

"Rasch-Andrich thresholds can be thought of as 'local dichotomies' between adjacent Likert-scale steps."

Figure 1 is an abbreviated variable map, based on real data from a WTC questionnaire administered to Japanese high school students. In this pared down version, 18 people, labeled P01-P18, have responded to three items labeled I2, I26, and I43 from the WTC questionnaire. Note, however, that in contrast to variable maps for a dichotomous model, the items are listed with decimal postscripts. That is, item 2 appears on the map as I2.1, I2.2, and I2.3. The postscripts refer to the *thresholds* between the four steps of the response scale. A threshold is the point at which a respondent is as likely to choose the higher scale step as the lower. Rasch-Andrich thresholds can be thought of as "local dichotomies" between adjacent Likert-scale steps. Thus, Threshold 1 of Item 2 (I2.1 on the variable map) represents the point on our hypothesized construct at which a respondent has just enough WTC to hover indecisively between "almost certainly not willing" and "probably not willing" on the easiest item on the questionnaire, which is "tell someone the time in English." I2.2 represents the threshold between "probably not willing" and "probably willing," and I2.3 the point between "probably willing" and "almost certainly willing."

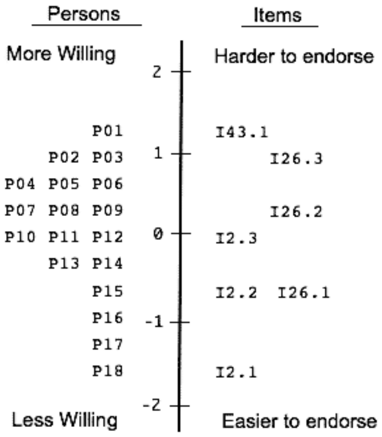

Figure 1. A variable map of willingness to communicate in English (abbreviated).

Item 26 was "help a native speaker read a menu in a restaurant." We see in Figure 1 that Threshold 1 of Item 26 has the same logit measure as Threshold 2 of Item 2, and person P15. We might imagine that person P15, at about 1.3 logits of WTC, is hovering between "2) probably not willing" and "3) probably willing" for Item 2, and "1) almost certainly not willing" and "2) probably not willing" for Item 26. Or in the language of inferential statistics, Person P15 has a fifty percent probability of endorsing either Step 2 or Step 3 for Item 2, and a fifty percent probability of endorsing Step 1 or Step 2 for Item 26. She is unlikely, however, to endorse Step 2 for Item 43, "make a welcoming speech to visiting foreign students." In this sample, only respondent P01 displayed enough overall WTC that she would, we infer, even consider endorsing Step 2 of our rating scale for this item.

Andrich's extension of the Rasch model to rating scales permits us to analyze both tests and questionnaires using a unified theory of measurement. The rating scale variable map helps to visualize the construct and how the questionnaire items define it, and can be a very useful tool for validating and increasing understanding of the construct. Moreover, complex attitudinal variables such as WTC are rarely unidimensional. Examination of Rasch fit statistics and residual structures provide tools that allow us to assess the degree to which secondary and tertiary dimensions are distorting or biasing the primary measurement dimension. An explanation of fit and residual analysis is beyond the scope of this installment, but for an excellent example of a Rasch analysis applied to a

WTC questionnaire, see Weaver (2005). For additional examples of questionnaire analyses using the Rasch-Andrich rating scale model, see Sick (2007).

The Rasch Partial Credit Model

Masters (1982) proposed a further generalization of the rating scale model, now known as the Rasch-Masters partial credit model. Masters pointed out that in the multiple-choice item format, some distractors may be closer to the correct answer than others. Rejecting distractors A and B and selecting distractor C, for example, might imply a greater degree of knowledge or ability than selecting A or B, even though the best choice is D. Similarly, in open item formats that require several steps for completion, such as algebra word problems, partial credit might be awarded for successful completion of some steps, even though the final answer is wrong due to a calculation error.

"the Rasch-Masters partial credit model permits each item to have it's own unique rating scale."

In the Rasch-Andrich rating scale model, all items have the same number of steps, and the modeled distance between adjacent steps is consistent across items. In contrast, the Rasch-Masters partial credit model permits each item to have its own unique rating scale. Items may vary in both the number of steps they have, as well the modeled distance between thresholds.

Figure 2 is another abbreviated variable map based on data from the Lunic Language Marathon (Sick and Irie, 2000), a multiple-choice language aptitude test. In this test, examinees had to induce grammar and syntax rules from samples of a pseudo-language and apply the rules to multiple choice translation items. For some items, distractors applied some but not all of the required rules and were considered worthy of partial credit. Differences between the rating scale and partial credit models are illustrated in Figure 2. Items 99 and 101 awarded 3 points for the best answer and two, one or zero points for the three distractors, on the rationale that choosing some distractors displayed greater control over the Lunic grammar rules than other distractors. This four-step configuration creates three Rasch partial credit thresholds. In Item 99 the logit measure of Threshold 2 is roughly halfway between Thresholds 1 and 3. In Item 101, however, the Threshold 2 measure is closer to Threshold 3, implying it is closer, in a sense, to the correct answer. Moreover, notice that some items, such as 98 and 103 have only two thresholds, while others such as 94, and 105 have none. This is because items 94 and 105 did not give partial credit for any distractors, and Items 98 and 103 allowed partial credit for only one of the three distractors. The partial credit model allows for greater flexibility in how items are modeled. This is especially useful for a multi-format test. A disadvantage is that it requires a larger sample size to estimate stable parameters because distances between thresholds are estimated separately for each item. In the rating scale model, a single set of rating scale thresholds is estimated using all of the item data and applied to all items.

