NIH Toolbox
Assessment of Neurological and Behavioral Function

Using Item Response Theory (IRT)-Based Instruments and Computerized Adaptive Testing (CAT) for Assessment of Health

David Cella, PhD
Center on Outcomes, Research, and Education (CORE)
October 27, 2008

NIH Blueprint for Neuroscience Research
Tools, Resources and Training

For more information, please visit www.nihtoolbox.org
Richard C. Gershon, PhD, PI gershon@northwestern.edu

This study is funded in whole or in part with Federal funds from the Blueprint for Neuroscience Research, National Institutes of Health under Contract No. HHS-N-260-2006-00007-C

Why is Toolbox Project Considering IRT?

- Some things being measured can be expressed as a range of function on a definable continuum
- That continuum can be characterized more or less equally by any number of questions or items
- Those items can be ordered from low to high, or easy to difficult
- IRT provides a way to organize the information so that a precise score can be obtained from just a few items
- Accuracy is nice; brevity is essential

Item Response Theory (IRT)

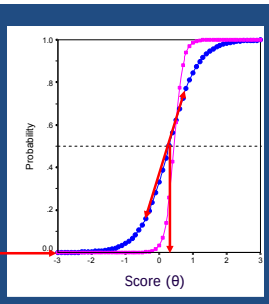
- Unlike classical test theory, IRT describes the association between where a respondent falls on a given concept or trait (θ) and the probability of a particular response to an item
  - Reliability is not a property of a "test;" it varies across the measurement continuum
  - Scores obtained can be treated as independent of the exact questions asked
- An IRT approach can be useful for examining
  - Item-level properties of an instrument ("item bank")
  - Information provided by administered items (or groups of items) across θ
- This adds up to FLEXIBILITY for the researcher

IRT Models

- IRT models differ with respect to the number of item parameters that are estimated. (3PL, 2PL, 1PL (Rasch))
- Polytomous models are used to describe items with 3+ response options, while dichotomous models address items with binary responses.
- Assumptions
  - unidimensionality
    - the instrument measures one dominant construct
  - local independence
    - the likelihood of answering an item in one direction is unrelated to the probability of answering another item in the same direction, at constant θ

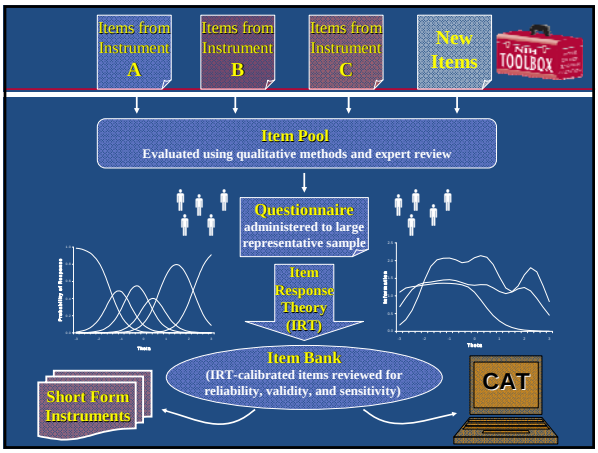
IRT Parameters

- A person's trait level and their likelihood of responding to an item can be described in terms of several parameters.
  - The point at which a participant has 50% likelihood of responding in the keyed direction
    - *b, or difficulty*
  - Its ability to distinguish low trait from high trait individuals
    - *a, or discrimination*
  - The likelihood of low attribute participants responding in the keyed direction
    - *c, or guessing*



Clinical and Health Services Research Applications of IRT

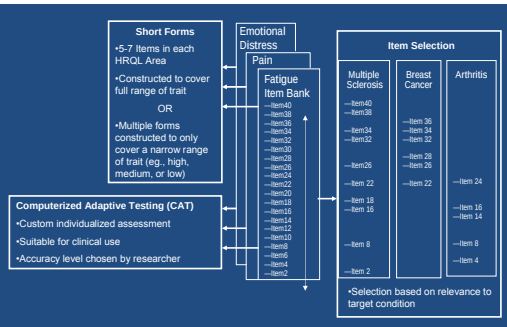
- Efficient collection of health outcomes data in clinical trials
  - Comparing health interventions and strategies
  - Comparing pharmaceutical treatments
- Monitoring the health outcomes of populations
  - Health plan members
  - Medicare beneficiaries
  - US general population
- Publicly available, adaptable and sustainable
  - Online common item repository
  - Online CAT and short form delivery



