**Comparing "Partial Credit Models" (PCM) and "Rating Scale Models" (RSM)**

"I'd like to compare a Rasch rating scale model and a partial credit model with the same data. Are there ways to compare the two models?"

*Hiroyuki Yamada*

A "rating scale" model is one in which all items (or groups of items) share the same rating scale structure. A "partial credit" model is one in which each item has a unique rating scale structure. This increases the number of free parameters to be estimated by (L-1)\*(m-2) where L is the number items and *m* the number of categories in the rating scale. A statistician might reply "Comparing models is easy! Compute the difference between the chi-squares for the models. Then test the hypothesis that the partial credit model fits no better than the rating scale model." A measurement practitioner might suggest "Compare the sample separation indices. Has there been a noticeable improvement in measure discrimination?" In practice, however, estimate stability and the communication of useful results all push towards fewer rating scale parameters.

**Estimate stability.** With the partial credit model, there may be only a few, or maybe no, observations in some categories of some items. The difficulty estimates of those items will be insecure, and a weak basis for inference (which is why we are bothering to do this analysis, isn't it?). A slightly better fit of the data to the model is of no value if it does not lead to a stronger basis for inference. [At least 10 ratings per category are recommended.]

**Communication of useful results.** If all the items or groups of items use the same response format, e.g., Strongly Disagree, Disagree, Agree, Strongly Agree, then the test constructors, respondents, and test users all perceive those items to share the same rating scale. To attempt to explain a separate parameterization for the rating-scale structure of each item would be an exercise in futility. It may be that one or two items clearly have unique structures. For instance, they may be badly-worded "true-false" item stems with Likert responses. Model most items to share the common rating scale(s), and allow just those very few idiosyncratic items to have their own scales. Even then, expect your audience to stumble as you try to explain your construct and your substantive findings. It will probably be more productive simply to omit those aberrant items.

*John Michael Linacre*

*See also*[*Model selection: Rating Scale Model (RSM) or Partial Credit Model (PCM)?, 1998, 12:3 p. 641-2*](http://www.rasch.org/rmt/rmt1231.htm)*.*

*Comparing "Partial Credit Models" (PCM) and "Rating Scale Models" (RSM), Linacre, J.M. … Rasch Measurement Transactions, 2000, 14:3 p.768*

*Later note:*

PCM vs. RSM is a difficult decision. Here are some considerations:  
  
1. Design of the items.  
If items are obviously intended to share the same rating scale (e.g., Likert agreement) then the Rating Scale Model (RSM) or Group Rating Scale Model is indicated, and it requires strong evidence for us to use a Partial Credit Model (PCM).  
But if each item is designed to have a different rating scale, then PCM (or grouped-items) is indicated, and it requires strong evidence for us to use RSM.  
  
2. Communication with the audience.  
It is difficult to communicate many small variants of the functioning of the same substantive rating scale (e.g., a Likert scale). So the differences between items must be big enough to merit all the extra effort.   
Sometimes this difference requires recoding of the rating scale for some items. A typical situation is where the item does not match the response-options, such as when a Yes/No item "Are you a smoker?" has a Likert-scale responses: "SD, D, N, A, SA".  
  
3. Size of the dataset.  
If there are less than 10 observations in a category used for estimation, then the estimation is not robust against accidents in the data. In most datasets, RSM does not have this problem. But a dataset of items with 4 response categories, but only 62 measured persons, indicates that some items may have less than 10 observations in some categories. This may be evidence against PCM.  
  
4. Construct and Predictive Validity.  
Compare the two sets of item difficulties, and also the two sets of person abilities.  
Is there any meaningful difference? If not, use RSM.  
If there is a meaningful difference, which analysis is more meaningful?  
For instance, PCM of the "Liking for Science" data loses its meaning because not all categories are observed for all items. This invalidates the item difficulty hierarchy for PCM.  
  
5. Fit considerations.  
AIC is a global fit indicator, which is good. But global-fit also means that an increase in noise in one part of the data can be masked by an increase in dependency in another part of the data. In Rasch analysis, underfit (excess non-modeled noise) is a much greater threat to the validity of the measures than overfit (over-predictability, dependency). So we need to monitor parameter-level fit statistics along with global fit statistics.  
  
Global log-likelihood is intended to compare descriptive (explanatory) models, for which the model fits the data. Rasch models are prescriptive models for which the data fit the model. We "prescribe" the Rasch model with the properties we need for our measures to be useful, and then fit the data to that model. If the fit is poor, then the data are deficient. We need better data (not a better model) for the purposes for which our data are intended.  
  
6. Constructing new items.  
With the partial-credit model, the thresholds of the rating-scale are unknown in advance of data collection. With the rating-scale model, they are the current thresholds.  
  
7. Unobserved categories.  
With the partial-credit mode, unobserved (in this sample) categories for an item distort the rating-scale structure. With the rating-scale model, the functioning of an unobserved category for one item is inferred from observations of the same category for other items.

8. Statistical information.  
For rating scales with the same number of categories, the total statistical information for an item is the same for the partial-credit and rating-scale models  
  
9. Antifragility.  
"The simpler, the better. Complications lead to multiplicative chains of unanticipated effects." Nassim Nicholas Taleb, Antifragile: Things That Gain From Disorder. Random House/Penguin. 2012