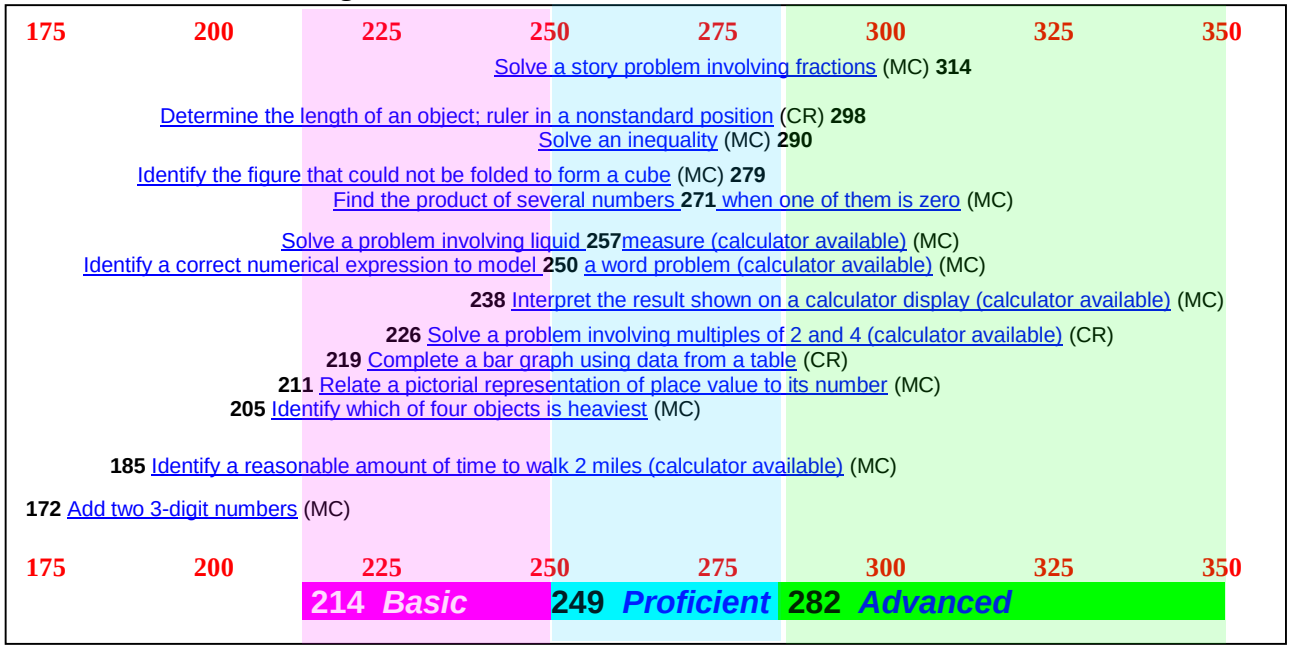
A Spectrum of Math Proficiency and the Specter of Word Problems

In mathematics, one does not understand anything. You just get used to them.
Johann Von Neumann

There are 10 kinds of people in the world: those who don't understand base 2 and those who do. More generally, the world has divided itself in two factions: those who think they don't understand math and those who think they do. But we're not talking about proving Fermat's Last Theorem or correcting Stephan Hawking's tensor algebra; we're talking about counting, applying the four basic operators, and solving the dreaded word problems using basic algebra, geometry, and perhaps a little calculus. That just about covers the range from counting your toes to determining the spot on which a fielder should stand to catch a batted ball and should be good enough to get you through freshman math.

Just as we have little need for one thermometer that covers the range from liquid hydrogen to the surface of the sun, we rarely need a math assessment that covers the range from counting to calculus. Figure II.3 shows a short segment of the math spectrum, like the "*roy*" portion of the visible light spectrum, with three-digit addition at the low end and word problems with fractions at the high end. Anyone familiar with elementary math education will recognize, even with these rather terse descriptors, a reasonable progression from easiest to hardest. This is the first step in making the scale meaningful. The performance bands add another layer of meaning, if you understand and accept what the developer means by *Basic*, *Proficient*, and *Advanced*. Performance Bands are more arbitrary than the order and spacing of the items, more like drawing a line between *red* and *orange* for color or between *just right* and *too hot* for temperature.

Figure II-3: A Short Segment of the Mathematics Spectrum



The scale locations of the items and Performance Bands are expressed in an arbitrary (and on its own meaningless) Scale Score metric, but there is one thing we can say easily: a student with a Scale Score equal to the scale location for an item has a 50% likelihood of answering that item correctly. A person with a Scale Score of 250 (just barely into the Proficient band) has a 50-50 chance of answering the item about identifying the correct mathematical expression. The chances are greater than 50% for any lower item and less than 50% for any higher item. We will need to get into the logit metric to be more specific.[1]

Because Rasch's method wasn't used to develop this assessment, the mathematics behind Figure II.3 are less straightforward, but we can tell a plausible story if we assume Rasch and a transformation from logits to scale scores of *Scale Score = 250 + 25(Logit)*. Based on this, Table II.7 shows the items' location in Scale Score and Logit metrics and the probability that a student at Scale Score 250 has of answering each item.