Advantages of Short-Forms Developed from Item Banks

- Flexibility in length and content
  - Select items matched to clinical features and severity in the target population
  - Select items known to provide the most information
- Any form selected or created produces scores on a common metric

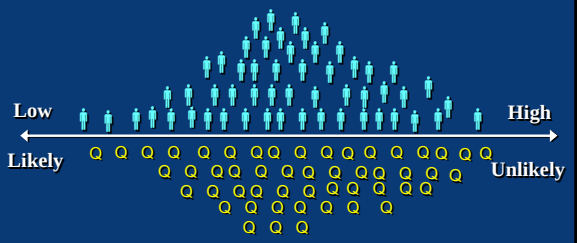
Three Diseases, Three Questionnaires, One Metric



Gershon et al, Exp Rev Pharmoecon Outcomes Res. (2003)

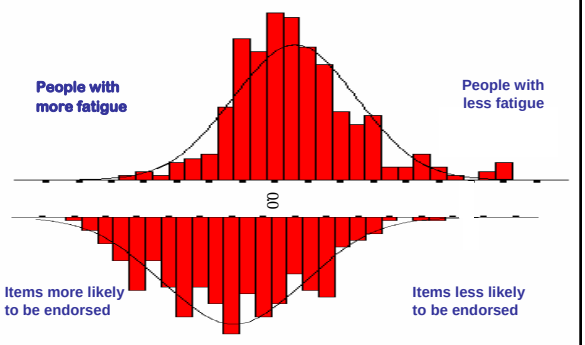
Interpretation Aids

PRO Bank Person Score



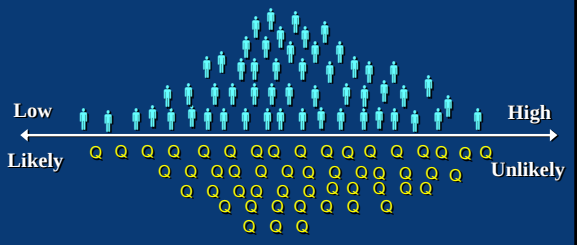
Item Location

People and Items Distributed on the Same Metric: Fatigue



Interpretation Aids

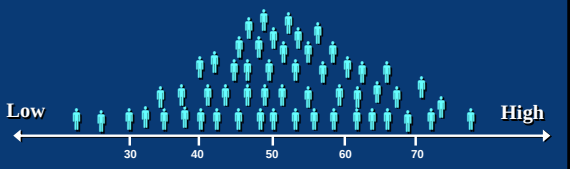
PRO Bank Person Score



Item Location

Interpretation Aids

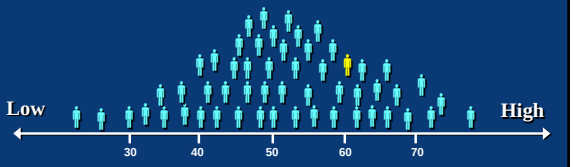
PRO Bank Person Score



M = 50, SD = 10

Interpretation Aids: Fatigue example

Fatigue Score=60



This person's fatigue score is **60**, significantly worse than average (50). People who score **60** on fatigue tend to answer questions as follows:

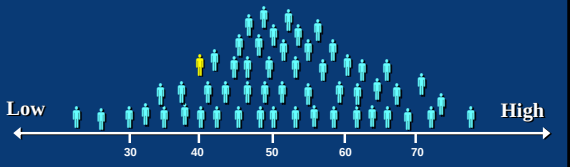
…"I have been too tired to climb one flight of stairs: VERY MUCH

