Computerized Adaptive Testing

W. Blaine Dowler

May 28, 2011

1 Limitations of Traditional Theory

Assessments which do not use item response theory, introduced in the previous lesson, have a number of limitations, many of which have been alluded to, but which have not yet been collected into a single list. Here is that list:

- All items on a single assessment must be treated as having the same difficulty, the same discrimination, and the same bias in order to produce norm-referenced models of the assessment. This is remarkably unlikely.
- Evaluating the difficulty, discrimination and bias of test items on a pencil and paper test by traditional means is more indicative of the class population than of the properties of the test itself.
- It is impossible to evaluate the performance of a single student outside the context of his or her classmates.
- Reliability is difficult to measure precisely and in isolation for individual test items. Only its secondary effects on discrimination can be seen, and these can be obscured by skill complexity.
- Statistical uncertainties in student performance cannot be measured, or even estimated precisely.

It is very difficult and time consuming to apply item response theory to a teacher-created pencil and paper test. Doing so can improve the quality of items over time, but a considerable amount of time is required to collect enough statistics to make the results useful.

Item response theory was first applied to standardized pencil and paper tests written for specific grade levels. Rather than looking at the probability of getting

a particular item correct, the student's percentage correct on the assessment is evaluated instead. A particular mark was mapped to a mathematical scale, and that was reported as the student's ability. It suffers from the same problems as before, particularly when dealing with students who are working significantly above or below the test level. Computerized adaptive testing can overcome these problems quite nicely.

2 Computerized Adaptive Testing - the Concept

With a computerized adaptive test (CAT), each student in a group can receive an entirely *different* but still perfectly valid (and reliable) test. The first step is to develop a very, very large number of assessment items for a database. Then, test creators give the items on the CAT to a large number of students to define all relevant Rasch parameters for the items in question. At this point, the CAT is ready to be administered. When a new student starts the test, he or she is given an assessment item from the database of items which is appropriate to the student's age grade level. Before the second question is served to the student, the first question is marked. If the student responded correctly, the computer draws a more difficult item from the database. If the student responded incorrectly, the computer draws a less difficult item from the database.

The process continues until the student has at least one response correct and one response incorrect.[1] At this point, the computer uses statistical methods to take its best guess at the actual ability level of the student. It then serves up a question from the database which provides the most information about a student's estimated ability and administers that question. With this new information, it refines the estimate of the student's ability. As more questions are served, the student's ability is known with greater and greater precision. Ultimately, the CAT ends, either because a predetermined level of precision is reached, or because a set number of questions have been administered.

This is a fantastic tool for norm referenced testing. If a grade eleven student is reading at the grade four level, the results obtained from a pencil and paper test designed for grade eleven would be almost worthless, while a CAT would move down through the database to find the items at the student's current functioning level. As all students are measured relative to the same absolute score, the norm referenced results would be accurate regardless of the student's ability.

[1] If a student gets every item correct or every item incorrect, then an ambiguous case occurs. The way this is handled depends on the individual test. In any event, the student either passed unconditionally or failed quite spectacularly. Either way, his or her fate is quite sealed.

A CAT is not as easily applied to criterion referenced testing. Questions are rarely cross-correlated to curricular outcomes, so the criterion referencing is more difficult to manage. Though certainly possible, it is more likely that the CAT would be used as a "locator" test to determine a student's ability, and any criterion referenced assessment would be used as a secondary follow-up assessment at that level. This combination of assessments could then be used to produce a very effective curriculum customized to the needs of that particular student, which is a tremendous asset to facilities which can provide individualized instruction.

3 Computerized Adaptive Testing - The Math

There are three mathematical components to the above conceptual description.

1. Determining the amount of information provided by a particular assessment item. This step requires calculus.
2. Estimating the student's ability. This can be done using statistics, but it most commonly accomplished through calculus.
3. Determining the precision to which a student's ability is known. Although this is derived from statistics, the application requires algebra alone.

The equations generated are all relatively straightforward combinations of statistics and calculus. Diverting into their specific derivations would require spending an excessive amount of time and paper on material not directly related to assessment, so they will be presented without proof.

