A Conceptual Overview of Computerized Adaptive Testing

Michael C. Edwards, PhD, Vector Psychometric Group, LLC, Chapel Hill, NC, USA, and Arizona State University, Tempe, AZ, USA, and **Carrie R. Houts, PhD**, Vector Psychometric Group, LLC, Chapel Hill, NC, USA

Michael C. Edwards, PhD

KEY POINTS . . .

Computerized adaptive testing is an advanced method of item selection using item response theory scoring technology.

Computerized adaptive testing, while using complex statistical theory and computations, is based on straightforward and intuitive ideas.

The use of computerized adaptive testing can reduce respondent burden in clinical outcome assessment without any loss of score precision.

This is the second of two articles in this issue of the topic of algorithmic advances in HEOR. Dr. Edwards and Dr. Houts highlight algorithmic advances in our field on computerized adaptive testing in clinical outcomes assessment.

Computerized adaptive testing, known commonly by the acronym CAT, refers to a collection of systems and statistical models that assemble an assessment in real time based on the observed item responses of each individual. This is quite different from standard assessments, many of which are still administered in a paper and pencil (P&P) format and have what are known as "static" forms (i.e., the assessment is fixed and doesn't change). While CAT has been showing up in the clinical outcome assessment (COA) literature for some time, it is still an under-utilized tool in real-world assessments in clinical trials. Although there are a variety of reasons for this, we suspect that it is in part due to a lack of comfort and familiarity with the technology. The goal of this article is to explore how CAT works, as well as why (and when) it is advantageous.

How does CAT work?

As the name implies, there are three things to consider when thinking about CAT: tests, computers, and adaptation. We'll begin with tests. Although the word "test" is often associated with educational assessment, it is meant as a generic term and covers the kinds of things seen in education, as well as the assessments used in the health domain (e.g., COAs).

Practically speaking, adaptive assessments require computers to be administered, so their computerized administration is usually a by-product of the desire to have an adaptive test as opposed to a driving design feature. However, computer-based assessment has a number of considerable advantages over the P&P mode of delivery. First, the data are captured directly into a database and do not require any manual data entry, which can be slow, expensive, and error prone. Second, with certain designs and systems it is possible to capture additional information about how the respondent behaved. For example, if there is one item per screen you can capture how long the respondent took for each page. While this is not exactly a response time, it does allow you to: 1) know how long respondents took, on average, to finish the assessment and 2) screen for individuals who completed it much more quickly (suggesting they weren't really paying attention), or much more slowly (suggesting some possible comprehension issues). Third, if you are using a computer to administer an assessment, the flexibility and power of the computer is at your disposal. This enables you to be much more creative with your "items". As an example, imagine that you are trying to measure how satisfied someone is with his or her vision. Rather than asking if he or she has trouble reading small print, you could actively manipulate the print size on the screen and have the participant choose the smallest one that he or she can see. Such flexibility could provide a real benefit when it comes to making validity arguments.

> "CAT can be used to reduce respondent burden in COA without any loss of score precision or validity of inferences."

While the adaptive feature of CAT is implemented by using complicated statistics, it is based on relatively straightforward ideas, many of which have analogs that are familiar to our readers. One critical idea in CAT is the item bank. This is a collection of items that could be used to assess a particular construct, but at the outset in a CAT environment our hope is to choose a subset (often very small) of the possible items to assess the construct in question. Again, while the specific way this is accomplished can be a bit complex, the idea is one nearly everyone has seen in practice. Many scales, when they were initially constructed, had a (relatively) large number of items. In many cases, users wanted to shorten these scales and selected a subset of the initial items to create a "short form." This is actually the same

>

basic idea of adaptive testing, just applied in a more static manner. Let's imagine there is a 30-item quality-of-life (QoL) measure that has been developed and shown to be useful in practice. However, ideally, researchers would like a 10-item measure as they are trying to balance reliability and respondent burden. How do we choose the best 10 items from the set of 30? You can think about the full set of 30 as an item bank. These are all items that could be used to assess QoL, but in this case you want to choose a subset of 10. In many cases, we choose the 10 items to keep reliability of the resulting score as high as possible. This is very similar to an item-selection strategy called "maximizing Fisher information" in the adaptive assessment world. The key difference between making a short form and an adaptive test is that the short form is created once and the same form is presented to everyone. In an adaptive environment, the hope is that it is possible to create numerous "person-specific short forms" which all achieve a targeted level of reliability with as few items as possible.

