SVS Society for Vascular Surgery

Psychometric Methods

Table of Contents

PRO: Basics

- **<u>Definition (FDA)</u>**: any report of the status of a patient's health condition that comes directly from the patient, without interpretation of the patient's response by a clinician or anyone else

PRO: Basics

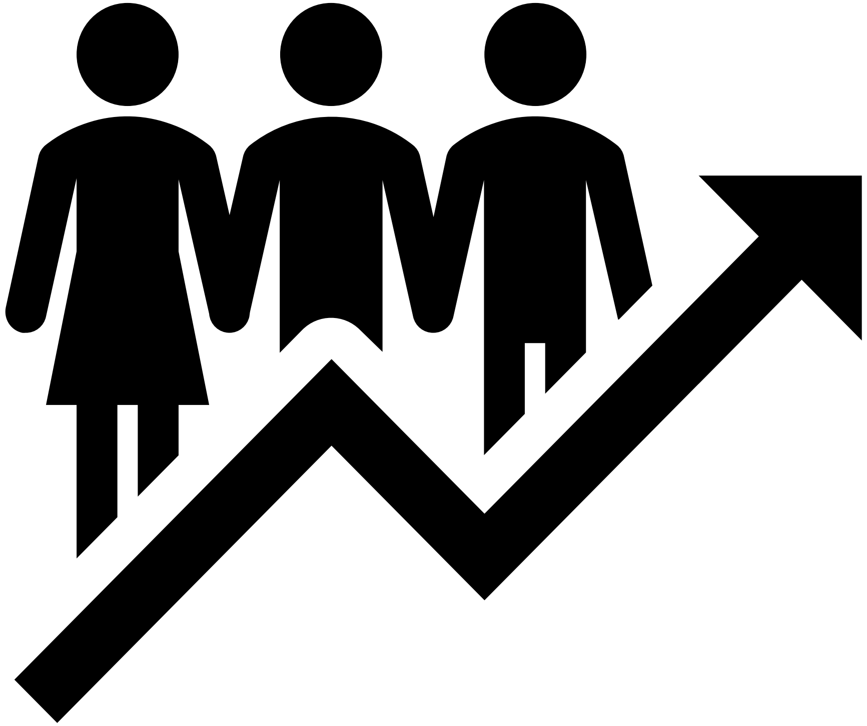
- **Description**: typically include information about health-related quality of life (HRQOL), symptoms, function, satisfaction with care or symptoms, adherence to prescribed medications or other therapy, and perceived value of treatment

PRO: Basics

- **Form**: questionnaire filled out by patient or given to patient
- **Items**: grouped into "domains" representing general category of assessment
- **Scoring**: unique to each PROM; relevant to its topic

PRO: Basics

- **Use**: data are used to inform and guide patient-centered care, clinical decision-making, and health policy decisions and are an important component in learning healthcare systems



PRO: Categories

Health related quality of life (HRQL)

- Assess how a disease and its treatment affect the physical, psychologic, and/or social aspects of life
- Objective assessments of functioning or health status: example; frequency of pain
- Subjective evaluation: example; extent to which pain hinders ability to engage in social activities

Satisfaction

- Entirely subjective
- Extent that the patient believes that high-quality health care was delivered
- Could potentially be defined differently by different people and by the same person at different times
- HCAHPS survey: random sample of discharged patients

PRO: Categories

Disease Specific

- Specifically designed to capture the symptoms, functioning, and quality of life as it relates to a specific disease state
- More sensitive to an individual’s experience as it relates to the particular condition

Generic

- Used to capture an individual’s overall health and is not specific to a particular disease
- Allow for comparison across different disease populations

PRO: Basics

- **Goal:** to improve clinical decision-making within the context of data-driven care

- **Successful integration:** continuous collection of accurate, valid, accessible, and reusable data in real time to support patient care, clinical research, quality improvement, and comparative effectiveness research (CER)

