Parametric Item Response Theory

W. Blaine Dowler

May 28, 2011

1 The Need For Better Analysis

The tools presented in the previous lesson are adequate for a single classroom, but may not be adequate for larger populations. The results are closely tied to the effectiveness of the instruction within that particular classroom, and that will vary from class to class. Furthermore, the difficulty of a question is tied specifically to the students in the analysis group. The difficulty of an assessment item given to a grade three classroom will probably not match the difficulty measured when giving the same item to a grade nine classroom, whatever the item is. We need a way to evaluate individual items on a larger scale, and with more accuracy and flexibility.

2 The Three Rasch Paramaters - Conceptual Overview

Georg Rasch was the first to develop and popularize a model for working with detailed analysis of assessment items which was consistently decoupled from any single individual classroom or grade level. His work actually formed a series of models, based on one, two, or three parameters.

Mathematically speaking, a parameter is a variable which is set for a particular question. It is, in essence, a number which varies from assessment item to assessment item, but not from student to student. The the first two of Rasch's three parameters will be familiar to those who have read lesson six.

2.1 Difficulty

Rasch's first step was to refine the definition of difficulty. He proposed a continuous ability scale, and that students at all stages of academia could be plotted somewhere on this scale. In other words, rather than having a measuring scale for each individual grade, all questions of all difficulties in all grade levels can be mapped continuously, as though schools were designed around year-round schooling. He also proposed that an individual student who had a 50% chance of answering a question of a given difficulty (say, θ) would have an ability equal to that question's difficulty. This is a reasonable assumption, and nicely sets up an infrastructure for students in a particular classroom who are not performing on the same level as their peers.

With this scale established, Rasch proposed that the difficulty of an assessment item be measured by the ability at which a student has a 50% chance of answering correctly. In other words, items and students are set along this scale independently, and item difficulty is rated by where the item falls on this scale. This now reports the intrinsic difficulty of the item, rather than the performance of a specific group of students on that particular item.

In some simple models, this is the only parameter used. All other parameters are set to some arbitrary value for every item on the assessment under scrutiny. When Rasch's third parameter is introduced, the difficulty parameter must be reinterpreted. Rather than representing the ability of a student who has an exact probability of 50% of correctly answering the question, it becomes the ability of a student who has moved half way from the introduction of the skill to complete mastery of the skill. The difference arises because, when a one or two parameter fit is used, it is assumed that students cannot respond to an assessment item correctly by guessing.

2.2 Discrimination

For any given curricular outcome, there is a point at which no students are expected to answer the question correctly, and a point at which virtually all students are expected to answer the question correctly. Rasch's discrimination parameter deals with the area between these points.

For example, think of a grade six level math question. Grade one students are not expected to answer it correctly, but grade twelve students are. Grade six students are expected to answer with a mix of results, some students answering correctly, and others not. Similarly, some of the most capable grade five students may deduce the steps needed to answer correctly, while other students will still struggle with the question when working at the grade seven or eight

level. Rasch's discrimination parameter is related to the width of this period in which some students answer correctly while others answer incorrectly. A higher discrimination value leads to a shorter difference in ability between students who answer correctly and students who answer incorrectly. More complex skills tend to have lower discrimination values.

2.3 Pseudo-chance Level

Rasch's third parameter is the one most often omitted from implemented models. The *pseudo-chance level* of an item is the probability that a student with minimal ability will answer the item correctly with no knowledge of the skill involved. In other words, it is the probability of guessing correctly. In a four question multiple choice question, for example, this parameter should be at or near 25%. In a true-false question, this parameter should be at or near 50%. In most cases of omission, it is set to 0%, as though the only students who provide the correct response to an item are those who understand the skill well enough to arrive at the correct answer through actual implementation of the skill.

3 The Logic of Omission

All three of Rasch's parameters are reasonable, and would be expected to apply to any assessment item. Yet, the difficulty parameter is the only one consistently applied. Why is this done, and how realistic are the assumptions behind this?

Most decisions to limit things to a single parameter are based on the same, single criteria: lack of resources. Assessing these parameters with any useful level of accuracy requires administering the items to hundreds or thousands of students. If you double the number of parameters, you need to more than double the relevant student sample population. Furthermore, you need the computer processing resources to perform the analyses, and computing time scales more quickly than the student population, too. Moving from one to three parameters could mean increasing the required budget by more than a factor of 10. Others choose to reduce the parameters because of a mathematical anomaly which may occur only with the three parameter fit, as a result of students with inconsistent results. These cases will be described in more detail in lesson eight.

