Advanced Scientific Capabilities |
Patient-Reported Outcomes and Assessment

Use of static and flexible PRO measures in global cancer trials: Challenges and opportunities

February 29th, 2024

Jim Shaw, PhD, PharmD, MPH
Head, Patient-Reported Outcomes Assessment

Bristol Myers Squibb™

Disclosures

Employment
and stock ownership with
Bristol Myers Squibb

Outline

01

Static vs. flexible: Definitions and examples

02

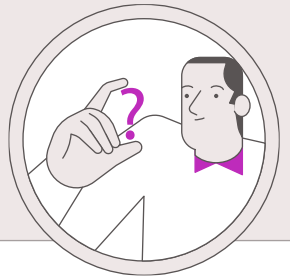
BMS experience with flexible measures

03

Comparative benefits of static and flexible measures

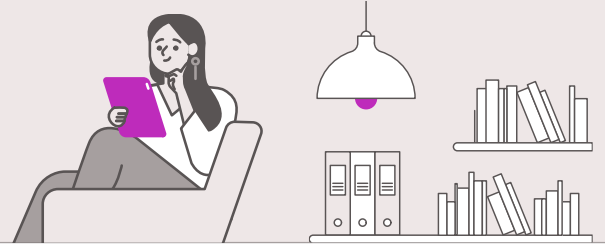
Static vs. flexible: Definitions and examples

Static questionnaires



- Fixed set of items intended to measure a specific concept or group of related concepts
  - The entire set of questions are administered, and the content and order of the items typically does not vary
  - Some questionnaires may incorporate conditional branching logic that allows skipping or branching of questions
- Can be generic or disease-condition specific
- Can be administered across a wide range of modalities
- Psychometrically validated to generate valid and reliable scores using classical test theory
- Examples: FACT-G, EORTC QLQ-C30, EQ-5D-5L

Item pools, banks, and libraries



Item Libraries/Item Pools:

- Collections of single items or multi-item scales that measure HRQoL domains.
- In contrast to static questionnaires, researchers can select specific items (or groups of items) from the library to measure on relevant PRO domains for a given context or target population.

Examples: PRO-CTCAE, EORTC Item Library

Item Bank:

- A special case of item libraries in that all the items included for each HRQoL domain have been calibrated with an item response theory model.
- Item banks allow investigators to generate multiple short forms from the same item bank, and they allow for CAT, which tailors the PRO measures on the basis of how a patient answers each item.

Example: PROMIS Item Banks

Piccinin et al,. Recommendations on the use of item libraries for patient-reported outcome measurement in oncology trials: findings from an international, multidisciplinary working group. Lancet Oncol 2023; 24: e-86-95

Short-Forms

- When bank items are analyzed using IRT, each item that fits the IRT model is calibrated on the trait based on the location of the items on the measurement continuum

- For individual measurement, precision is calculated for each unique level along that continuum.
  - When the IRT measurement model fits the item bank data, one can select any subset of questions in that bank. Since items in the bank are calibrated onto the same continuum, the scores obtained from the derived short-forms are comparable to that from the complete bank
  - In contrast, classical test theory typically requires that an entire test be administered to appropriately represent the concept being measured and a single precision is reported across the whole continuum

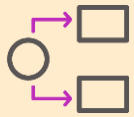
- Example: PROMIS Short-Form, PROMIS Profile Measures

Cella D, Gershon R, Jin-Shei L, Choi S, The Future of Outcomes Measurement: Item Banking, Tailored Short-Forms, and Computer Adaptive Assessment, Qual Life Res 2007: 16;133-141

Computer adaptive testing

- CAT is a specific type of computer-based testing that has seen popularity in certification, licensure and educational testing.
- Computer-based assessment offers such advantages as immediate data entry; ease of scoring; and almost immediate plotting of results and/or changes over time.
- Using IRT measurement models, item selection is guided by an individual's response to previously administered questions from a large item bank. The respondent need only answer a small number of informative items to accurately estimate what would have been obtained had the entire set of items been administered
- Example: PROMIS CAT

Cella D, Gershon R, Jin-Shei L, Choi S, The Future of Outcomes Measurement: Item Banking, Tailored Short-Forms, and Computer Adaptive Assessment, Qual Life Res 2007: 16;133-141