Developing PRO

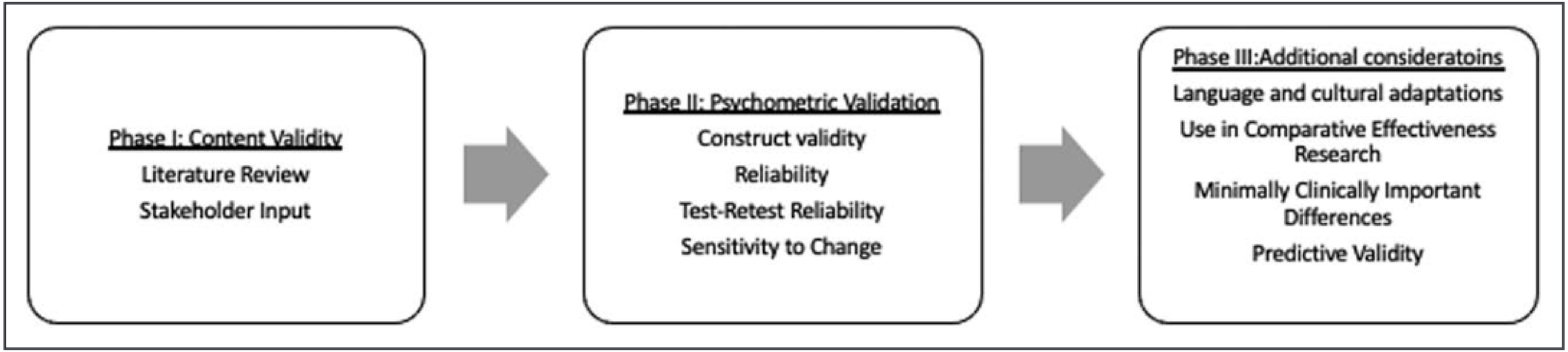
- **When to use:** the concept being measured is best known by the patient or best measured from the patient's perspective (example: Wong-Baker FACES scale to communicate a self-assessed measure of discomfort or pain to a healthcare provider)

- **Before using:** determine whether an adequate instrument exists to address and measure the concepts of interest, or whether an existing instrument could be modified appropriately (may involve combining, modifying, or developing new instruments)

Developing a PRO

- **Requirement:** provide documentation of patient input during the development process.

- **Evidence:** demonstrate instrument's performance in the specific application for which it was intended



Phase I: Content Validity

Literature Review

Stakeholder Input

Phase II: Psychometric Validation

Construct validity

Reliability

Test-Retest Reliability

Sensitivity to Change

Phase III:Additional consideratoins

Language and cultural adaptations

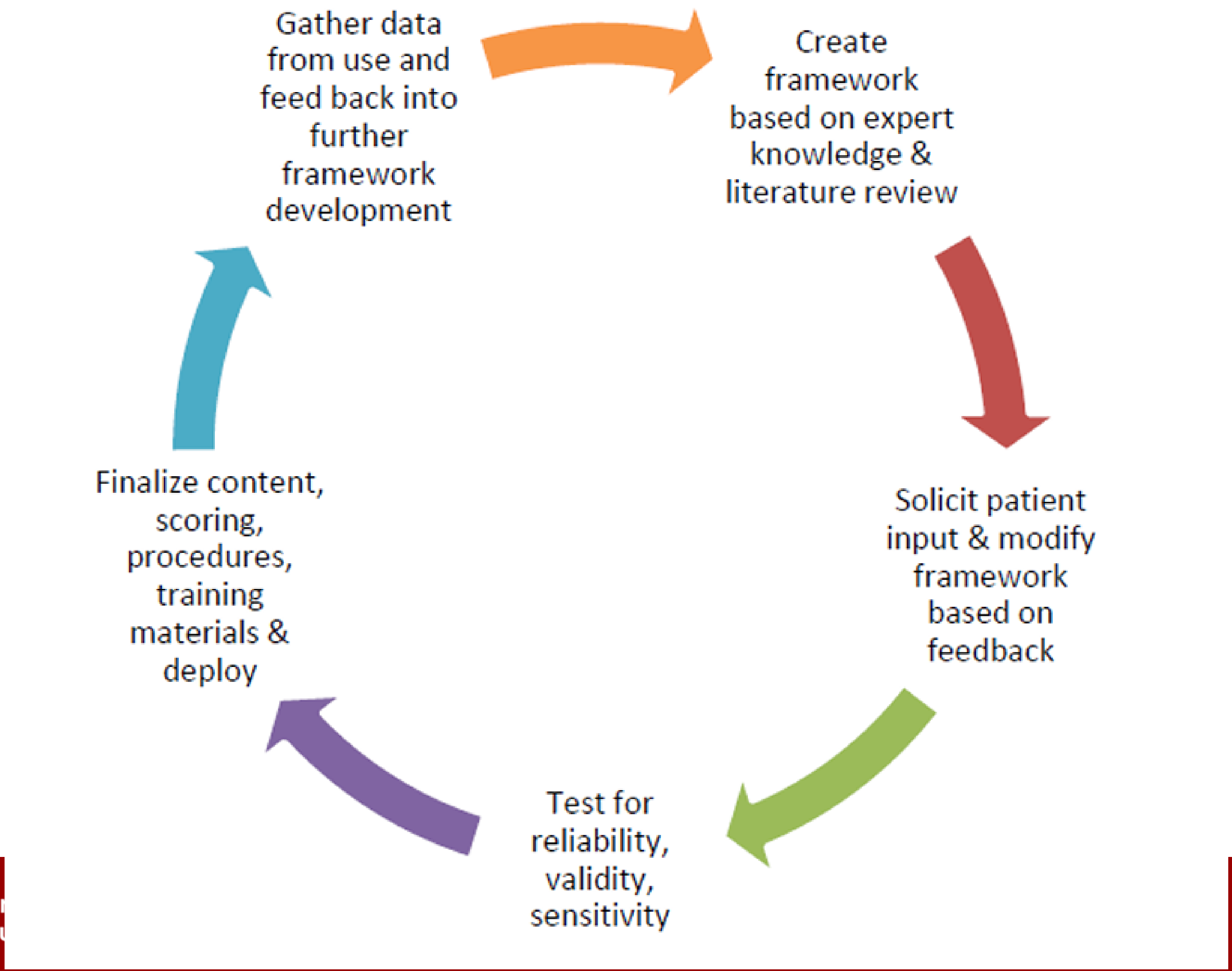
Use in Comparative Effectiveness Research

