Viiic: More than One; Less than Infinity

For many testing situations, simple zero-one scoring is not enough and Poisson-type counts are too much. Polytomous Rasch models (*PRM*) cover the middle ground between one and infinity and allow scored responses from zero to a maximum of some small integer *m*. The integer scores must be ordered in the obvious way so that responding in category *k* implies more of the trait than responding in category *k-1*. While the scores must be consecutive integers, there is no requirement that the categories be equally spaced; that is something we can estimate just like ordinary item difficulties.

Once we admit the possibility of unequal spacing of categories, we almost immediately run into the issue, Can the thresholds (i.e., boundaries between categories) be disordered? To harken back to the baseball discussion, a four-base hit counts for more than a three-base hit, but four-bases are three or four times more frequent than three-bases. This begs an important question about whether we are observing the same aspect with three- and four-base hits, or with underused categories in general; we'll come back to it.

To continue the archery metaphor, we now have a number, call it *m*, of concentric circles rather than just a single bull's-eye with more points given for hitting within smaller circles. The case of $m=1$ is the dichotomous model and $m \to \infty$ is the Poisson, both of which can be derived as limiting cases of almost any of the models that follow. The Poisson might apply in archery if scoring were based on the distance from the center rather than which one of a few circles was hit; distance from the center (in, say, *millimeters*) is the same as an infinite number of rings, if you can read your ruler that precisely.

Rating scale model (RSM)

The rating scale model (Andrich, 1978; Wright & Masters, 1982) characterizes the person's responses as a simple function of the person's condition (e.g., attitude, preference, status, level of pain, ability), the item's strength, and several levels up or down for the response categories. The model is used more frequently for attitude, preference, or evaluation questionnaires than achievement testing. One common format is a series of statements that the respondent is asked to react to on, say, a five-point scale from "*strongly disagree"* to *"strongly agree*".

If we are considering, for example, the statement:

"*The Rasch model is the very definition of measurement*"

and the response format is:

Strongly Disagree	*Disagree*	*Don't Care*	*Agree*	*Strongly Agree*

and we intend to respond either in category "*agree*" or category "*strongly agree*", the probability of choosing "*strongly agree*" over "*agree*" is:

13. $$p^*(k \mid \beta_v, \delta_i, \tau_k) = \frac{p(k)}{p(k-1)+p(k)} = \frac{e^{\beta_v-(\delta_i+\tau_k)}}{1+e^{\beta_v-(\delta_i+\tau_k)}},$$

where response category *k* is "*strongly agree*" and *p(k)* is the unconditional probability of responding in category *k*, which we have not yet revealed. Because, at this point, we are considering only two categories, expression (13) is identical to the dichotomous case with the

item difficulty δ_i replaced by $\delta_i + \tau_k$. The categories other than *k* and *k-1* do not enter into the equation.

The distinction between *p* and *p** is that *p* is the probability of responding *"strongly agree"* without restriction but *p** is the probability of *"strongly agree"*, given the response is either *"agree"* or *"strongly agree"*. We have, in effect, already dismissed the less positive responses from our consideration. It may help some to say that, $\sum_{x=0}^{m} p_x = 1$ and $\sum_{x=k-1}^{k} p_x^* = 1$.

Apply a little algebra to expression *(13)* and we have a recursive expression for *p(k)*:

14. $$p(k) = e^{\beta_v - (\delta_i + \tau_k)} p(k-1).$$

Significantly, the likelihood of moving from *(k-1)* to *(k)* involves only the parameter for *(k)* and none of the others. It does not matter hard it was to get to *(k-1)* or how easy it might be to get beyond *(k)*.

Equivalent and sometimes more convenient than *(14)*, the log odds of *k* versus *k-1* (i.e., logit):

15. $$\ln\left\{\frac{p(k)}{p(k-1)}\right\} = \{\beta_v - (\delta_i + \tau_k)\}.$$

Once again the logit is the log odds of one outcome versus the other.

While we now have an expression for *p(k)*, we need a starting point. It is convenient, and no more arbitrary than any other value, to define the logit for category *0* as *0*, and then the probabilities can be developed as in Table 3[1].

