Using large banks of items and computer adaptive testing to improve clinical outcome assessment

Nina R. Schooler, PhD

SUNY Downstate Health Sciences Center

Brooklyn, NY

nina.schooler@gmail.com

Outline of the Presentation

- What is Computerized Adaptive Testing (CAT)?
  - Taking the same test twice
- Examples of CAT programs
  - NIH – PROMIS initiative for Patient Reported Outcomes
  - Adaptive Testing Technologies
    - CAT-MH battery of of adaptive tests for psychiatry
- Example of an Adaptive Test for Psychosis Severity
- Application of Adaptive Testing in Drug Development
  - FDA guidance
- Developing Adaptive Tests for psychopathology symptoms
  -

PROMIS
Computerized Adaptive Test for pain behavior
In the past 7 days …

- When I was in pain, I moved extremely slowly
  - Often
- I had pain so bad it made me cry
  - Rarely
- Pain caused me to curl up in a ball
  - Never
- When I was in pain it showed on my face (squinching eyes, frowning)
  - Sometimes
- Score 60 and worse than
  - 82% of general population
  - 79% of females

PROMIS
Computerized Adaptive Test for pain behavior

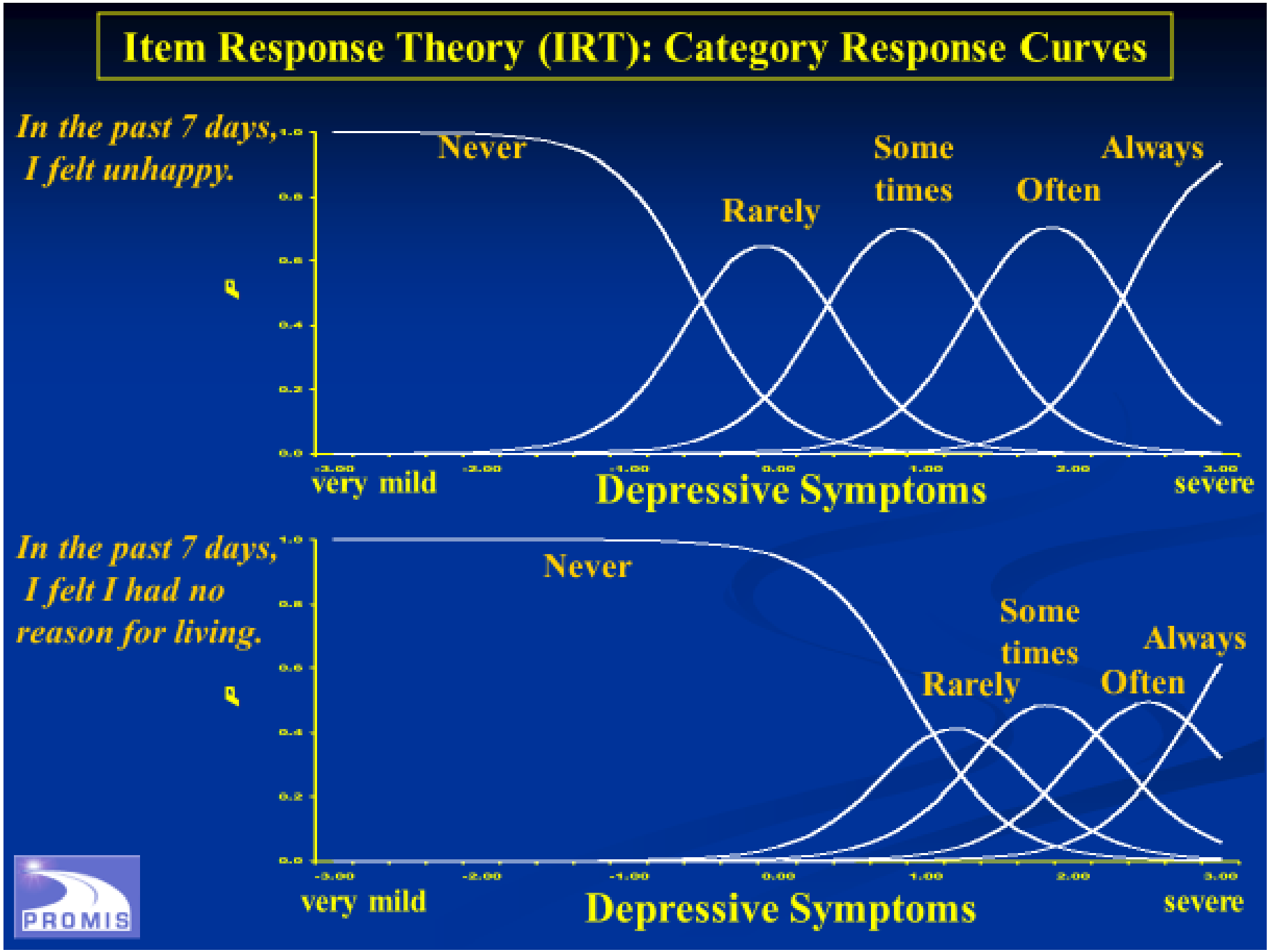
In the past 7 days …

- When I was in pain, I moved extremely slowly
  - Often
- I had pain so bad it made me cry
  - Sometimes
- Pain caused me to curl up in a ball
  - Sometimes
- When I was in pain, I appeared sad or upset
  - Sometimes
- Score 63 and worse than
  - 92% of general population
  - 81% of females

NIH – PROMIS – Patient Reported Outcome Measurement Information System

- Program initiated in 2004
  - ISCTM conference featured PROMIS in a session titled "Patient Reported Outcomes: New Approaches Using Item Banks and Computerized Adaptive Assessments" in 2008.
  - Developed both Computerized Adaptive Tests (CATs) and static scales for a wide range of patient and clinician reports of symptoms and functioning
- CAT model
  - Designed for brief administration – generally shorter administration times than formal static scales
  - Incorporates more items in large item banks
    - Final PROMIS item banks range from short - 4 or 5 items to long 165