So, how does one ensure that the results are accurate with fewer parameters in use? This is done by performing a little trick common to the social sciences but abhorred in the harder sciences: if you cannot change the model to fit the data, you change the data to fit the model. Assessment items which do not conform to a one or two parameter fit are discarded entirely. This practice is

effective for norm-referenced assessments, but is open to dispute for criterion-referenced assessments, particularly with the more complex skills. The more complex the skill, the lower the discriminating power, the more likely students will just give up and guess, and the less likely it is to conform to a one or two parameter fit. Thus, norm-referenced assessments based on only one or two parameters tend to have extremely high statistical uncertainties at the highest levels of difficulty where the average complexity of a skill becomes large. Still, at lower levels and with less complex norm referenced skills, the one and two parameter fits work well enough.

Most pencil and paper based norm-referenced assessments are modeled using only the two parameters listed here. Very few incorporate Rasch's third parameter. Moreover, older pencil and paper assessments are often modeled as though all items on the assessment are at the same difficulty level with the same discrimination, whatever that may be. This is why the extrapolation methods mentioned in lesson four tend to become inaccurate when a student departs too greatly from the anticipated ability of students writing that individual assessment tool: when the third Rasch parameter is assumed to be zero, and when all items are treated as equivalent, the small errors that are present in the ability regions of interest become large errors when one departs that region. What the author will never understand is why Rasch's third parameter is set to 0 when not explicitly studied. It takes virtually identical amounts of computing power to set it to 0 as to 0.2, 0.25, 0.5, etc. to represent the number of options presented to the student on a machine scored assessment. If all assessment items with n choices are to have a single, arbitrary value, that value should be $\frac{1}{n}$.

4 The Three Rasch Paramaters - Mathematical Overview

The remainder of this lesson will incorporate some mathematics from late high school or early post-secondary, depending on the region. It can be omitted without reducing the accessibility of the conceptual aspects of future lessons.

With all combinations of parameters, items are modeled in terms of the probability P that a student of ability θ will correctly respond to the assessment item.

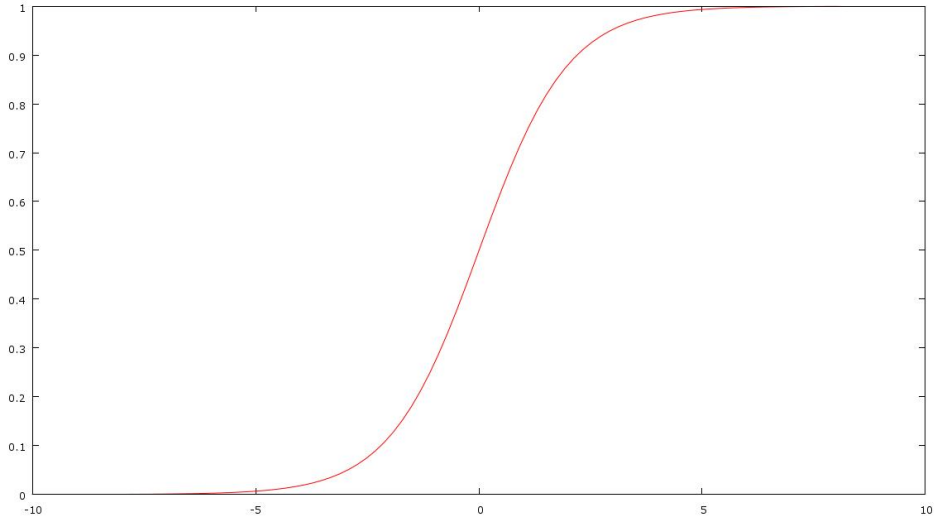

Figure 1: A single parameter Rasch model

4.1 Difficulty

If we label difficulty with variable b, then the formula describing the probability of a student of ability θ correctly responding to the assessment item is given by:

$$P(\theta) = \frac{e^{\theta - b}}{1 + e^{\theta - b}} \tag{1}$$

On mathematical scales, $0\% = 0$ and $100\% = 1$, such that a plot of such a graph for an assessment item of difficulty 0 would look like the graph seen in figure 1.

This graph is the item characteristic curve for the assessment item. In a model using only this single parameter, the only distinguishing features among the graphs would be their positions along the horizontal axis, such that three different item characteristic curves graphed side by side would look like figure 2.