Table 3: Response Category Probabilities for a Rating Scale Model

k	*Logit*	*Numerator*	Pr*obability*
0	0	1	$1/\gamma$
1	$\alpha_1 = \beta_v - (\delta_i + \tau_1)$	e^{α_1}	e^{α_1}/γ
2	$\alpha_2 = \beta_v - (\delta_i + \tau_2)$	$e^{\alpha_1+\alpha_2}$	$e^{\alpha_1+\alpha_2}/\gamma$
3	$\alpha_3 = \beta_v - (\delta_i + \tau_3)$	$e^{\alpha_1+\alpha_2+\alpha_3}$	$e^{\alpha_1+\alpha_2+\alpha_3}/\gamma$
4	$\alpha_4 = \beta_v - (\delta_i + \tau_4)$	$e^{\alpha_1+\alpha_2+\alpha_3+\alpha_4}$	$e^{\alpha_1+\alpha_2+\alpha_3+\alpha_4}/\gamma$
Total	$m(\beta - \delta) - \sum\tau$	γ	1.0

For the sake of completeness and the compulsively mathematical, the relationships of *Table 3* can be captured in the standard expression of the rating scale model for the probability that person v taking item *i* will respond in category *k*, given the person parameter β_v, the item parameter δ_i, and *m* category parameters τ_j:

[1] The γ term in Table 3 is a normalizing constant to make the probabilities sum to one. It is nothing more or less than the sum of the numerators. If we were being more rigorous, a form of this constant would be introduced in expressions (12) and (13).

16. $$p\{k \mid \beta_v, \delta_i, (\tau_{j=1,m})\} = \frac{e^{\sum_{j=1}^{k}(\beta_v - \delta_i - \tau_j)}}{1 + \sum_{x=1}^{m} e^{\sum_{j=1}^{x}(\beta_v - \delta_i - \tau_j)}} = \frac{e^{k(\beta_v - \delta_i) - \sum_{j=1}^{k}\tau_j}}{1 + \sum_{x=1}^{m} e^{x(\beta_v - \delta_i) - \sum_{j=1}^{x}\tau_j}}.$$

The summation in the exponent represents the summing of logits in Table 3; the summation in the denominator is the summing of the numerators, the numerator of *k=0* being one. This is the *p* needed for expressions 13 to 15.

Figure 2 shows the Category Characteristic Curves (colors) for a five-category item with nicely spaced categories. The category parameters used to create the plot are (-3, -1, 1, 3). These parameters appear in the figure as the intersections between adjacent categories. The curves for category *Fail* and category *Marginal* cross at *-3*; the curves for category *Marginal* and category *Pass* cross at *-1; etc*. The only intersections that matter are those for adjacent categories. Items never look like this in real life.

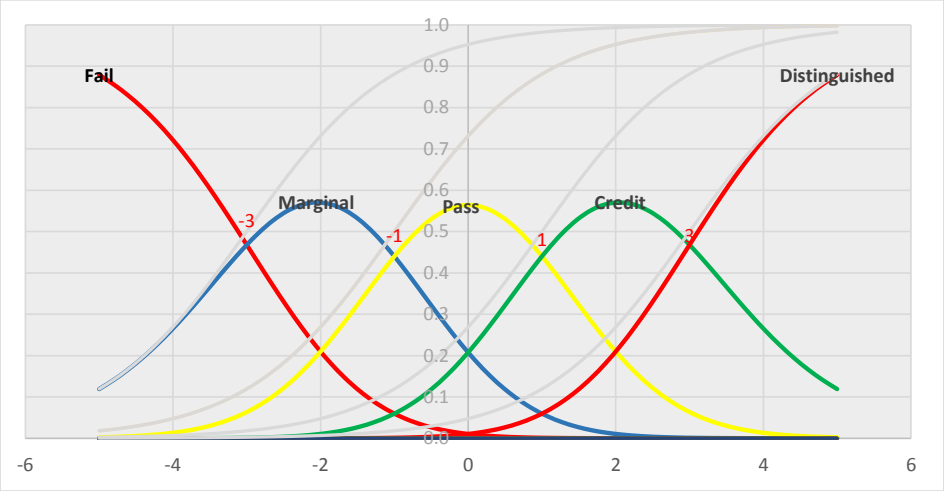

Figure 2: Rating Scale with Four Equally Spaced thresholds: parameters = (-3, -1, 1, 3)

The lighter gray curves are *Threshold Characteristic Curves* that show the probability of being in the higher of two adjacent categories. These look just like the *Item Characteristic Curves (ICC)* we are familiar with for dichotomous items because that is what they are: the probability of being in category *k* rather than *k-1*. They will always cross 0.5 on the vertical axis at threshold value on the horizontal. In the language of expression (13), these are the *p**; the more colorful Category Curves are the *p*'s. As required by (13), *p** is always above *p*.