  - Uses Item Response Theory (IRT) analyses as the primary psychometric method for assessing items for inclusion

Item Response Theory (IRT): Category Response Curves



Computerized Adaptive Testing- MH(CAT-MH)

- A suite of CATs for mental health applications
- Designed for multiple applications
- CATs available for overall severity of Depression, Psychosis and Substance Use Disorder
  - Uses a “bifactor” IRT model that recognizes subdomains within the construct
- CAT for Diagnosis of Major Depressive Disorder
  - Uses a decision tree model

Gibbons RD, Weiss DJ, Frank E, Kupfer D. *Computerized Adaptive Diagnosis and Testing of Mental Health Disorders* Annu. Rev. Clin. Psychol. 2016. 12:83–104

Identified “steps” in creating a CAT within the CAT-MH framework

- Develop an item bank – can use a variety of sources
  - Requires ratings of the items
- Use IRT (item response theory) to “calibrate” a model of the CAT
  - Formally a two-factor model: 1st factor the domain of interest and 2nd factor is a sub-domain within the domain
  - Items that do not discriminate on the 1st factor – overall severity - are removed
- Simulation of CAT responses using the item response patterns generated during calibration
  - Beginning with at item in the middle of the severity continuum defined by IRT, items are administered iteratively to estimate severity and uncertainty of the severity estimate, until uncertainty drops below a specified threshold.
  - Results of the simulations generate tuning parameters which will be applied during adaptive testing to select the next item in the item bank to be administered.
- Create the live program in the cloud allowing administration
- Validate the CAT in an independent sample of affected and non-affected individuals
  - Administer the CAT
  - Administer a “:gold-standard” rating scale for convergent validity of severity
  - Re-administer the CAT for test-retest reliability
  - Evaluate discriminative validly by comparing scores between affected and non-affected individuals

Example: CAT-Psychosis for Clinicians and Patients Item Bank Identification and IRT analyses