This begs the question, how is the difficulty scale set? It seems odd to have 0 as the middle of the scale. There are methods to do this based on intrinsic data from the set and based on student ability levels, and these "natural" methods almost always have 0 in the middle of the scale. The ability scale here is the same as the scale score or standard score introduced in lesson four. It's arbitrary, and chosen by the entity creating the assessment. As the Rasch scale is an equal ability scale (meaning the difference in skill levels between scores 1.2 and 1.3

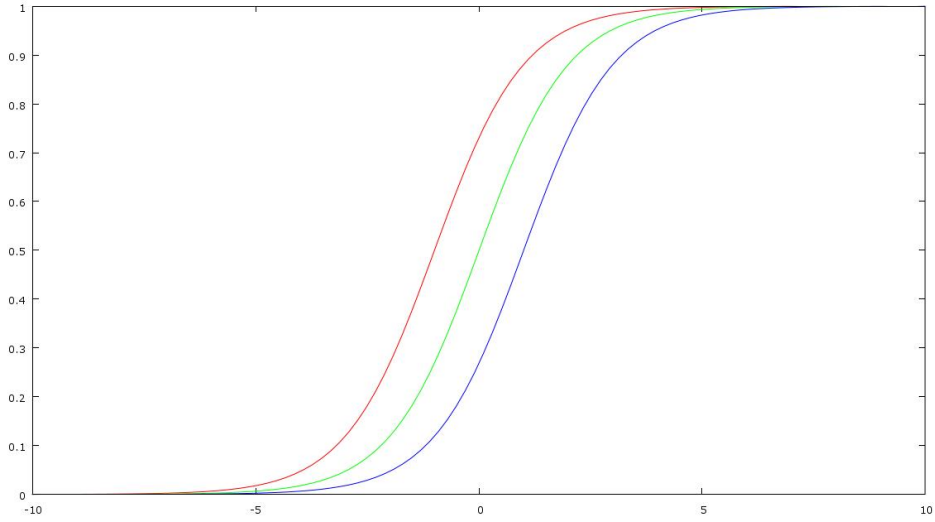

Figure 2: Three item characteristic curves modeled with a single parameter Rasch model. The only differences between the curves are their difficulties, which are the points at which the probability of answering correctly is exactly 0.5.

reflects the same difference in skill levels between scores 11.2 and 11.3), grade levels or grade equivalents are not appropriate scales to use, as the model fits will not conform to the actual student results. Typical practice is to fix two points on the scale at the extremes, with one point indicating an assessment item every student answered incorrectly (to define the upper limit of difficulty) and a second point indicating an assessment item every student answered correctly (to define the lower limit of difficulty.) If these points are set at 0 and 1000 (as is common practice) then many practical questions will fall in the 100-400 point range. Surprisingly, this range is not centered about 500: regardless of student ability, it is much more likely that assessments will include items every student will get right than items every student will get wrong. (i.e. the average student is more likely to approach the low end of the scale than the high end.)

4.2 Discrimination

The discrimination parameter a is added to equation 1 as follows:

$$P(\theta) = \frac{e^{a(\theta-b)}}{1+e^{a(\theta-b)}} \tag{2}$$

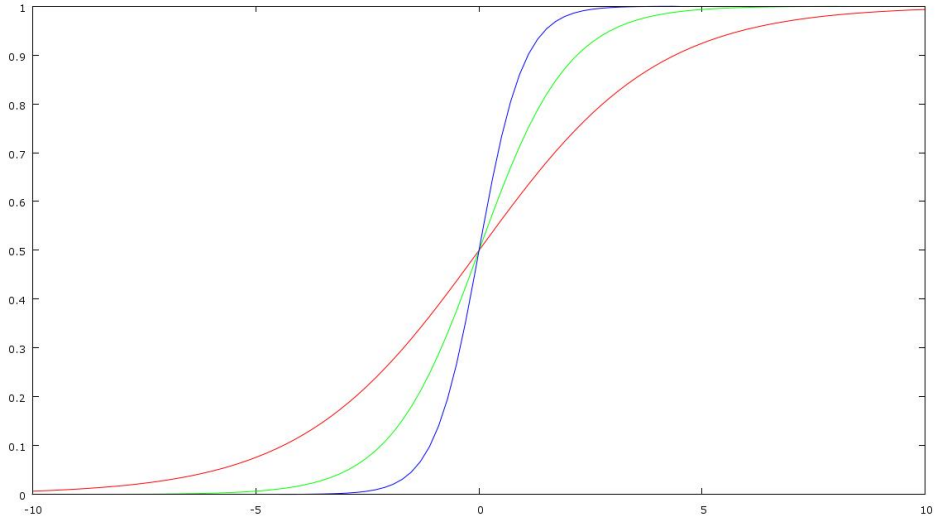

Figure 3: Three item characteristic curves with difficulty 0 but with three different discrimination values. (red: $a = 0.5$, green: $a = 1$, blue: $a = 2$)

For three item characteristic curves with the same difficulty (0) and three different discrimination values, see figure 3. Note that higher discrimination values lead to steeper item characteristic curves. These are less complex skills, which move from introduction to mastery in shorter periods of time.