…"I have had enough energy to go out with my family: A LITTLE BIT

Click here if you would like to see this person's individual answers

Interpretation Aids: Fatigue example

Fatigue Score=40



This person's fatigue score is **40**, significantly better than average (50). People who score **40** on fatigue tend to answer questions as follows:

…"I have been too tired to climb one flight of stairs: SOMEWHAT
…"I have had enough energy to go out with my family: VERY MUCH

Click here if you would like to see this person's individual answers

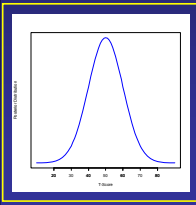
Computerized Adaptive Testing (CAT)

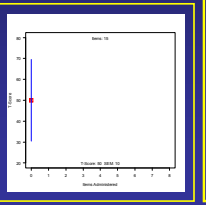
- Estimates location (severity; capability) of a person on a domain (concept) by selecting questions based on that person's prior answers
- Iteratively estimates a person's standing on the domain (e.g., depressive symptoms) and administers only the most informative items, achieving precision with a minimum possible number of questions.

Beginning of CAT:

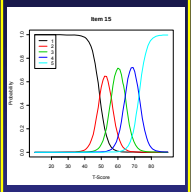
PROMIS Depression Bank

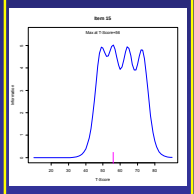
T-Score = 50 | SE = 10





Best Item–I felt depressed

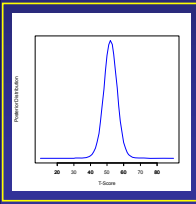


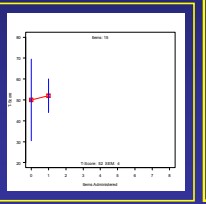


I felt depressed

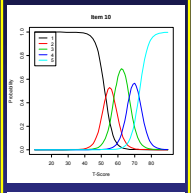
1. Never
2. Rarely
3. Sometimes
4. Often
5. Always

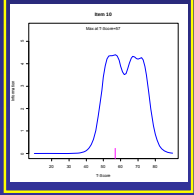
T-Score = 52 | SE = 4





Next Best Item–I felt like a failure

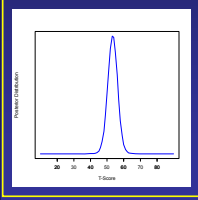


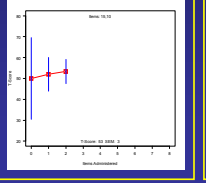


I felt like a failure

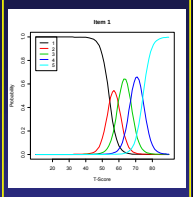
1. Never
2. Rarely
3. Sometimes
4. Often
5. Always

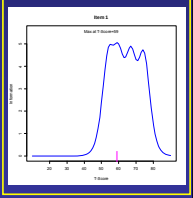
T-Score = 53 SE = 3





Next Best Item-I felt worthless

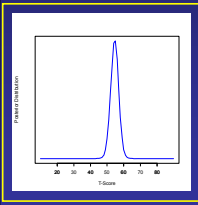


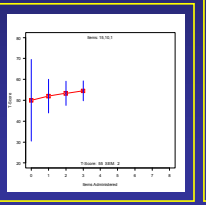


I felt worthless

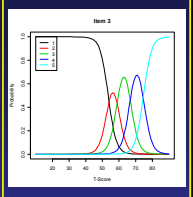
1. Never
2. Rarely
3. Sometimes
4. Often
5. Always

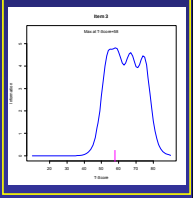
T-Score = 55 SE = 2





Next Best Item-I felt helpless

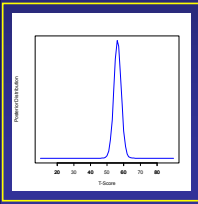


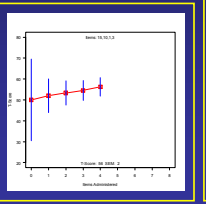


I felt helpless

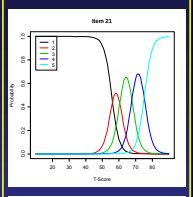
1. Never
2. Rarely
3. Sometimes
4. Often
5. Always