Minimally Clinically Important Differences

Predictive Validity

Figure. Stages of development and validation of a patient-reported outcome measure.

Development and Improvement Cycle for PRO



Gather data from use and feed back into further framework development

Create framework based on expert knowledge & literature review

Solicit patient input & modify framework based on feedback

Test for reliability, validity, sensitivity

Finalize content, scoring, procedures, training materials & deploy

1 Create conceptual model

- **Basis:** define interest and boundaries; align with research goals

- **Framework**: measurable items that collectively describe a domain

❷ Patient Input

- **Adjust:** solicit patient input and adjust framework based on response

- Focus groups and/or individual interviews

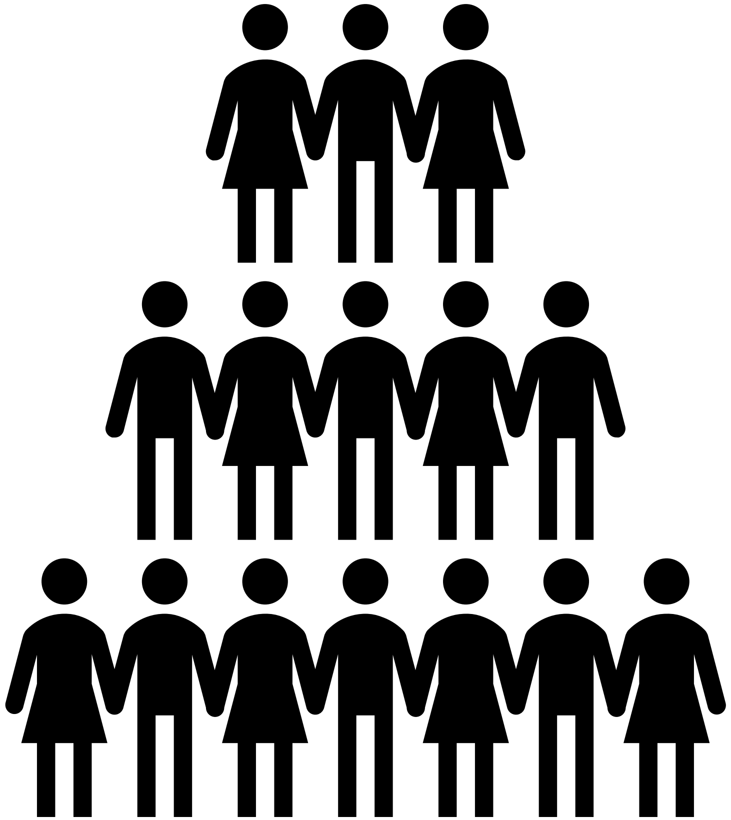
3 Testing

- Draft instrument
- Give to diverse patients in target group
- Evaluate for reliability, validity, and ability to detect change

DRAFT

4 Deploy

- Finalize content, scoring, procedures, and training materials
- Administer to large group
- Confirm that it measures what it intends to measure by comparing the responses with objective measures of health.



5 Gather Data

- **Iterative process:** modify instrument and repeat cycle according to data

- Translation and cultural adaptation and repeat of step four

Validating PRO

- Measurable items that collectively describe a domain
- Domain: specific feeling, function, or perception being measured
- Obtain feedback from patients and modify accordingly

Description of Terms

- **Validity:** degree to which an instrument measures what is intended to measure

- **Reliability:** degree to which measures are reproducible and consistent over time in patients with a stable condition

- **Responsiveness:** degree to which an instrument detects meaningful change over time