3.1 Information Functions

Assessment items only provide useful information if their rated difficulty is close to the student's ability. If an item is too easy or too difficult, learning that the student gets the item wrong or right tells us next to nothing. Therefore, it is not surprising that a mathematical formalization of the information function would depend on the same θ variable that the item characteristic curve $P(\theta)$ depends on. Regardless of the number of parameters involved in the Rasch fit, the information function $I(\theta)$ related to the item characteristic curve $P(\theta)$ is given by

$$I(\theta) = \frac{\left(\frac{dP(\theta)}{d\theta}\right)^2}{P(\theta)\left(1 - P(\theta)\right)} \tag{1}$$

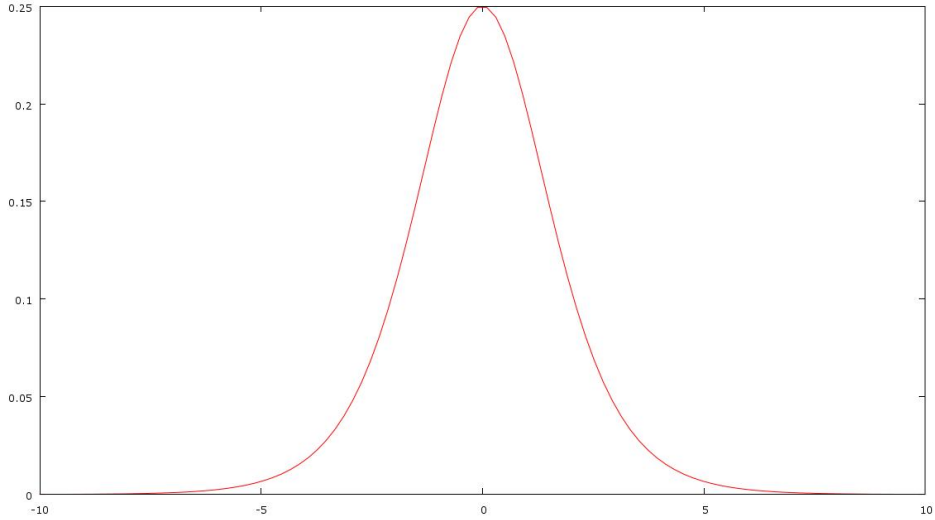

Figure 1: The information function corresponding to an item characteristic curve with $a = 1$ and $b = c = 0$.

where $\frac{dP(\theta)}{d\theta}$ is the usual derivative from calculus. For a full three parameter fit,

$$\frac{dP\left(\theta\right)}{d\theta} = a\left(1-c\right)\frac{e^{a(\theta-b)}}{\left(1+e^{a(\theta-b)}\right)^2}$$

If the fit has less than three parameters, then simply substitute the assumed values of a and c into the expression. For the special case with $a = 1$, $b = c = 0$, the information function can be seen in figure 1.

Although the information function and item characteristic curve have entirely different vertical scales and interpretations, they are both determined by the same set of Rasch parameters and θ. As such, it is useful to plot them on the same axes, despite the incomparable vertical scales. This has been done in figure 2. Notice that the information function peaks at $\theta = b$ and drops rapidly to each side. This is consistent with intuition: information is only garnered about a student's ability from an assessment item of similar difficulty.

Notice also that the information function depends on the square of the derivative of the item characteristic curve. Less discriminating items have shallower slopes, as do problems with high pseudo-chance levels. These problems provide less information about a student's ability, and should be avoided if the sole purpose of the assessment is to generate norm referenced results.

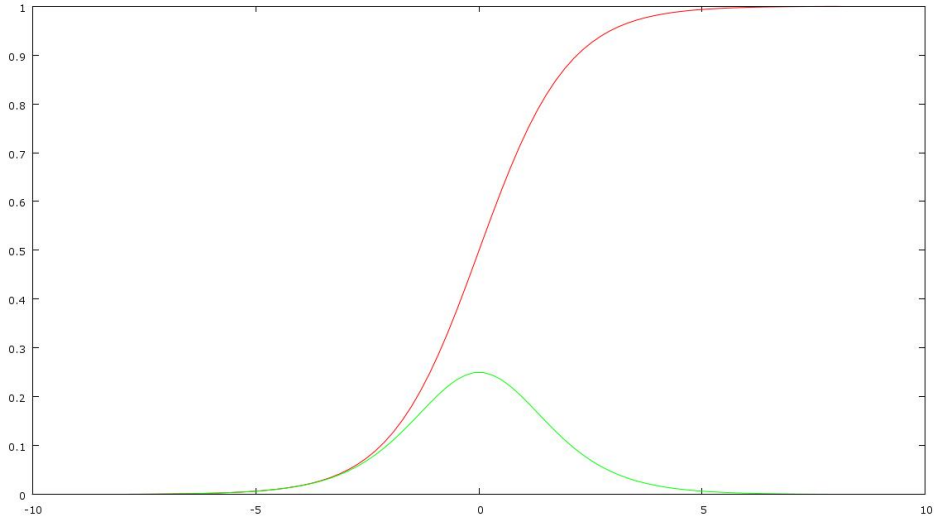

Figure 2: An item characteristic curve and its information function plotted on the same axes to compare their dependence on θ, despite the incompatibility of their vertical scales.

Finally, it is important to note that information functions add linearly when multiple assessment items are administered. This will be important when we discuss the calculation of uncertainties and standard errors two subsections from now.