A person in category *k* is not described adquately by the category parameter. For this example, although τ_2 is *-1.0*, the most likely value for a person's location, given an observed category of *2*, is a logit of *0.0*. Because of the symmetry of this example, this estimate happens to be half-way between adjacent threshold values.

Table 4 illustrates some of the calculations behind Figure 2; specifically, the calculations needed for a point on each curve where $\beta_v - \delta_i = 1$. Column *1* is the category score *k*. Column *2* is the threshold parameter τ_k; as with the dichotomous case that had two categories and one parameter, there is one fewer threshold than categories. The third column, $\exp(\beta_v - \delta_i - \tau_k) = \exp(1 - \tau_k)$, is the exponentiation at the point on the logit continuum where the person parameter exceeds the

item parameter by one logit. The *Numerator* is the exponentiation times the previous numerator. The *Probability* is the *Numerator* divided by the sum of the numerators. This is a repeat of Table 3 with numbers instead of symbols.

Table 4: Category probabilities for $\beta-\delta=1$ and $\boldsymbol{\tau}$ = (-3, -1, 1, 3).

Category	Logit	$exp(1-\tau_k)$	Numerator	Probability
0			1	0.001
1	-3	54.60	54.60	0.060
2	-1	7.39	403.43	0.440
3	1	1.00	403.43	0.440
4	3	0.14	54.60	0.060
Sum			917.06	1.000

Partial credit model (PCM)

The *Partial Credit* model (Masters, 1980; Wright & Masters, 1982) looks almost identical to the rating scale model. When looking at a single item, the models are indistinguishable. There is nothing about Figure 2 that says rating scale, not partial credit. Tables 3 and 4 can be used here just as well if an *i* is added to the subscript of each τ_j. If we continue to belabor the archery metaphor, in addition to concentric circles of different sizes, different targets may use different numbers or patterns for the concentric circles.

Based on the original rationalization of partial credit scoring, the category parameters are typically referred to as *steps*. This is the point on the continuum where the person has completed one *step* in the problem solution, receives credit for that work, and begins work on the next *step*. As with the rating scale formulation, this is the point on the scale at which the two adjacent categories are equally likely. We will follow Andrich and refer to the category parameters as *thresholds*.

For the mathematically inclined, the partial credit model for the probability of person ν responding in category *k* on item *i*, given the person parameter β_ν and the m_i item parameters $\delta_{ij}=\delta_i+\tau_{ij}$, may be written as:

17. $$p\{k \mid \beta_\nu, (\delta_{i,j=1,m_i})\} = \frac{e^{\sum_{j=1}^{k}(\beta_\nu-\delta_{ij})}}{1+\sum_{x=1}^{m_i} e^{\sum_{j=1}^{x}(\beta_\nu-\delta_{ij})}} = \frac{e^{k\beta_\nu-\sum_{j=1}^{k}\delta_{ij}}}{1+\sum_{x=1}^{m_i} e^{x\beta_\nu-\sum_{j=1}^{x}\delta_{ij}}}$$

The distinction between the rating scale model (16) and the partial credit model (17) is that the category parameters, τ_j, have now been subsumed under the item parameters δ_{ij}. The practical implication of this change is that the response categories can differ across items; they can be different formats or have different numbers of categories. For attitude or preference questionnaires, this may mean that different response categories are used for each statement (e.g., *agree-disagree* versus *never-always; four-point scales* versus *five-point*). For achievement testing, it may mean *zero* points are given for completely wrong answers, m_i points are given for completely right answers, and integer scores between *zero* and m_i are given for partially correct answers according to the item rubric, with the maximum points m_i and the scoring rubric for the partial credit specific to each item.

It is a matter of style or context whether the partial credit model is written in terms of δ_{ij} or $\delta_i+\tau_{ij}$. The first form implies the item is best described by the m_i threshold values; then $\beta_v - \delta_{ij}$ looks like an extension of the dichotomous case. The second form implies the item can be described by a single location with the thresholds given as offsets around that; then $\beta_v - (\delta_i + \tau_{ij})$ looks like a generalization of the rating scale model. While it doesn't much matter which version, δ_{ij} or τ_{ij}, of the thresholds you use, it is useful to know which one your software gave you.