## Computerized Adaptive Testing: the easy part

If you are reading this in the 21st Century and are planning to launch a testing program, you probably aren't even considering a paper-based test as your primary strategy. And if you are computer-based, there is no reason to consider a fixed form as your primary strategy. A computer-administered and adaptive assessment will be more efficient, more informative, and generally more fun than a one-size-fits-all fixed form. With enough imagination and a strong measurement model, we can escape from the world of the basic, text-heavy, four- or five-foil, multiple-choice item. For the examinee, the test should be a challenging but winnable game. While we may say we prefer ones we can win all the time, the best games are those we win a little more than we lose.

If you live in my *SimCity* with infinite, calibrated item banks of equally valid and reliable items, people with known logit abilities, and responses from an unfeeling and impersonal random number generator, then *Computerized Adaptive Testing* (CAT) is not that hard. The challenge of CAT has very little to do with simple logistic models and much to do with logistics and validity. It has to do with how do you get the person and the computer to communicate, how do you ensure security, how do you avoid using the same items over and over, how do you cover the content mandated by the powers that be, how do you replenish and refresh the bank, how do you allow reviewing answered items, how do you use built-in tools like rulers, calculators, dictionaries, and spell checkers, how do you deal with aging hardware, computer crashes, hackers, limited band width, skeptical school boards, nervous teachers, angry parents, *gaming* examinees, attention-seeking legislators, or investigative "journalists." In short, how do you ensure a valid assessment for anyone and everyone?

I'm not going to help you with any of that. You should be reading van der Linden[1] and visiting the *International Association for Computerized Adaptive Testing*[2].

In my simulated world, an *infinite item bank* means I can always find exactly the item I need. *Equally valid* items means I can choose any item from the bank without worrying about how it fits into anybody's test blueprint. *Equally reliable* items means I can pick the next item based on its logit difficulty, not worry about maximizing any *information* function. Actually in my world of Rasch measurement, picking the next item based on its logit difficulty is the same as maximizing the information function. The standard approach is to administer and score an item, calculate the person's logit ability based on the items administered so far, retrieve and administer an item that matches the person's logit (and satisfies any content requirements and other constraints,) and repeat until some stopping rule is satisfied. The stopping rule can be that the standard error of measurement is sufficiently small, or the probability of a correct classification is sufficiently large, or you have run out of time, items, or patience.

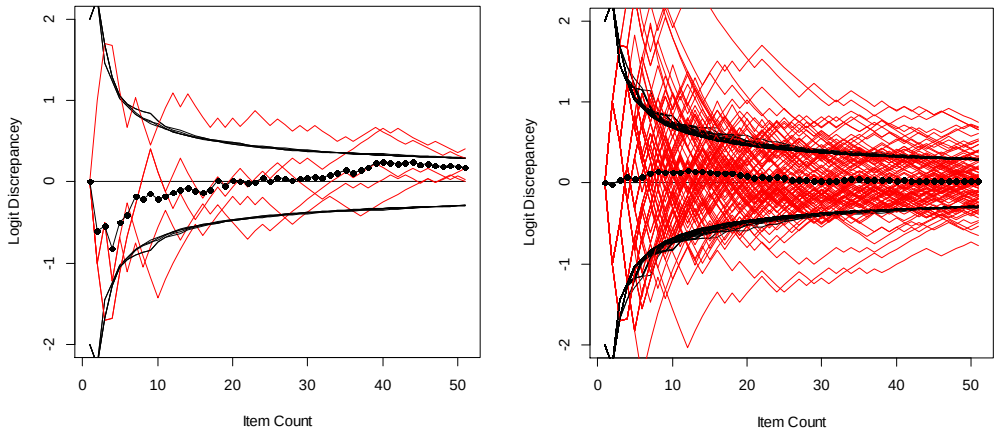
The process works on paper. The left chart shows the running estimates of ability (red lines) for five simulated people; the black curves are the running estimates of the standard error of measurement. The red lines should be between the black lines two thirds of the time. The black

---

[1] van der Linden, W. J. (2007). The shadow-test approach: A universal framework for implementing adaptive testing. In D. J. Weiss (Ed.), *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing*.