	Static	Customizable	Dynamic
Flexibility 	• All items from instrument need to be completed • Examples: — EORTC QLQ-C30 — FACT-G	• Can select a subset of items • Examples: — PRO-CTCAE — PROMIS Custom Short-Forms	• Computer software selects the items • Examples: — PROMIS CAT measures
Scoring 	• Generates scale and/ or subscale scores according to instrument scoring rules/ procedures	• Scoring at either the individual item level or scale level (if sufficient items from a given scale are co-administered)	• Trait scoring using IRT (requires computer to calculate score)

BMS experiences with flexible assessment: PRO-CTCAE

PRO-CTCAE

NCI Patient Reported Outcomes version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE®) Measurement System

- Designed to evaluate symptomatic toxicities by self-report in adults, adolescents and children in cancer trials
- The adult item library holds 124 items covering 78 symptoms
  - Items cover frequency, severity, interference and presence/absence of symptom
- The library provides flexibility in two ways:
  - Only items for relevant symptoms are selected for a particular use
  - The majority of symptoms have more than one item but conditional branching allows for only the relevant items to be asked to a particular patient
    - Example: A symptom may have a frequency and a severity item, but the severity item doesn't get asked if the patient reports in the frequency item that they never have the symptom

Use of PRO-CTCAE at BMS

BMS was an early adopter, with use in clinical trials as early as 2016

Fourteen hematology/oncology trials including 6 Phase 3s

Experience with creating de novo items

In the last 7 days, what was the **severity** of the **swelling or puffiness around your eyes** at its **worst**?				
○ *None*	○ *Mild*	○ *Moderate*	○ *Severe*	○ *Very severe*
In the last 7 days, how much did the **swelling or puffiness around your eyes interfere** with your usual or daily activities?				
○ *Not at all*	○ *A little bit*	○ *Somewhat*	○ *Quite a bit*	○ *Very much*

Learnings from our experience

Strategy

- Clinical teams are often receptive to using the PRO-CTCAE, especially in Phase 2 trials, but costs and patient burden are a concern, especially because the measure should be done frequently to capture symptomatic toxicities and requires eCOA
- While the FDA has recommended their use for dose selection, there is currently no regulatory incentive to do so nor is there a clear model on how to incorporate them in decision-making
- The appropriate/unbiased selection of PRO-CTCAE items is an ongoing concern

Analytical

- Analyzing PRO CTCAE data can be a challenge due to varying outcomes, response metrics, and branching logic. There is a need to educate stats personnel on analysis and other stakeholders on interpretation of results

Learnings from our experience

Operational

Early implementations involved paper use, which can cause problems with patients answering questions that aren't relevant to them

eCOA implementation is now standard but makes the ability to capture other symptoms – Beyond the items chosen – Difficult as free text keyboards can be cumbersome or unavailable on eCOA devices in some languages

eCOA vendors have differing levels of familiarity with the measure and enacting conditional branching correctly is a challenge for some

Comparative benefits of static and flexible measures

Static measures

Easier to interpret and communicate results to internal and external stakeholders

Greater comparability of results with those achieved in competitor trials (excepting flexible measures that use IRT-based scoring)

Potential efficiencies had in developing eCOA builds

HTA bodies demand “validated” tools. The issue of validation becomes contentious when considering the use of instruments in part or batteries varying in content across trial participants (e.g., CAT)

Regulatory perspective/guidance on static measures is clearer

Not all eCOA suppliers have the capability to implement flexible measures. Impact of this depends on sponsor flexibility to employ different suppliers

Needed translations may not exist for all item bank/library items, and processes for deriving these may be more convoluted and/or time consuming than developing translations for static forms

Provenance of bank items may lead to questions of content validity (e.g., “Are you able to pass a 20-pound (10 kg) turkey or ham to other people at the table?”). Nevertheless, one can also question the content validity of static forms given their development prior to the era of modern cancer therapies

IRT-based scoring of flexible measures may be a challenge if sponsor is unable to run trial data through external systems

Flexible measures

Tailor item content to specific trial and item difficulty to respondent capability and understanding

Potentially reduced respondent burden and administration time, which may impact data completeness, participant retention, and site enrollment

IRT-based scoring provides greater interpretability

CAT provides increased score precision across continuum of respondent ability, arguably making assessments fairer for high- and low-ability respondents

Conclusions

- Flexibility should be viewed as a continuum
- With flexible assessment, item content can be tailored to trial specifics or respondent ability. However, there are numerous barriers to using flexible measures that need to be weighed against potential benefits
- The application of flexible measures requires the support of appropriate sponsor roles, processes, and platforms as well as supplier technology
- Insufficient regulatory guidance and precedents as well as HTA concerns need to be addressed

Questions