4.3 Pseudo-chance Level

The pseudo-chance level parameter c is added to equation 2 as follows:

$$P(\theta) = c + (1 - c)\frac{e^{a(\theta - b)}}{1 + e^{a(\theta - b)}} \tag{3}$$

As mentioned in section 2.1, the introduction of this parameter alters the meaning of the difficulty parameter. In a one parameter fit, the probability of a student of ability b correctly answering the item is given by

$$P(b) = \frac{e^{b-b}}{1 + e^{b-b}} = \frac{e^0}{1 + e^0} = \frac{1}{1 + 1} = \frac{1}{2}$$

or 50%. Once the third parameter is introduced, this becomes

$$P(b) = c + (1 - c)\frac{e^{a(b-b)}}{1 + e^{a(b-b)}} = c + (1 - c)\frac{1}{2} = \frac{1 + c}{2}$$

which is the probability exactly half way between the probability c of guessing correctly with no knowledge of the skill and 100%. This is exactly half way between the lower and upper limits, which places the student half way between introduction and mastery of the skill.

Note also that many texts interpret the discrimination parameter directly in terms of the slope of the line. The definition used here, related to the time elapsed between introduction to and mastery of a skill, has two advantages:

1. It is easier to understand to a non-mathematical audience.
2. It does not require reinterpretation with the introduction of the third parameter. The higher the pseudo-chance parameter, the lower the slope of the line, but the difference in ability levels between when the line "looks" straight at the lower limit to the naked eye and when it "looks" straight at the upper limit remains unchanged.

Three different item characteristic curves with $a = 1$ and $b = 0$ are seen in figure 4. The red curve is a standard Rasch fit with $a = 1, b = c = 0$. The green is the same curve with $c = 0.25$, as it would be in most four question multiple choice assessment items. The blue curve shows $c = 0.5$, as it would be on a typical true-false type of question.

5 Upcoming Lessons

In lesson eight, we will discuss further analyses made possible by the Rasch models, and discuss how they pertain to computerized adaptive testing, while will become much, *much* more common in the future. In our final lesson, we discuss item response theory which does not involve any parameters.

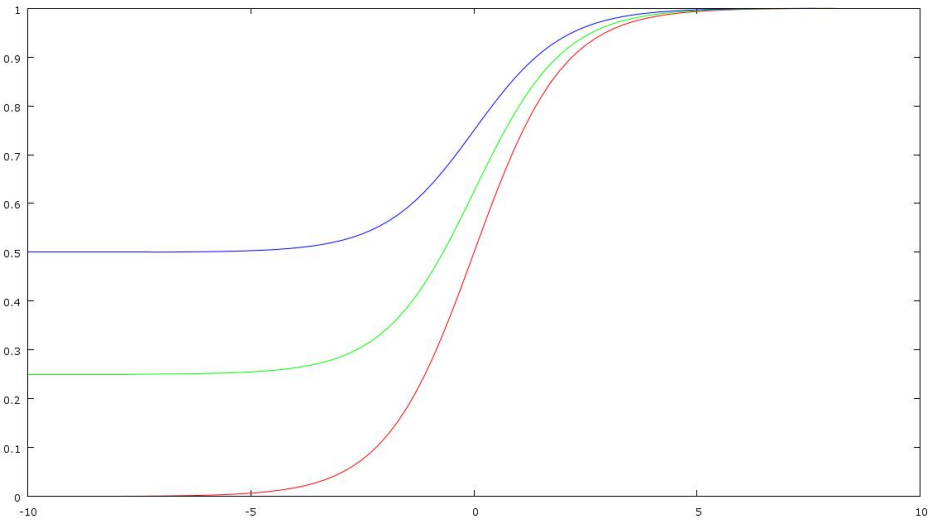

Figure 4: Three different item characteristic curves. The only difference in the parameters is found in the c parameters. (red: $c = 0$, green: $c = 0.25$, blue: $c = 0.5$)