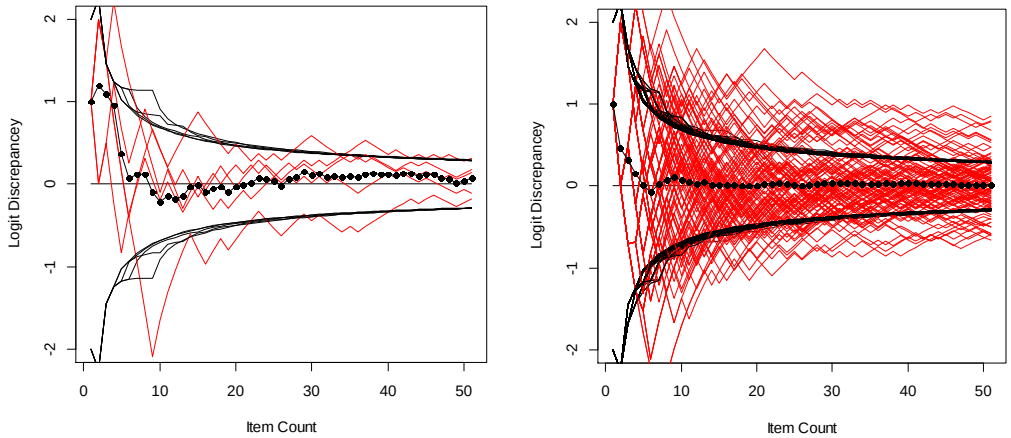
[2] www.iacat.org

dots are the means of the five examinees. The only stopping rule imposed here was 50 items. The right chart shows the same things for 100 simulated people.




With only five people, it's fairly easy to follow the path of any individual. They tend to vacillate dramatically at the start but most settle down between the standard error lines pretty much. Given the nature of the business in general, there will always be considerable variability in the estimated measures. With the 50 items that we ended on, the standard error of measurement will be roughly 0.3 logits (no lower than $2/\sqrt{50}$), which is hardly laser accuracy, but it is a reliability approaching 0.9 if you are that old school.

We started assuming a logit ability of zero, which is exactly on target and completely general because the items are relative to the person anyway. This may not seem quite fair because we are beginning right where we want to end up. But the first item will either be right or wrong so our next guess will be something different anyway. If we hadn't started right where we wanted to be, our first step will usually be toward where we should be. For example, if we start one logit away, we get pictures like these:




A curious artifact of this process is that if our starting guess is right, our second guess will be wrong. If our starting guess is wrong, we have a better than 50% chance of moving in the right

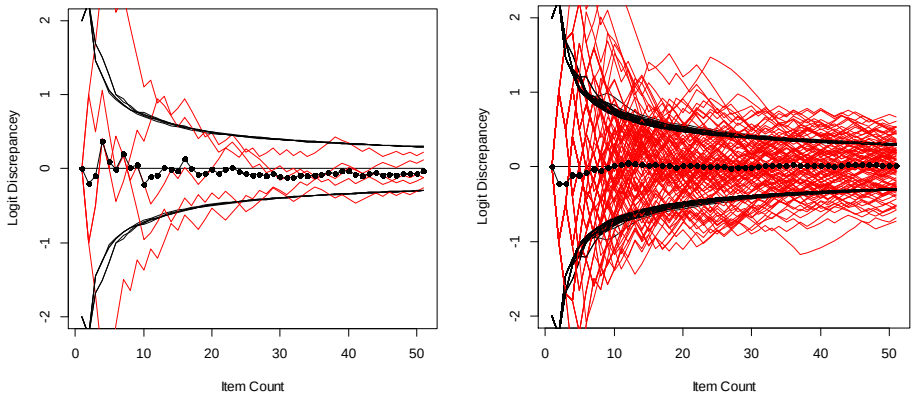
direction on our second guess; the further off we are, the more likely we are to move in the right direction. Maybe we should always begin off target.

Which says to me, when we are off by a logit in the starting location, it doesn't much matter. On average, it took 5 or 6 items to get on target, which causes one to wonder about the value of a five-item locator test, or maybe that's exactly what we have done. One implication of starting one logit high for a person is there is a good chance that the first four or five responses will be wrong, which may not be the best thing to do to a person's psyche at the outset.

The basic algorithm is choose the item for step $k+1$ such that $d^{[k+1]} = b^{[k]}$, where $b^{[k]}$ is the ability estimated from the difficulties of and responses to the first $k$ items. There is the start-up problem; we can't estimate an ability unless we have a number correct score $r$ greater than zero and less than $k$. I dealt with this by adjusting the previous difficulty by $\pm 1/\sqrt{k}$ while $r*(k-r) = 0$. One rationale for this is the adjustment is something a little less than half a standard error. Another rationale is that the first adjustment will be one logit and moving one logit changes a 50% probability of the response to about 75% (actually 73%). We made a guess at the person's location and observed a response. That response is more likely if assume the person is one logit away from the item rather that exactly equal to it. We're guessing anyway at this point.