- **Acceptability:** degree to which the instrument is acceptable to the patient

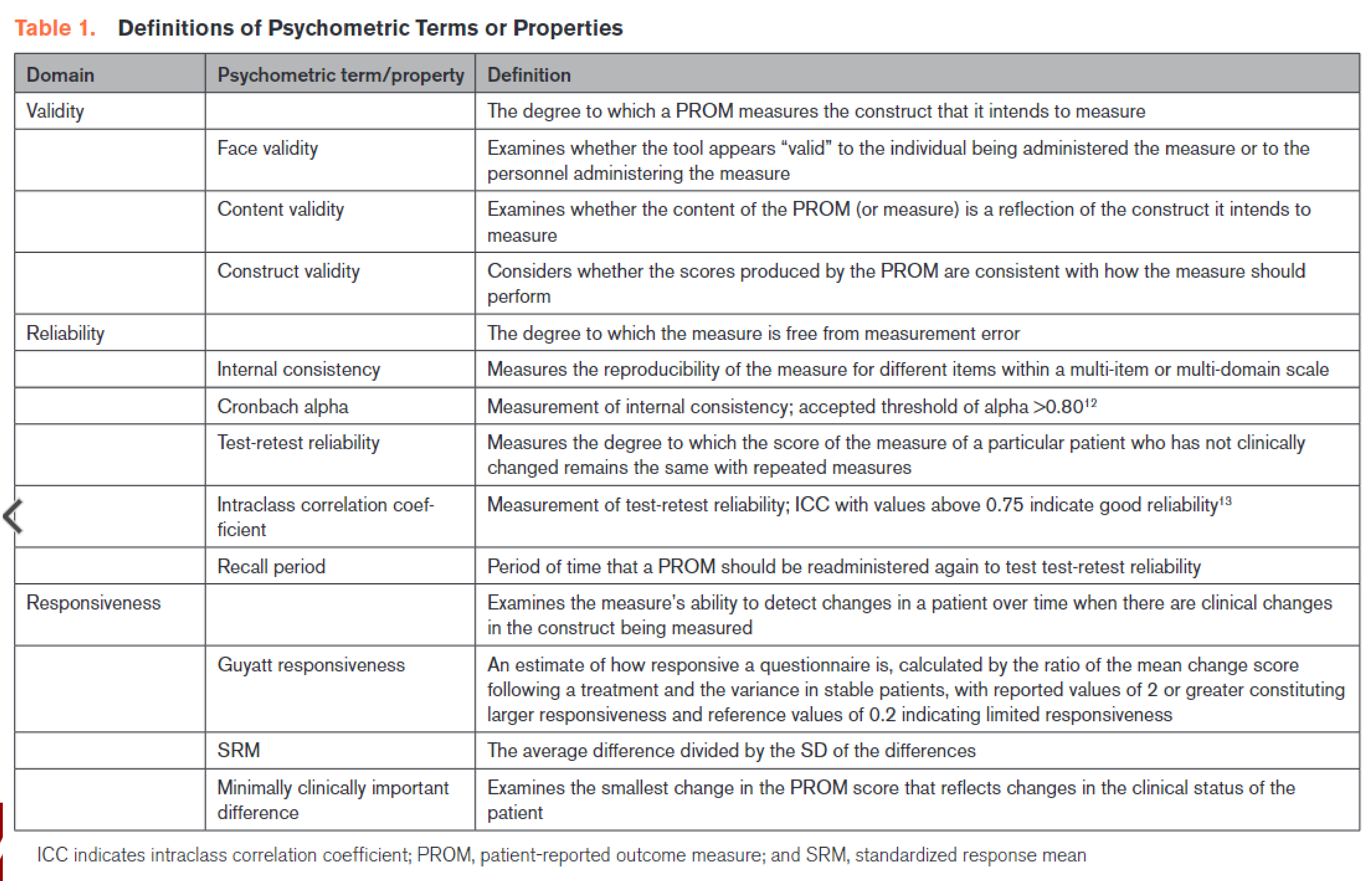


Table 1. **Definitions of Psychometric Terms or Properties**

Domain	Psychometric term/property	Definition
Validity		The degree to which a PROM measures the construct that it intends to measure
	Face validity	Examines whether the tool appears "valid" to the individual being administered the measure or to the personnel administering the measure
	Content validity	Examines whether the content of the PROM (or measure) is a reflection of the construct it intends to measure
	Construct validity	Considers whether the scores produced by the PROM are consistent with how the measure should perform
Reliability		The degree to which the measure is free from measurement error
	Internal consistency	Measures the reproducibility of the measure for different items within a multi-item or multi-domain scale
	Cronbach alpha	Measurement of internal consistency; accepted threshold of alpha >0.80[12]
	Test-retest reliability	Measures the degree to which the score of the measure of a particular patient who has not clinically changed remains the same with repeated measures
	Intraclass correlation coefficient	Measurement of test-retest reliability; ICC with values above 0.75 indicate good reliability[13]
	Recall period	Period of time that a PROM should be readministered again to test test-retest reliability
Responsiveness		Examines the measure's ability to detect changes in a patient over time when there are clinical changes in the construct being measured
	Guyatt responsiveness	An estimate of how responsive a questionnaire is, calculated by the ratio of the mean change score following a treatment and the variance in stable patients, with reported values of 2 or greater constituting larger responsiveness and reference values of 0.2 indicating limited responsiveness
	SRM	The average difference divided by the SD of the differences
	Minimally clinically important difference	Examines the smallest change in the PROM score that reflects changes in the clinical status of the patient

ICC indicates intraclass correlation coefficient; PROM, patient-reported outcome measure; and SRM, standardized response mean

Psychometric Methods

Why do I need to know about psychometric methods for developing PROMs?