3.2 Determining Student Ability

Determining a student's ability is the most mathematically involved portion of item response theory and computerized adaptive testing. Imagine the student has completed n assessment items thus far, and the response provided for assessment item j is represented by u_j, where $u_j = 1$ for a correct response and $u_j = 0$ for an incorrect response. The likelihood L that the student who has ability θ and has provided responses u_1 through u_n for the n assessment items answered thus far is given by

$$L\left(u_1, \ldots, u_n, \theta\right) = \prod_{j=1}^{n} P_j^{u_j} \left(1 - P_j\right)^{1-u_j} \tag{2}$$

where the θ dependence of $P(\theta)$ has been omitted to make subscripts and superscripts more clear, and where $\prod$ is the standard product notation.[2] The most likely ability level for a student to have is the maximum value of this function. The maximum value occurs where $\frac{dL}{d\theta} = 0$.

This is not a fun integral to evaluate, particularly given the rapidly changing nature of L with each completed assessment item. It can be simplified somewhat by noting that the maximum of L will correspond to the same θ as the maximum of $\log L$, which can turn the function being differentiated into a sum rather than product, which is computationally preferable. Still, numeric methods are used to find the interesting point. Things get even more complicated when a student's responses are inconsistent in a three parameter fit. Although impossible in a one or two parameter fit, a three parameter fit combined with student responses which include correct responses to relatively difficult questions and incorrect responses to relatively easy questions can result in a likelihood function which has a maximum at no finite value of θ. Some assessments use one or two parameter fits specifically to make this problematic situation mathematically impossible. Unfortunately, inconsistent results are something of a hallmark of students with learning disabilities,[3] so omitting the third parameter to ensure these situations are never represented means ensuring a group of students who most need a test which will adapt to student performance will not be accurately assessed by them.

This is the step which ensures application of item response theory to assessment will likely always be performed through the use of computer technology instead of by human hand. Determining which question a student needs to do next just takes too much time during a testing situation.

3.3 Computing Uncertainties

Now that we have a method by which to estimate a student's ability, we need to determine the accuracy of that estimate. The standard error $SE(\theta)$, or statistical uncertainty, in the measure of a student's ability as θ after administering

[2] If the reader is unfamiliar with $\prod$, it is to repeated, sequential multiplication what $\sum$ is to repeated, sequential addition. For example, $\prod_{j=1}^{5} j = 1 \times 2 \times 3 \times 4 \times 5 = 5! = 120$. One starts by substituting the $j = 1$ starting value that appears below the $\prod$ into the expression (j) to the right of the $\prod$ and evaluating that. Then you add 1 to the value to obtain $j = 2$, and substitute that into the expression to the right of $\prod$. This continues until you reach the value $j = 5$ listed above the $\prod$. At this point, you take the values of everything you calculated after each substitution (1, 2, 3, 4 and 5) and multiply them together.

[3] The author personally hates the term "learning disability." In his experience, the term "learning anomaly" is far more accurate.

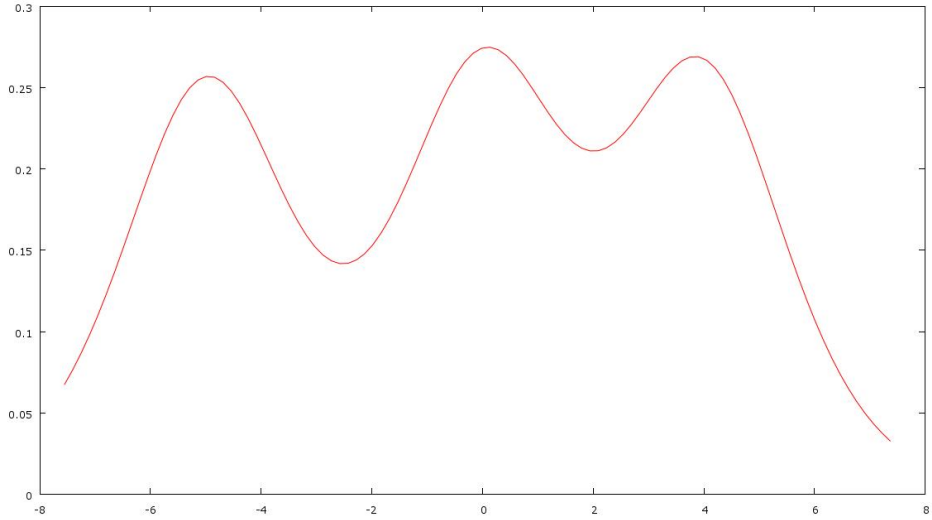

Figure 3: The combined information function after administering three assessment items of difficulties 0, 4 and -5.

n assessment items is given by

$$SE\left(\theta\right) = \frac{1}{\sqrt{\sum_{j=1}^{n} I_j\left(\theta\right)}} \tag{3}$$

where we have used the previously mentioned fact that information functions add linearly.