Development of Item Bank from clinician rated scales

- SADS, SAPS, SANS and BPRS yielded 144 items
  - Pruned to 73 items by eliminating items with poor discrimination
- ”Bifactor” IRT analysis assessed two latent factors overall severity and highest loading subdomain among positive, negative, disorganized and manic items
- Items reworded to create patient facing language
  - **Extent to which the patient’s ability to communicate is affected**
  - *I have trouble communicating with others.*
  - **Claims power, knowledge or identity beyond the bounds of credibility**
  - *I feel that I am a particularly important person or that I have special powers or abilities.*
- Simulation showed that administering 12 items correlated highly with full 73 item set

Guinart D, de Filippis R, Rosson , Patil B, Prizgint L, Nahal Talasazan N, Meltzer H, Kane JM, Gibbons RG *Development and Validation of a Computerized Adaptive Assessment Tool for Discrimination and Measurement of Psychotic Symptoms* Schizophrenia Bulletin,47: *644–652, 2021*

Example: CAT-Psychosis for Clinicians and Patients
New Sample of Patients and Clinicians Assessment Battery

- 140 patients: 37 affective psychosis – 123 non-affective psychosis
- Assessment battery and rationale
  - Patient CAT Psychosis – discriminant validity, convergent validity
  - Patient CAT Psychosis 1 to 7 days later – test-retest reliability
  - Clinician CAT Psychosis twice on the same day – inter-rater reliability
  - Clinician CAT Psychosis 1 to 7 days later – test-retest reliability
  - Brief Psychiatric Rating Scale-Anchored - convergent validity
  - Structured Clinical Interview for DSM-5 – discriminant validity
- 40 healthy controls
  - Patient CAT Psychosis – discriminant validity

Example: CAT-Psychosis for Clinicians and Patients Results of Reliability/ Validity Study

- Median administration time and number of items
- Patient: 1 minute 4 seconds with 12 items
- Clinician: 5 minutes, 2 seconds with 12 items
- CAT-Psychosis Patient and Clinician
  - Convergent validity with BPRS
  - Convergent validity with SCID (n = 79)
  - Correlated with each other
  - Had good test-retest reliability
  - Discriminant validity compared to normal controls

Example: CAT- Psychosis in Use

- ESPRITO network of Early Intervention for First Episode Psychosis
  - Part of a larger group of networks called EPINET funded by NIMH
  - Provides NAVIGATE – an integrated set of services and treatment for FEP
    - Individual Resiliency Training
    - Family psychoeducation
    - Supported employment and education
    - Prescriber using Measurement based Care
- Uses the CAT-Psychosis Patient version in preparation for prescriber visits
  - 12 sites across the US
  - Three items that are key in psychosis treatment are “forced”
    - Follows Guinart et al study – appear randomly in the sequence

Robinson DG, Kane JM. R01MH120594 Early-phase Schizophrenia: Practice-based Research to Improve Outcomes (ESPRITO)

Computerized Adaptive Testing – Going Forward

- “Patient-Focused Drug Development: Incorporating Clinical Outcome Assessments Into Endpoints For Regulatory Decision-Making”
  - FDA Draft Guidance issued April 2023
  - Includes consideration of Computerized Adaptive Testing
- “Because a CAT is based on IRT modeling, sponsors who wish to use CAT should demonstrate that:
  - the underlying IRT parameters are statistically sound and come from the population of interest
  - the assumptions of the IRT model and CAT are tenable
  - the adaptive and scoring algorithms were correctly implemented.”
- Guidance suggests use of hybrid scales
  - Some items required, some administered following the algorithm
  - CAT-Psychosis does exactly that
- Guidance suggests justification for CAT as opposed to a short, fixed scale perhaps chosen from the same item bank

How CAT Can Contribute to New Scale Development

- A CAT can be developed for ANY symptom or concept of interest
- Challenges
  - Identify and define the symptom/sign for measurement
  - develop a robust item bank that can withstand the calibration requirements of IRT
    - If the symptom/sign has pre-existing ratings use those for IRT analyses
    - If not, administer the item bank to appropriate cohorts of patients
  - Persuade the scientific community to use the measure
  - Meet the requirements of regulatory agencies