Table II.7: Locations and Probabilities for Item on the Assessment for Student at 250

Item Location		Probability for Student at 250
Scale Score	Logit	
314	2.56	0.07
298	1.92	0.13
290	1.60	0.17
279	1.16	0.24
271	0.84	0.30
257	0.28	0.43
250	0.00	0.50
238	-0.48	0.62
226	-0.96	0.72
219	-1.24	0.78
211	-1.56	0.83
205	-1.80	0.86
185	-2.60	0.93
172	-3.12	0.96

This is a simple extension of what we just said, that a person at 250 (i.e., logit = 0) will have a 50% likelihood on the item at 250 (logit = 0). For any other probability, we use the same expression we used for football; the odds for the person divided by the odds for the person plus the odds for the item:

7. *Probabilty = exp(person logit) / {exp(person logit) + exp(item logit)}*

With the same type of arithmetic, one can also say that our 250 student has a 22% chance of answering an item on the line between Proficient and Advanced (Scale Score = 282; logit = 1.24) and a 19% chance of missing an item on the line between Basic and Proficient (Scale Score = 214; logit = -1.44). These results can't be read directly from the Table II.7 and you need to use logits in expression 13.

[1] This assessment was developed with methods that do not necessarily conform to Rasch principles, so what I am saying here, which assumes Rasch, may not be strictly true. For example, the ordering of the items may not be consistent when viewed from different points on the scale, but these methods, in spite of the mathematical contortions, are a reasonably good approximation to Rasch in practice.

The thing that really matters in these calculations is not the locations but the distance between locations. We can make the discussion a little more general if we switch to differences rather than locations. We already know that a difference of zero (in whatever metric) means a probability of 0.5. Table II.8 provides a few more values. For example, if the person exceeds the item by 75 Scale Scores (or equivalently 3 logits), the probability of success will be 0.95. The relationship is symmetric; if the item exceeds the person by 75 Scale Scores, the probability that the item wins is 0.95 (or 0.05 that the person wins.)

Table II.8: Differences between Person and Item and the Probabilities

Difference		Probability
Scale Score	Logit	
-75.0	-3.0	0.05
-62.5	-2.5	0.08
-50.0	-2.0	0.12
-37.5	-1.5	0.18
-25.0	-1.0	0.27
-12.5	-0.5	0.38
0.0	0.0	0.50
12.5	0.5	0.62
25.0	1.0	0.73
37.5	1.5	0.82
50.0	2.0	0.88
62.5	2.5	0.92
75.0	3.0	0.95

There is also a back-of-the-envelop calculation that can be done with the Rasch logit for a person and the performance bands. If we are using a test of 50 items that are like the ones in Table III.7 and an approximation for the error of measurement[2], we can determine the likelihood that our standard 250 person, who is now classified as *Proficient*, would be classified as *Basic* if tested again by resorting to the standard normal distribution[3]. For this person:

8. $Z = (249 - 250) / 9 = -0.11$.

The standard Normal deviate of -0.11 has 46% of the area under curve to its left, which in our context means the probability that the student could have landed in *Basic* rather than *Proficient* is 0.46.

We might also ask about a student located right in the middle of the *Proficient* Band at 265. The same process would give a probability of 4% that student would be in *Basic* and a probability of 3% in *Advanced*, if retested. None of this is particularly surprising and should not be disturbing, just cautionary. Fifty items is not a particularly long test and the whole process is imperfect at best. And for the *265* person, we are 93% certain we have the right Performance Band.

---

[2] Wright and Douglas (1974) give an approximate logit standard error equal to $2.5/\sqrt{L}$ where L is the test length. This will be discussed later in more detail along with more precise formulas. We are also making a few other simplifying assumptions about the Normal distribution applying. And there is something disconcertingly Bayesian about my language in this section.

[3] Note we are not assuming a normal distribution of person abilities, which would require specifying a population. We are resorting to the *normal distribution of errors*, which is where Gauss started and that's a big enough leap of faith in this case.

Before Science, Measurement

The process that we have gone through, albeit rather crudely, to obtain useful evidence is:

- What aspect of the people are we trying to understand?
- What types of evidence might we collect that would relate to the status of the people in this aspect?
- What groups of people do we intend to measure?
- Can we reasonably expect this evidence to be *valid* for comparing any and all members of these groups?

These deliberations are basic instrument development and hardly invented by Rasch. They are recounted here to emphasize the crucial roles of the instrument and the rationale supporting it. One cannot expect to take a poorly thought-out instrument and salvage it through mathematical gymnastics or psychometric incantations. The Rasch analysis of the data is an important tool but certainly should not start the discussion and definitely does not end it.