The psychometric properties of a tool determine its value

Does the PROM reflect what it aims to measure?

VALIDITY

Is the PROM stable over time?

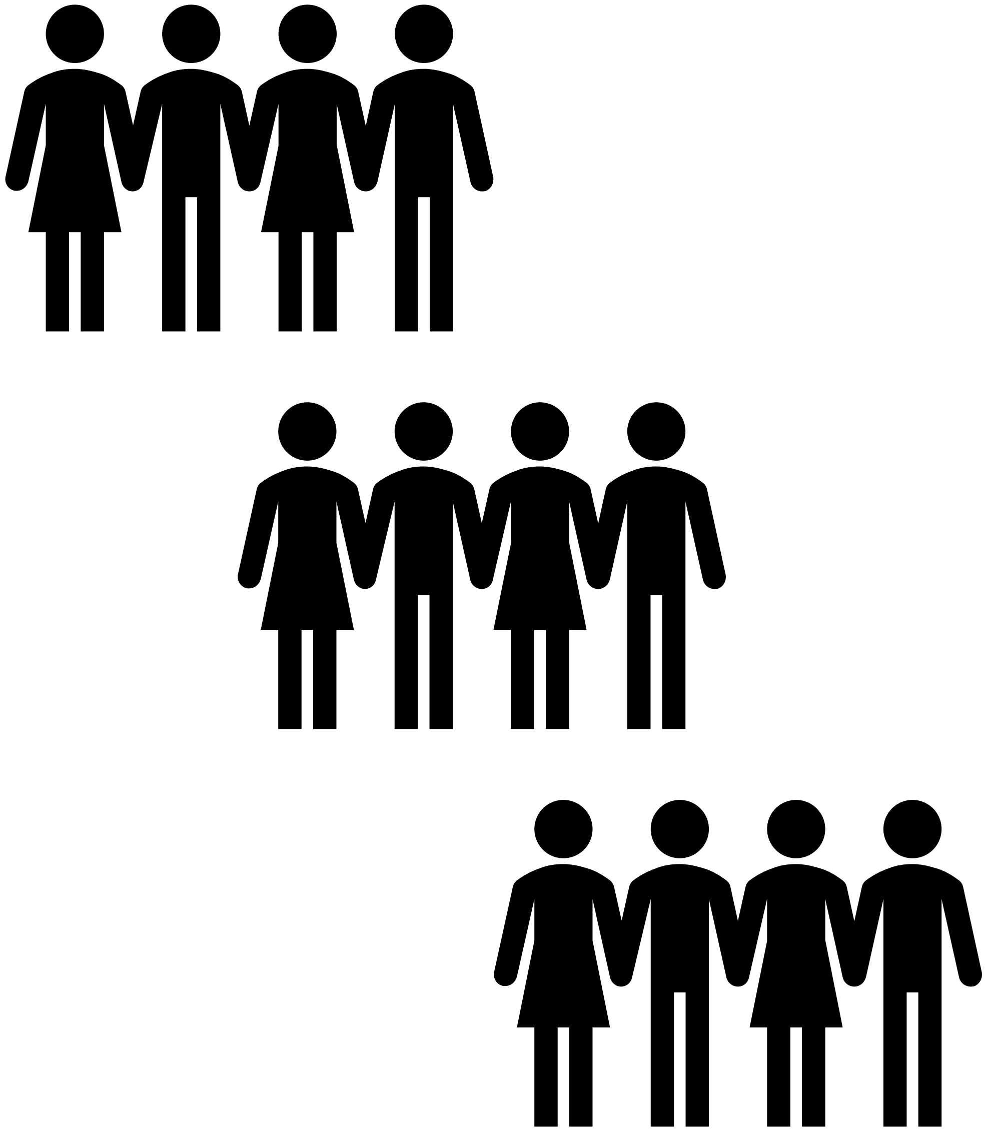
RELIABILITY

Can the PROM detect changes over time?