So why would you want short forms that are different for each person? It turns out that targeting the assessment to the person's level of the construct is the most effective way psychometricians have found to maximize reliability. Intuitively this makes sense. If you imagine that 10 of our 30 QoL items measure low levels of QoL, 10 items measure moderate levels, and 10 items measure high levels, you can start to see that you may be able to save some time by not giving someone all the items. For example, if someone is relatively healthy, you probably won't get any useful information from asking them items that measure very low levels of QoL.

Let's imagine an (absurd) example called the Monotonous Quality of Life Scale (MQLS). The MQLS consists of the following ten items:

1. My quality of life is at least a 1 on a 1 to 10 scale.
2. My quality of life is at least a 2 on a 1 to 10 scale.
3. My quality of life is at least a 3 on a 1 to 10 scale.
4. My quality of life is at least a 4 on a 1 to 10 scale.
5. My quality of life is at least a 5 on a 1 to 10 scale.
6. My quality of life is at least a 6 on a 1 to 10 scale.
7. My quality of life is at least a 7 on a 1 to 10 scale.
8. My quality of life is at least an 8 on a 1 to 10 scale.
9. My quality of life is at least a 9 on a 1 to 10 scale.
10. My quality of life is at least a 10 on a 1 to 10 scale.

Respondents are instructed that 1 is bad and 10 is good and each item has a dichotomous yes/no response format. Someone who has a QoL of 2 will say "yes" to the first two and then "no" to the remaining eight items. While the "no" response to the third question is useful, the additional "no" responses aren't actually providing new information. Similarly, if someone stops saying "yes" at Item 8, the "yes" responses to Items 1 to 6 don't provide information above and beyond the "yes" to Item 8.

In this example, all respondents see the same ten items, but only one or two of the responses are actually informative. In an adaptive context, we would generally start in the middle (Item 5). If the respondent says "yes" to Item 5, then we would ask him or her Item 6. If the response is "no", then we would ask Item 4. If we don't know anything in advance about the respondent, this is the fastest way to "zero in" on his or her level of QoL while asking the fewest questions possible.

The MQLS scale is obviously an over-simplified example, but hopefully it has shown some of the basic intuitions necessary to understand CAT. In a CAT environment, we choose items based on their severity and our current best estimate of the respondent's level of the construct being measured. This means that, if measuring anxiety, you begin with some questions that indicate moderate levels of anxiety and then progress from there. If the person endorses those, you don't gain any information from asking lower severity questions, so you move to higher severity items to figure out where on the anxiety spectrum he or she lies.

When moving to a CAT system from the static assessments most people are familiar with, one big change is that items can no longer simply be added up to create a scale score. Because different respondents will see different items, it is necessary to weight items to create comparable scores. Fortunately, the technology for doing this (item response theory [IRT]) has been around for a long time, and we have decades of experience (with millions of assessments) making this all work smoothly in practice. Indeed, despite the complex calculations taking place behind the scenes, the technology for CAT— and IRT scoring in general — is such that regardless of the system or software being used, the obtained scores should the same. That is, item responses and item parameters (e.g., item severity values mentioned earlier) are simply numbers plugged into fixed equations and (assuming similar technical settings across programs) the scores that come out are solely determined by the numbers that went in. That is, whether scores are from proprietary systems for assessments such as the NIH-sponsored Patient Reported Outcomes Measurement Information System (PROMIS), the GRE, or various employment screeners, from independent CAT packages, stand-alone IRT software, or from self-written code, given the same data and item information, the scores should all match.

Why is CAT useful?

Stated simply, adaptive testing provides the most efficient (i.e., fewest items) assessment to achieve a targeted level of score reliability for all respondents. CAT can be used to reduce respondent burden in COA without any loss of score precision or validity of inferences. This means that respondents will see the fewest items possible, which allows for additional data collection or the minimization of assessment times. The latter is valuable situationally when dealing with, for example, individuals experiencing a great deal of pain. The minimization of assessment time is also globally valuable in a world where we are increasingly being surveyed and suffering from "survey fatigue." CAT is also useful in that patients are not presented with numerous questions which would be perceived as not relevant to their situation. CAT provides a way to minimize the burden on respondents while maintaining the quality of the data being collected. It is our hope that as people become more familiar with CAT and its underlying principles they will take advantage of this technology, which can improve measurement from both the patient and scientific perspectives. ■