Computer-Administered Test to Learn

One of the political issues with computer administered tests (CAT) is what to do about examinees who want to revisit, review, and revise earlier responses. Examinees sometimes express frustration when they are not allowed to; psychometricians don't like the option being available because each item selection is based on previous successes and failures, so changing answers after moving on has the potential of upsetting the psychometric apple cart. One of our more diabolical thinkers has suggested that a clever examinee would intentionally miss several early items, thereby getting an easier test, and returning later to fix the intentionally incorrect responses, ensuring more correct answers and presumably a higher ability estimate. While this strategy could sometimes work in the examinee's favor (if receiving an incorrect estimate is actually in anyone's favor), it is somewhat limited because many right answers on a easy test is not necessarily better than fewer right answers on a difficult test and because a good CAT engine should recover from a bad start given the opportunity. While we might trust in CAT, we should still row away from the rocks.

The core issue for educational measurement is *test as contest* versus a useful self-assessment. When the assessments are infrequent and high stakes with potentially dire consequences for students, schools, districts, administrators, and teachers, there is little incentive not to look for a rumored edge whenever possible[1]. Frequent, low-stakes tests with immediate feedback could actually be valued and helpful to both students and teachers. There is research, for example, suggesting that taking a quiz is more effective for improved understanding and retention than rereading the material.

The issue of revisiting can be avoided, even with high stakes, if we don't let the examinee leave an item until the response is correct. First, present a multiple choice item (hopefully more creatively than putting a digitized image of a print item on a screen). If we get the right response, we say "Congratulations" or "Good work" and move on to the next item. If the response is incorrect, we give some kind of feedback, ranging from "Nope, what are you thinking?" to "Interesting but not what we're looking for" or perhaps some discussion of why it isn't what we're looking for (recommended). Then we re-present the item with the selected, incorrect foil omitted. Repeat. The last response from the examinee will always be the correct one, which might even be retained.

The examinee's score on the item is the number of distractors remaining when we finally get to the correct response[2]. Calibration of the thresholds can be quick and dirty. It is convenient for me here to use the "rating scale" form for the logit $[\beta_v - (\delta_i + \tau_{ij})]$. The highest threshold, associated with giving the correct response on the first attempt, is the same as the logit difficulty of the original multiple choice item, because that is exactly the situation we are in, and $\tau_{im} = 0$ for an item with m distractors (i.e., $m+1$ foils.) The logits for the other thresholds depend on the attractiveness of the distractors. (usually when written in this form, the τ_{ij} sum to zero but that's not helpful here.

To make things easy for myself, I will use a hypothetical example of a four-choice item with equally popular distractors. The difficulty of the item is captured in the δ_i and doesn't come into

[1] Admission, certifying, and licensing tests have other cares and concerns.

[2] We could give a maximum score of one for an immediate correct response and fractional values for the later stages, but using fractional scores would require slightly different machinery and have no effect on the measures.

the thresholds. Assuming an item with a p-value of *0.5* and equally attractive distractors, the incorrect responses will be spread across the three, with 17% on each. After one incorrect response, we expect the *typical* examinee to have a *[0.5 / (0.5+.017+0.17)] = 0.6* chance of success on the second try. A *0.6* chance of success corresponds to a logit difficulty *ln [(1 – 0.6) / 0.6] = –0.4.* Similarly for the third attempt, the probability of success is *[0.5 / (0.5+.017)] = 0.75* and the logit difficulty *ln [(1 – 0.75) / 0.75] = –1.1*. All of which gives us the three thresholds τ *= {-1.1, -0.4, 0.0}.*

This was easy because I assumed distractors that are equally attractive across the ability continuum; then the order in which they are eliminated doesn't matter in the arithmetic. With other patterns, it is more laborious but no more profound. If, for example, we have an item like:

1. *Litmus turns what color in acid?*
   A. *red*
   B. *blue*
   C. *black*
   D. *white,*

we could see probabilities across the foils like (*0.5, 0.4, 0.07, and 0.03*) for the *standard* examinee. There is one way to answer correctly on the first attempt and score 3; this is the original multiple choice item and the probability of this is still 0.5. There are, assuming we didn't succeed on the first attempt, three ways to score 2 (*ba, ca,* and *da*) that we would need to evaluate. And even more paths to scores of 1 or zero, which I'm not going to list.

Nor does it matter what p-value we start with, although the arithmetic would change. For example, reverting to equally attractive distractors, if we start with *p=0.75* instead of *0.5*, the chance of success on the second attempt is *0.78* and on the third is *0.875*. This leads to logit thresholds of *ln [(1 – 0.78) / 0.78] = –1.25*, and *ln [(1 – 0.875) / 0.875] = –1.95.* There is also a non-zero threshold for the first attempt of *ln [(1 – 0.7) / 0.7] = –0.85*. This is reverting to the "partial credit" form of the logit ($\beta_v - \delta_{ij}$). To compare to the earlier paragraph requires taking the *-0.85* out so that *(-0.85, -1.25, -1.95)* becomes *-0.85 + (0.0, -0.4, -1.1)* as before. I should note that this not the partial credit or rating scale model although a lot of the arithmetic turns out to be pretty much the same (see Linacre, 1991). It has been called "*Answer until Correct*;" or the *Failure* model because you keep going on the item until you succeed. This contrasts with the *Success* model[3] where you keep going until you fail. Or maybe I have the names reversed.

Because we don't let the examinee end on a wrong answer and we provide some feedback along the way, we are running a serious risk that the examinees could learn something during this process with feedback and second chances. This would violate an ancient tenet in assessment that the agent shalt not alter the object, although I'm not sure how the Quantum Mechanics folks feel about this.

[3] DBA, the *quiz show* model.