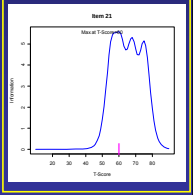
T-Score = 56 SE = 2





Next Best Item-I felt hopeless

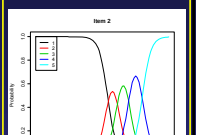


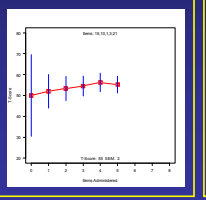


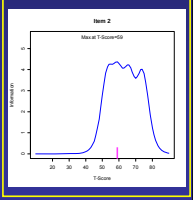
I felt hopeless

1. Never
2. Rarely
3. Sometimes
4. Often
5. Always

T-Score = 55 SE = 2





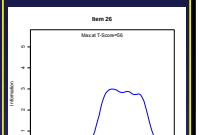


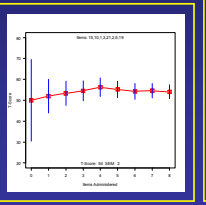
Next Best Item-nothing to look f

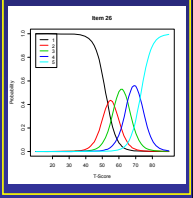
I felt unhappy

1. Never
2. Rarely
3. Sometimes
4. Often
5. Always

T-Score = 54 SE = 2



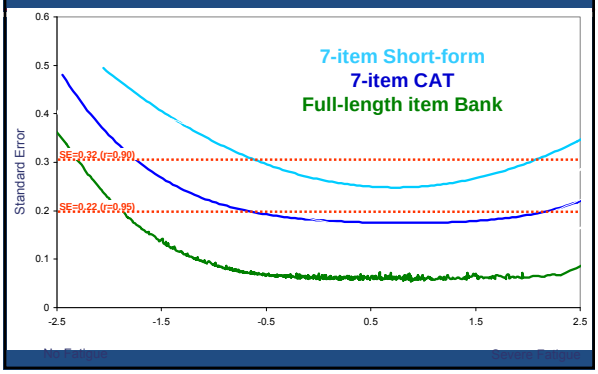


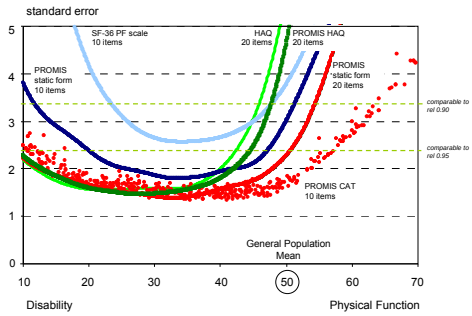


Next Best Item-life was empty

Comparison of Measurement Precision

Full-length Item Bank vs. CAT vs. Short-form





Conclusion

Item Response Theory (IRT) models enable reliable and precise measurement

- Fewer items needed for equal precision
  - Making assessment briefer
- More precision gained by adding items
  - Reducing error and sample size requirements
- Error is understood at the individual level
  - Enabling practical individual assessment

NIH Toolbox
Assessment of Neurological and Behavioral Function

Using Item Response Theory (IRT)-Based Instruments and Computerized Adaptive Testing (CAT) for Assessment of Health

David Cella, PhD
Center on Outcomes, Research, and Education (CORE)
October 27, 2008

NIH Blueprint

For more information, please visit www.nihtoolbox.org
Richard C. Gershon, PhD, PI gershon@northwestern.edu

This study is funded in whole or in part with Federal funds from the Blueprint for Neuroscience Research, National Institutes of Health under Contract No. HHS-N-260-2006-00007-C