SIGNANT HEALTH
THE EVIDENCE GENERATION COMPANY™

The Signal

Computerized Adaptive Tests: A Primer

Most patient-reported outcome measures (PROMs) are composed of a fixed number of items – that is, a question and its associated response scale. While they may be simple to administer, it is thought that some PROMs may lead to increased patient burden due to the quantity of questions required to be answered, especially when patients are asked to compete multiple PROMs together.

Computerized adaptive tests (CATs) differ in that these use a variable number of items picked from a superset of items (an item bank). The item bank typically provides information across the full range of severities of a particular concept of interest, for example, physical function. Each subsequent item presented depends upon the responses provided to all the previous items. In this way, the response to each subsequent item hones in, with increasing confidence, to the final estimate (score) of the concept of interest. When sufficient confidence in the reliability of the estimate is achieved, the test can stop, and no further questions are required.

CATs offer the potential to provide sufficiently precise estimates (scores) of the concept of interest while delivering fewer items, and this may benefit patients, especially in reducing burden when frequent assessment is required, or when multiple PROMs are administered together.

In this blog, we look a little deeper into how CATs work, explore their performance compared to static PROMs, and explore regulatory opinion on the use of CATs.

How CATs Work

As described above, CATs operate quite differently from static list PROMs. Static list PROMs, where participants respond to every item within the measure (and rules for missing responses to be imputed are defined), represent measures constructed using classical test theory (CTT). Most of the PROMs we see in clinical research fall into this category. With CTT, the overall measure provided by a PROM might be calculated, for example, as a total sum of the scores of all the individual items.

However, there are a number of CATs that are of interest to clinical trials, in particular those developed using the PROMIS item bank (PROMIS-CAT – Northwestern University, Evanston, IL, USA). CATs use Item Response Theory (IRT) as opposed to CTT. Measures/scores delivered by an IRT model are estimated based on a probability model that identifies the most likely score

for the concept of interest, given the items to which the patient has already responded.

Without going into the statistics, in essence, an IRT model works in the following way – in this example let's consider estimation of physical functioning:

1. Select and deliver the first item to be administered. Typically, this will be an item in the middle of the range for the attribute being measured, in this case physical function.
2. The probability model is used to estimate the overall score for physical function.
3. The model is then used to select the next item from the item bank considered optimal to refine the estimated score.
4. The model is then used again to recalculate the score, based on the responses to the two items.
5. This process continues to administer items until the model's stopping rules are met. Stopping rules vary, but, as an example, some of the PROMIS-CAT measures stop estimating when a maximum number of items have been administered (n=12), or when the standard error of the estimate falls below a defined threshold (ideal stopping condition).

CATs require interaction with the underlying IRT model throughout the testing process, and so they are typically conducted using solutions with direct web connectivity to the IRT model.

While the complexity of CATs may require careful consideration, they promise some attractive properties:

- They may require fewer items to be answered, and so reduce patient burden associated with PROM completion.
- They may provide estimates of the attribute of interest that are more precise than obtained using CTT.
- The items presented to each patient are tailored based on their responses, and so there may be less chance of asking questions that do not seem relevant to each individual patient.

CAT performance

There are a number of studies that explore the performance of CATs in comparison to static-list PROMs. One example used simulation to explore the accuracy and number of items administered using the PROMIS CAT in comparison to the 4-, 6-, and 8-item PROMIS short forms for several domains: physical function, anxiety, depression, fatigue, sleep disturbance, social function, and pain interference [1].

Number of items

This study [1] found that the average number of items administered in CAT was 4.7 across all domains, showing a reduction in the number of items delivered in comparison to the 6-item and 8-item short forms.

Accuracy

The simulation study indicated that CAT administration generated higher percentages of accurate scores in comparison to the short form versions [1]. Further, PROMIS-CAT correlated with a wider accurate range compared to the short forms. This was manifested by better performance at the very poor health and very good health ends of the range of severity for each domain – for example, showing much lower percentages of scores at the floor and ceiling of each scale. However, most short forms examined, especially the 8-item versions, provided reasonably wide accurate range.

In summary, this study is helpful to illustrate the potential to realize some of the benefits of CATs in reduced numbers of questions and improved accuracy properties.

Regulatory view

There has been little published opinion on the use of CATs by the regulatory bodies. However, recently FDA included a section on CATs in the third section of their patient-focused drug development draft guidance series [2], and this is perhaps the most information we have on the agency's current opinion.

While much of the validity work needed for a CAT is common to that of a static PROM, FDA point out that there are additional evidentiary requirements needed to demonstrate that the IRT model is well fitting, and that (for example) changes in patient's scores over time are due to true changes in the underlying concept of interest, and not as a result of the application of a different set of items.

In the draft guidance, FDA also state they recommend not making changes to an item bank mid-trial. If this is required, it will be important that CAT authors can demonstrate that the item bank remains well calibrated with respect to the original concept being measured.

FDA also state that sponsors must be able to describe and justify the stopping rules used for the CAT in terms of the minimum level of measurement precision sought. Stopping rules, in the opinion of the agency, should also cap the total number of items to be administered, to ensure patient burden is considered.

Conclusions

There is interest in the use of CATs to improve precision of measures, to reduce the number of items administered to patients, and to eliminate the delivery of less relevant questions / items. There is some evidence in the literature that CATs can deliver these enhancements. There are additional demands in terms of evidentiary requirements to support the use of a CAT in regulatory submissions, including data to support the goodness-of-fit of the IRT model, and to defend the accuracy of scores returned when different sets of items are used in their calculation.

As our industry considers the role of CATs in clinical trials, perhaps the most valuable areas for consideration are those requiring frequent assessments, or requiring lengthy

combinations of PROMs to be delivered concurrently, or in diseases where the burden of completion is of highest concern for other reasons, such as treatment- and disease-related symptoms and side effects.

References

1. Segawa E, Schalet B, Cella D. A comparison of computer adaptive tests (CATs) and short forms in terms of accuracy and number of items administrated using PROMIS profile. Qual Life Res. 2020; 29(1): 213-221.

2. FDA. Patient-Focused Drug Development: Selecting, Developing, or Modifying Fit-for-Purpose Clinical Outcome Assessments: Draft Guidance for Industry, Food and Drug Administration Staff, and Other Stakeholders. June 2022. Patient-Focused Drug Development: Selecting, Developing, or Modifying Fit-for-Purpose Clinical Outcome Assessments (fda.gov)

Bill Byrom, Ph.D.

VP Product Intelligence and Positioning

Richard Sutherland

Product Manager