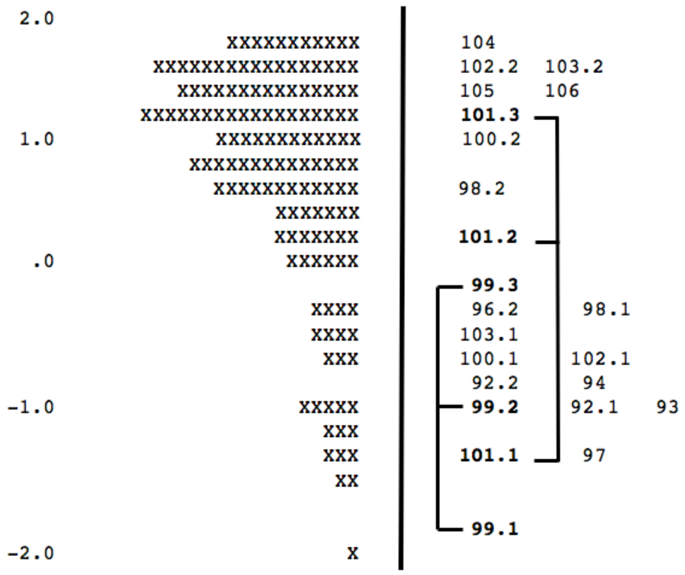

Figure 2. A variable map of a language aptitude test using the partial credit model.

The Many-Facets Rasch Model

The fourth and perhaps best known Rasch model in language testing is the many-facets Rasch model (Linacre, 1992). A many-facets Rasch model is usually used for performances that are awarded subjective ratings, such as essays or speaking assessments. It is designed to control for the effect of confounding "measurement facets" that influence scores, such as rater severity, or topic difficulty. A many-facets Rasch analysis does not, as is often erroneously believed, increase the reliability of an assessment. Improving reliability generally requires increasing the number of observations. A many-facets Rasch analysis increases the *accuracy* of the measurement of the latent variable by simultaneously creating measures of the confounding facets, and adjusting the person measures to compensate. If the assessment involves multiple raters jointly observing and rating performances, a measure of rater severity is constructed by observing how raters differ when observing the same performance. If the performance includes multiple topics

> *"A many-facets Rasch analysis increases the accuracy of the measurement of the latent variable by simultaneously creating measures of the confounding facets, and adjusting the person measures to compensate."*

or essay prompts, a measure of topic difficulty can be constructed by observing how examinee performances differ when they write or speak on different topics. The modeled effect of these confounding facets can then be added to or subtracted from the person measures in order to provide us with a more accurate measure of the abilities that produced the performances. Because the variable maps for many-facet Rasch models become very complex, they are usually displayed as separate maps, linked to a common logit scale, for each facet.

Conclusion

Although they differ in complexity and are designed to handle different measurement problems, the family of Rasch models are linked conceptually through the notion of person and task. Before attempting to construct a Rasch measure of a psychological attribute, one should perform a logical test: does it make sense to say that some people have more of this construct than others, and that some tasks *require* more of it than other tasks? If yes, we attempt to assemble a set of tasks of increasing difficulty that span the range of ability or attitude in our target population. The various Rasch models differ in the types of tasks they use. In the dichotomous model, the task is to produce the one acceptable response to a test item. In the rating scale model, the task is to honestly endorse a response category on a questionnaire. In the partial credit model, it is to choose the better of several alternatives or to complete sequential steps in a complex task. For the many-facets Rasch model, the task is essentially to please a rater. Or to state it with the complexity that a many-facets Rasch analysis entails, the task is to produce a performance sufficient to induce a rater of severity *Z* to award a desirable score on a task of difficulty *D*.

Rasch measures are constructed through reverse inference. We observe a pattern of responses or performances for a set of ordered tasks, and make a rational and systematic inference about the ability or attitude of the person who produced these responses. Table 1 provides a summary of the four major Rasch models discussed in this installment. Future installments will deal with practical issues involved in conducting Rasch analyses, such as sample size and test length, and the problematic issue of guessing in multiple choice test formats.

Table 1. *A summary of the major Rasch models*

Model	Example tasks	Fundamental inference
Dichotomous model	Select or produce the one acceptable answer to a test item	How much of the attribute is required to succeed at this task?
Rating scale model	Honestly endorse a response category on a questionnaire	How much of this attribute (attitude, trait, orientation, etc.) is required to (strongly) agree or disagree with this statement?
Partial credit model	Choose or produce the better of several possible responses	How much of the attribute is required to enable one to reject some attractive distractors?
	Complete one or more sequential steps in a complex task	How much of the attribute is required to complete this many steps of the task?
Many-facets model	Attain a desirable score from a rater observing a performance, usually further defined by a set of prompts and descriptors	How much of the attribute is required to generate a performance sufficient to induce a rater of severity *Z* to award a score of *Y* on a performance task of difficulty *D*?

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrica, 43,* 561-573.

Linacre, J. M. (1992). *Many-facet Rasch measurement*. Chicago, IL: Mesa Press.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrica, 47,* 149-174.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Denmarks Paedagogiske Institut.

Sick, J. R. (2007). *The learner's contribution: Individual differences in language learning in a Japanese high school.* Unpublished Doctoral Dissertation, Temple University Japan, Tokyo.

Sick, J. R., & Irie, K. (2000). The Lunic Language Marathon: A new aptitude instrument for Japanese learners. In S. Cornwell & P. Robinson (Eds.), *Individual differences in foreign language learning: Effects of aptitude, intelligence, and motivation* (pp. 173-187). Tokyo: Aoyama Gakuin University.

Weaver, C. (2005). Using the Rasch model to develop a measure of second language learners' willingness to communicate within a language classroom. *Journal of Applied Measurement, 6* (4), 396-415.