Ordered Categories, Disordered Thresholds

When the experts all agree, it doesn't necessarily follow that the converse is true. When the experts don't agree, the average person has no business thinking about it. B. Russell

The experts don't agree on this topic and I've been thinking about it anyway. But I may be less lucid than usual.

The categories, whether rating scale or partial credit, are <u>always</u> ordered: *0* always implies less than *1; 1* implies less than *2; 2* implies less than *3* . . . The concentric circle for *k* on the archery target is always inside (smaller thus harder to hit) than the circle for *k-1*. The transition points, or thresholds, might or might not be ordered in the data. Perhaps the circle for *k-1* is so close in diameter to *k* that it is almost impossible to be inside *k-1* without being inside *k*. Category *k-1* might be very rarely observed, unless you have very sharp arrows and very consistent archers.

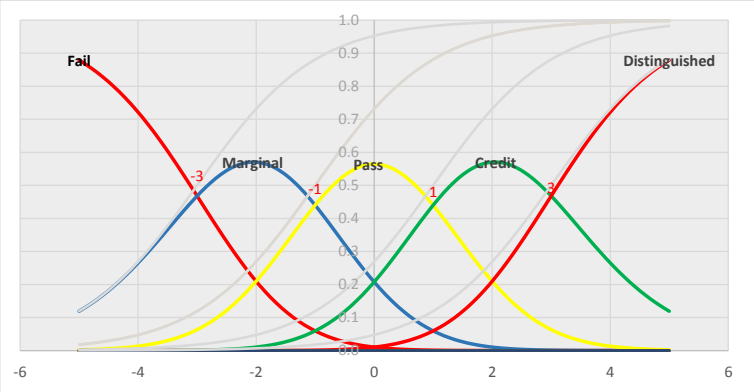

Figure 2: Rating Scale or Partial Credit with Five Categories and Ordered Thresholds: parameters = (-3, -1, 1, 3)

In Figure 2 repeated above, everything was ordered nicely; Figure 3, below, illustrates another four-point item but the second and third thresholds have been reversed giving disordered threshold values of *(-3, 1, -1, 3)*. Category *Pass* becomes more likely than *Marginal* at a logit value of *1* but category *Credit* became more likely than *Pass* at a logit value of *-1*.

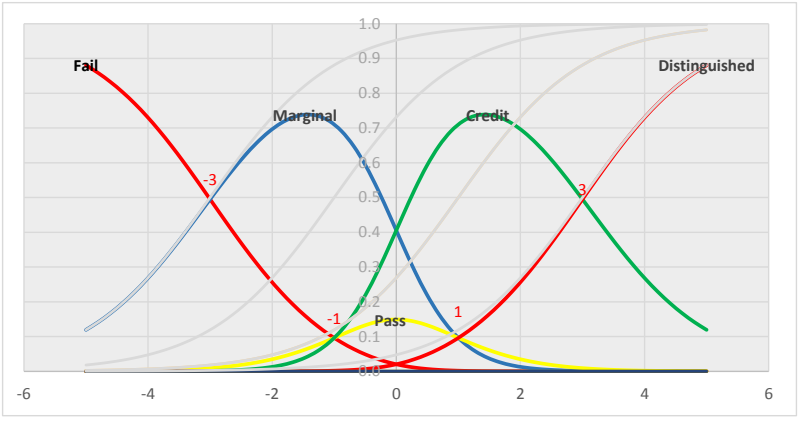

Figure 3: Rating Scale or Partial Credit with Five Categories and Disordered Thresholds: parameters = (-3, 1, -1, 3)

There is no point on the continuum for which category *Pass* is the most likely response for the person. The person who is most likely to be in *Pass* has a logit location of -0.5; however, a

person at this location is more likely to be in either *Marginal* or *Credit*. A person who is strong enough to leave *Marginal* is unlikely to stop at *Pass* but is expected to go immediately to *Credit*. In spite of this confusion of threshold parameters, the category curves are still in the natural order: being in *Pass* implies more than being in *Marginal* and less than being in *Credit*.

If we go half way and set the second threshold equal to the third (*i.e.*, *-3, 0, 0, 3*), we get Figure 4 with the three curves crossing at zero. The *Pass* category is more likely than it was but still never the most likely.

Figure 4: Rating Scale or Partial Credit with Five Categories and Ordered Thresholds: parameters = (-3, 0, 0, 3)

Setting the thresholds very far apart (-3, -2, 2, 3) makes the *Pass* category very likely over a wide range as shown in Figure 5.