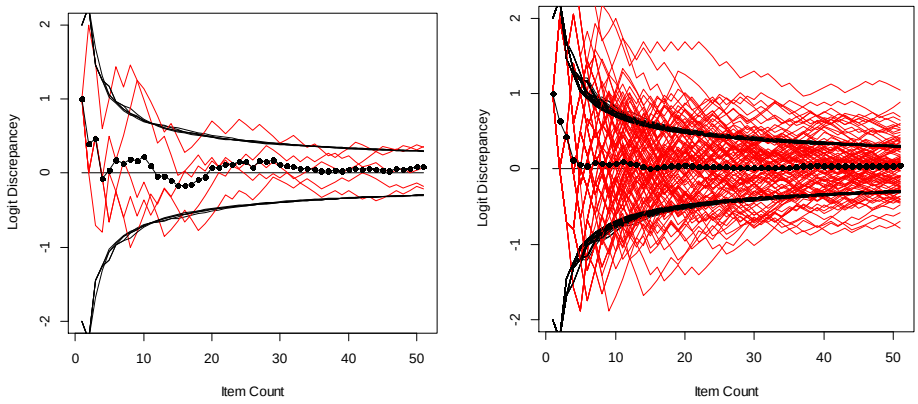
The standard logic, which we used in the simulations, seeks to maximize the information to be gained from the next item by picking the item for which we believe the person has a 50-50 chance of answering correctly. Alternatively, one might stick with the start-up strategy and look only at the most recent item, choosing a logit ability that makes the person's result on it *likely* by adjusting the difficulty of the chosen item without bothering with estimating the person's ability. The following charts adjust the difficulty by plus or minus one standard error, so that $d^{[k+1]} = d^{[k]} \pm s^{[k]}$, where $s^{[k]}$ is the standard error[3] of the logit ability estimate through step $k$.

First we tried it starting with a logit of zero:




Then we tried it starting with a logit of one:

[3] We are somewhat kidding ourselves when we say we didn't need to bother estimating the person's logit ability at every step of the way because we need that ability to calculate the standard error and check the stopping rule. We could approximate the standard error with $2/\sqrt{k}$ (or $1/\sqrt{k}$ or $2.5/\sqrt{k}$; nothing here suggests it matters very much) but that doesn't avoid the *when to stop* question.




The pictures for the two methods give the same impression. The results are too similar to cause anyone to pick one over the other and begin rewriting any CAT engines. Or to put it another way, these analyses are too crude to pick a winner or even know if it matters.

The viability of CAT in general and Rasch CAT in particular is sometimes debated on seemingly functional grounds that you need very large item banks to make it work. I don't buy it[4]. First, if your entire item bank consists of the items from one fixed form, the CAT version will never be worse than the fixed form and may be a little better; the worst that can happen is you administer the entire fixed form. You can do a better job of tailoring if you have the items from two or three fixed forms but we are still a long way from thousands. Second, with computer-generated items and item engineering templates coming of age, items can become far more plentiful and economical. We could even throw crowd sourcing item development into the mix.

Rasch has gotten some bad press in here because it is so demanding that it is harder to build huge banks; it requires us to discard or rewrite a lot more items. This is a good thing. A large bank of marginal items isn't going to help anyone[5]. The extra work up front should result in better measures, teach us something about the aspect we are after, and not fool us into thinking we have a bigger functional bank than we really do.

As with everything Rasch, the arithmetic is too simple to distract us for long from the bigger picture of defining better constructs and developing better items through technology. But that leaves us with plenty to do. Computer administration, in addition to helping us pick a more efficient next item, creates a whole new universe of possible item types beyond anything Terman (or Mead but maybe not Binet) could have envisioned and is much more exciting than minimizing the number of items administered.

The main barriers to the universal use of CAT have been hardware, misunderstanding, and politics. The hardware issue is fading fast or has morphed into how to manage all the hardware we have available. Misunderstanding and politics are harder to dismiss or even separate. Those aren't my purview or mission today. Well, maybe misunderstanding.

---

[4] I will concede a *very large* item bank is nice and desirable if it is filled with nice items.

[5] In its favor, any self-respecting *3pl* engine will try to avoid the marginal items but it would be better for everyone if they didn't get in the bank in the first place. It have never been explained to me why you would put the third (*guessing*) parameter in a CAT model, where we should steer clear of the asymptotes.