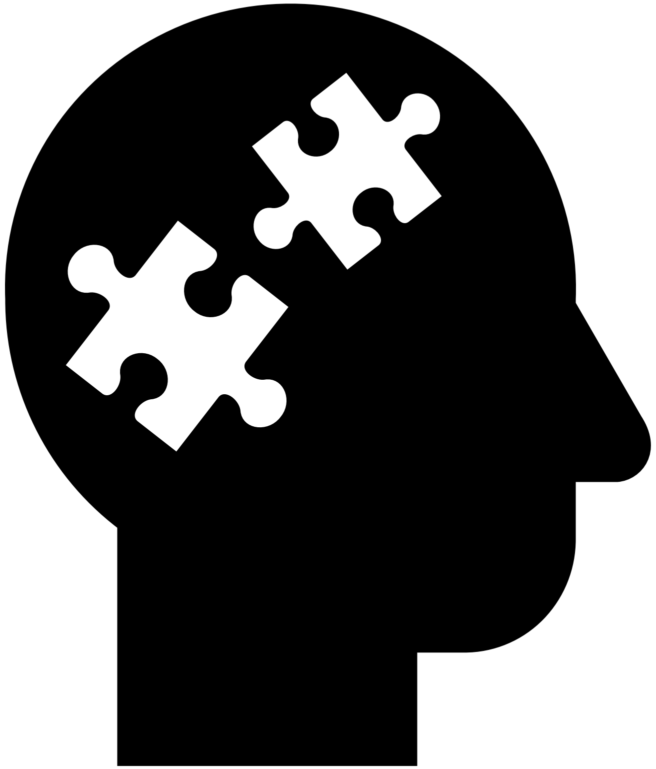
RESPONSIVENESS

A PROM can be considered relevant, valid and reliable only if it has proven psychometric qualities in all these dimensions

Understanding psychometric properties provides a way to develop, evaluate, and improve PROM's



Psychometric theory offers a range of tests that can be used as supportive evidence of both validity and reliability of a PROM



Types of psychometric analysis

- Classical test theory (CTT)
- Item response theory (IRT)

Types of psychometric analysis

- **Classical test theory (CTT)**

- Item response theory (IRT)

Classical test theory (CTT)

- A quantitative approach to testing the reliability and validity of a scale based on its items

CTT Assumes:

- Observed score of a PROM is the sum of the True score and random error

CTT Assumes:

- **True Score** is the attribute of interest
- **Error** is completely random and uncorrelated with true score

CTT Concepts:

- Descriptive assessment
- Item discrimination
- Dimensionality
- Reliability
- Sample size

CTT Concepts:

Descriptive assessment

- Means and std dev
- Z score for severity

+ Higher variability

+ Mean closer to median

+ Spread across categories

CTT Concepts:
Item Discrimination

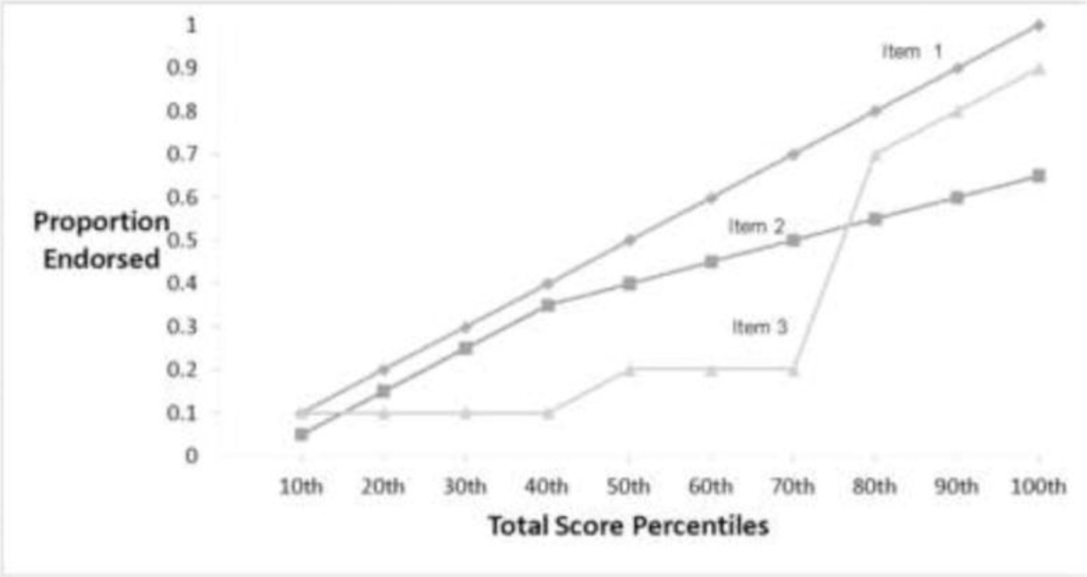
1. Partition into highest and lowest overall scores (e.g. top and bottom 25%)
2. Determine proportion of each item endorsed by upper and lower group
3. Discrimination index = upper group proportion minus lower group proportion

CTT Concepts:

Descriptive assessment

- Corrected item-to-scale correlation

+ Large: > 0.37



item response curve

CTT Concepts: Dimensionality

Extent an item measures a property distinctly

- ☑ Multi-trait scaling analysis
- ☑ Factor analysis

CTT Concepts: Reliability

- If responses are inconsistent (not reliable) it implies invalidity

- Converse is NOT true: consistent responses do not imply validity

CTT Concepts:
Reliability

Test-retest reliability

- Kappa statistic: categorical responses
- Intraclass correlation: continuous responses

Multi-item scales

- Cronbach’s coefficient alpha
- Covariance and correlation based formulas

CTT Concepts:

Sample Size

Early stage

- 30-50 subjects
- Add more if no trends
- More categories need more subjects
- Recruit for representation

Later stage

- 5 cases / item; min 300 cases
- No. of subjects = 10x no. of items

Types of psychometric analysis

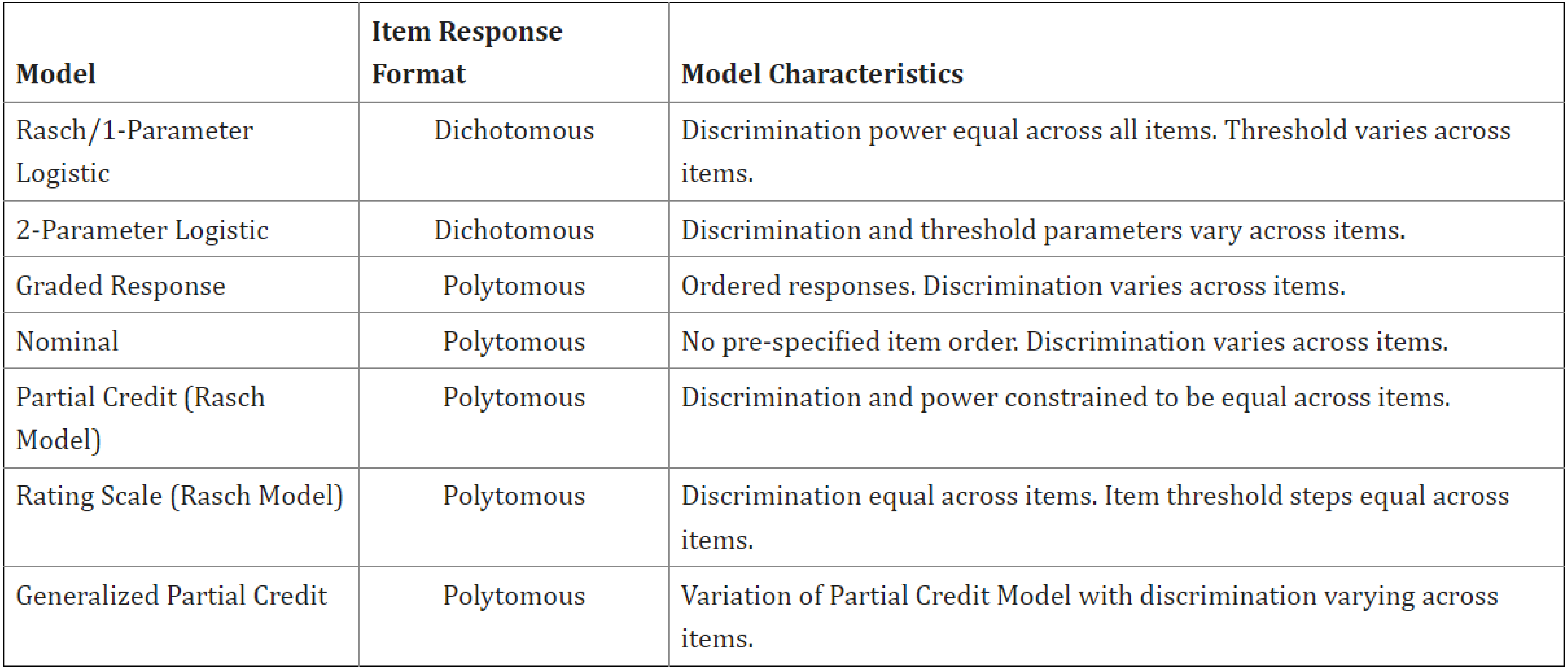
- Classical test theory (CTT)
- **Item response theory (IRT)**

Item Response Theory (IRT)

- A collection of measurement models that attempt to explain the connection between observed item responses on a scale and an underlying property

IRT Concepts: Models

- Mathematical equations describing the association between subjects' levels on a latent variable and the probability of a particular response to an item



Model	Item Response Format	Model Characteristics
Rasch/1-Parameter Logistic	Dichotomous	Discrimination power equal across all items. Threshold varies across items.
2-Parameter Logistic	Dichotomous	Discrimination and threshold parameters vary across items.
Graded Response	Polytomous	Ordered responses. Discrimination varies across items.
Nominal	Polytomous	No pre-specified item order. Discrimination varies across items.
Partial Credit (Rasch Model)	Polytomous	Discrimination and power constrained to be equal across items.
Rating Scale (Rasch Model)	Polytomous	Discrimination equal across items. Item threshold steps equal across items.
Generalized Partial Credit	Polytomous	Variation of Partial Credit Model with discrimination varying across items.