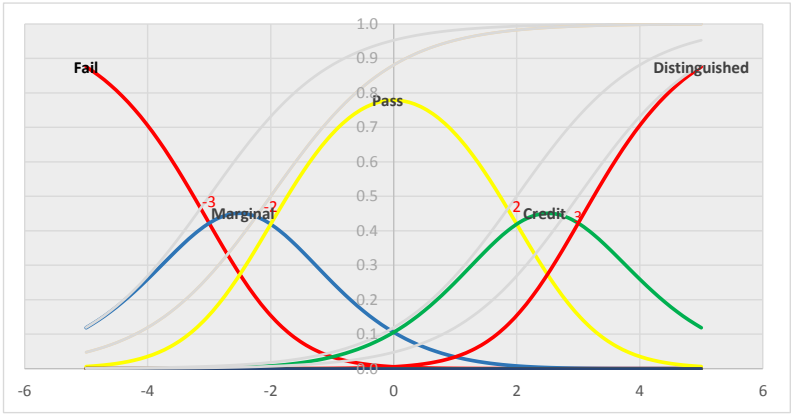

Figure 5: Rating Scale or Partial Credit with Five Categories and Ordered Thresholds: parameters = (-3, -2, 2, 3)

The category curves in the figures pertain to individuals; they are the probabilities (*vertical axis*) that a person at each point on the continuum (*horizontal axis*) will respond in each of the categories. They are not empirical frequencies that have been or might be observed for any population, real or imagined. Whether or not any categories are under-represented or over-represented in any particular analysis depends on the distribution of the sample of people. The observed frequencies are sample-dependent. Estimates of the thresholds are not sample-dependent; we should be able to recover them consistently and objectively from any decent sample.

There is some "discussion" about how to react to disordered thresholds. Masters, and others, arguing from the data and tests of fit, contends that nothing in the model is violated and it simply reflects an under-used category, which can be informative if less than optimal. As the plots above show, we have no trouble talking about and visualizing disordered thresholds. The standard tests of fit do not object to the disorder.

Andrich, arguing from the underlying Guttmannesque nature of categories, asserts that the disordering of thresholds discloses the presence of an anomaly in the data, not the model, so the tests of fit are answering a different question. Ordering of the categories is mandated when they are created: *marginal* requires more than *fail; pass* requires more than *marginal; Distinguished* requires more than *Credit; Strongly Agree* requires more than *Agree*. Ordering of thresholds is presumed when we accept a response in a single category to represent the person's status on the item. We are effectively allowing only the Guttmann patterns shown in Table 5. Any items showing disordering are allowing non-Guttmann responses and should be reviewed, revised, or discarded.

The debate started with the appearance of the models circa 1980 and continues to the present.

The archery metaphor used a bull's eye target with concentric rings and the score was the smallest ring the arrow touched, numbered from the largest to smallest, zero if you miss it entirely. The notion of disordered thresholds was described in terms of two rings so close in size that it is difficult to touch the larger without touching the smaller. We would expect these thresholds to be disordered and have no reason not to expect the data would fit the model.

The target is in effect a very crude ruler for measuring how far the arrow is from the center. What the disordered thresholds are telling us is that we have a poor ruler (and poor target maker) and there could be a substantial increase in score for a very small improvement in performance. The problem isn't so much that one ring is too small but that the adjacent ring is too big. There would be no problem if, for example, we have lots of rings, say, one millimeter apart. The effect of this improvement[1] in our ruler simply underscores the inherent ordering of the categories and thresholds.

If our five-category item were decomposed into a series of four dichotomous items, the *16* theoretically possible response strings and related categories are shown in Table *5*. Only five of the 16 hypothetical response patterns match the Guttmann patterns. Our requirement that the categories be strictly ordered (i.e., *4* implies more than *3*, etc.) means that those 5 Guttmann patterns are the only ones consistent with the Rasch view of the world. And, because the person is only giving one compound response to the five imagined items, we logically associate one unique Guttmann pattern with each of the five response categories. And, if those are the only patterns allowed, how do we get any disordered thresholds?