For example, assume we have administered three assessment items. For simplicity, assume $a = 1$ and $c = 0$ for all three. If the three items have difficulties of 0, 4 and -5 respectively, then their combined information function can be seen in figure 3. Notice how the items add nicely when the relative difficulty levels are close together. The standard error function for this information function can be seen in figure 4.

For a better comparison of the θ dependence of the two functions, they can be plotted against the same horizontal axis with differing vertical axes as seen in figure 5.

The fit is complete when the standard error at the anticipated value of θ has dropped below a specific threshold value. In many cases, particularly those using only single parameter Rasch fits, the assessment continues until a specific number of questions have been administered. Because all such information functions are identical apart from position on the θ axis, if the test administra-

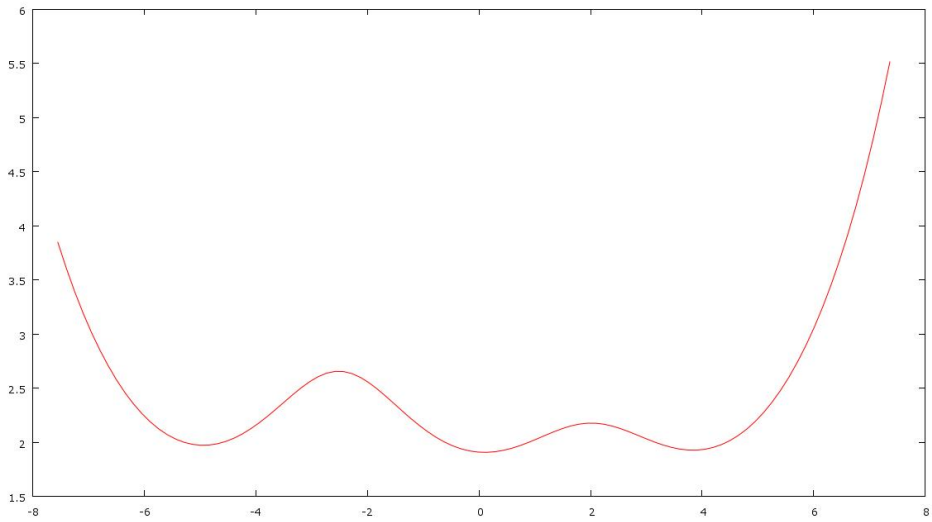

Figure 4: The standard error function corresponding to the information function in figure 3.

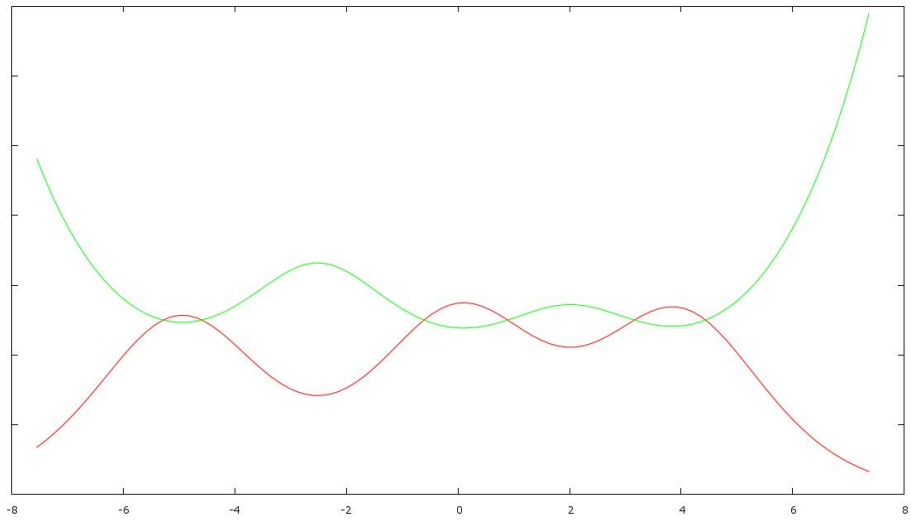

Figure 5: Information and standard error functions plotted against the same θ access with different vertical axes.

tors can accurately predict the rate at which the test will adapt to any given student's ability, then they can determine in advance how many questions are required to reach the required minimum standard error. This reduces the computational requirements during the test administration, as standard error and information functions would no longer need to be tracked. If one uses a more accurate two or three parameter fit, then the calculations must be done along the way.

4 The Upcoming Lesson

In our final lesson, we will describe methods of measuring student abilities with questions that do not fit into any parameterized model of any kind.