Andrich (2013, 2014) frames the argument in terms of a judge classifying a student essay into one of five[2] categories: *Fail, Marginal, Pass, Credit,* and *Distinguished*, which are clearly ordered, both in terms of the labels and the protocols the judges use to sort the essays. The thresholds have a natural, intrinsic order. For a paper to be considered for *Distinguished* over *Credit,* any consideration about *Credit* versus *Pass* should be a foregone conclusion. The data

[1] There is some point when we should look at a *Poisson* model rather than *Rating Scale Model.*

[2] Andrich discusses three or four categories but I've added "Marginal" to make it match my diagram.

may or may not confirm this ordering. If they don't, then the model has exposed an anomaly that should be investigated.

Table 5: Hypothetical and Guttman Response Patterns

Category	One	Two	Three	Four	Score
0	0	0	0	0	0
1	1	0	0	0	1
2	1	1	0	0	2
3	1	1	1	0	3
4	1	1	1	1	4
	0	0	0	1	1
	0	0	1	0	1
	0	0	1	1	2
	0	1	0	0	1
	0	1	0	1	2
	0	1	1	0	2
	0	1	1	1	3
	1	0	0	1	2
	1	0	1	0	2
	1	0	1	1	3
	1	1	0	1	3

When faced with an essay whose true location is, say, *-0.95* logits (just above the threshold between *Pass* and *Credit* in Figure 3) a perfect judge would reason, *I believe this paper is more likely to be* Credit *rather than* Pass *but I know enough about conditional probabilities to know it really belongs in* Marginal. If we think about independent judges[3] making dichotomous decisions (i.e., *Marginal* vs. *Pass; Pass* vs. *Credit;* and *Credit* vs. *Distinguished*,) a judge for the *Marginal* vs. *Pass* decision would be comfortable declaring the essay *Marginal*; the judge for the *Pass* vs. *Credit* decision would lean toward *Credit* over *Pass*. Andrich believes we would have difficulty defending this to the parents or their attorney.

I will leave the math and debate to authors more invested (Adams, Wu, & Wilson, 2012; Andrich, 2013, 2014) and take refuge in my two analogies, archery and baseball. For baseball as a *Rating Scale Model*, the batter will get a score on the batting task of *0* to *4* based on how many bases are touched. These categories are clearly ordered; you can't get to second without touching first first, and so on. You don't actually score a point in baseball until you touch all four bases, but your likelihood of doing that improves with each base reached.

Table 6: Reasonable Likelihoods of Bases Reached

	RELATIVE FREQUENCIES		
CATEGORY	Team	Batter 1	Batter 4
0	0.68	0.65	0.75
1	0.23	0.30	0.20
2	0.05	0.04	0.01
3	0.01	0.01	0.00
4	0.03	0.00	0.04

[3] Andrich (2013) established the equivalence between four judges making independent, dichotomous decisions and one judge making a single, non-independent, compound decision.

As noted in an earlier example, a reasonably good team might have category frequencies for the five categories *0* to *4* respectively that are shown in the second column of table 6. This is the data and it would typically lead to disordered thresholds. It may or may not pass the Pearsonian tests of fit but that's a different question. Category frequencies for two not necessarily typical batters are shown in the last two columns.

If we know nothing else about the data, we might speculate (and it is speculating) that category 4 is very close to category 3 so that it is easy to reach 4 if we reached 3. But we do know some other things, if you grew up with baseball rather than cricket. Four-base hits (in organized baseball) happen almost exclusively by hitting the ball over the fence and out of the park. Having done that, it does not matter how long you take to touch the bases in the proper order. Three-base hits happen exclusively by hitting the ball in the field of play and running fast enough to reach third before the fielders have managed to retrieve it. The batter with lots of *4*s depends heavily on power; the batter with any *3*s depends more on speed.

We might have speculated earlier that the data arose because the distribution of batters, while unidimensional, is bimodal. If there were a few very proficient batters who almost always get *4*s, then the model could fit fine. However, this speculation is about the empirical distribution of all batters on the team, not category probabilities of individuals. If we try to put *batter 4* into the class of very proficient because of all the *4*s, we would have to explain why so many *0*s. The basic idea is that there are two very different ways to score well (speed or power) and they lead to very different response patterns in the data.

Returning briefly to archery and measuring the distance from the edge, disordered thresholds are analogous to having an arrow on the line between, say, 99 and 100 millimeters and assigning a score of 98 based on the conditional probabilities. With my grounding in Fisher, I am inclined toward the view of using the model to expose anomalies in the data and using the theory to explain them rather than the more Pearsonian approach of fitting a model to the data and calling that "explaining." *Models must be used but never believed*. (M. Wilk)