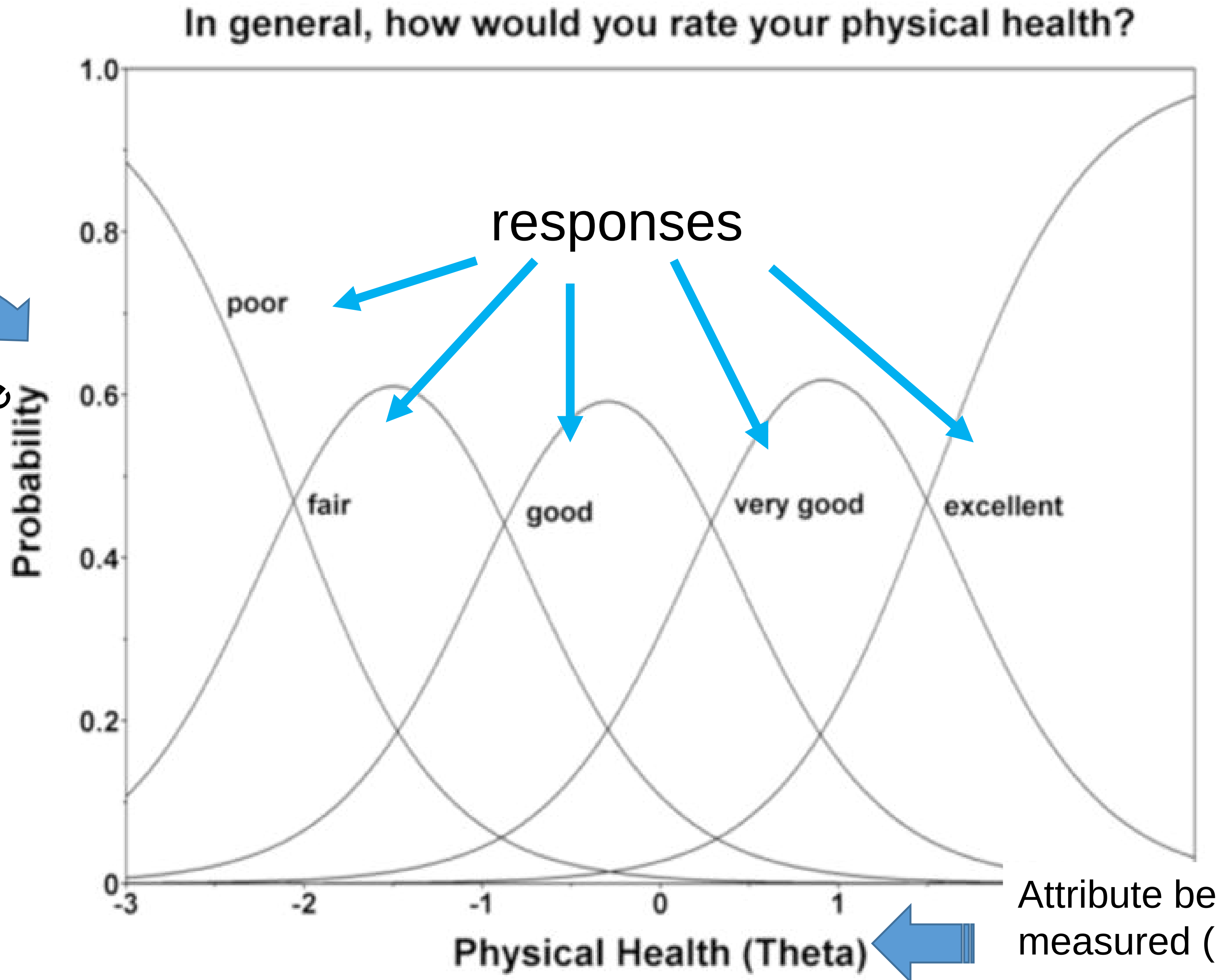
IRT Concepts:

- Category Response Curves
- Item information
- Person Item Map
- Sample size

IRT Concepts: Category Response Curves

- Display relative position of each category along continuum of concept being measured

➕ Ideal: each response category being most likely to be selected for some segment of the underlying continuum of the attribute



responses

of selecting response

Attribute being measured (Z scale)

IRT Concepts: Item Information

- Assessment of precision of item to distinguish subjects across different levels of property being measured

➕ Higher item information implies more precision

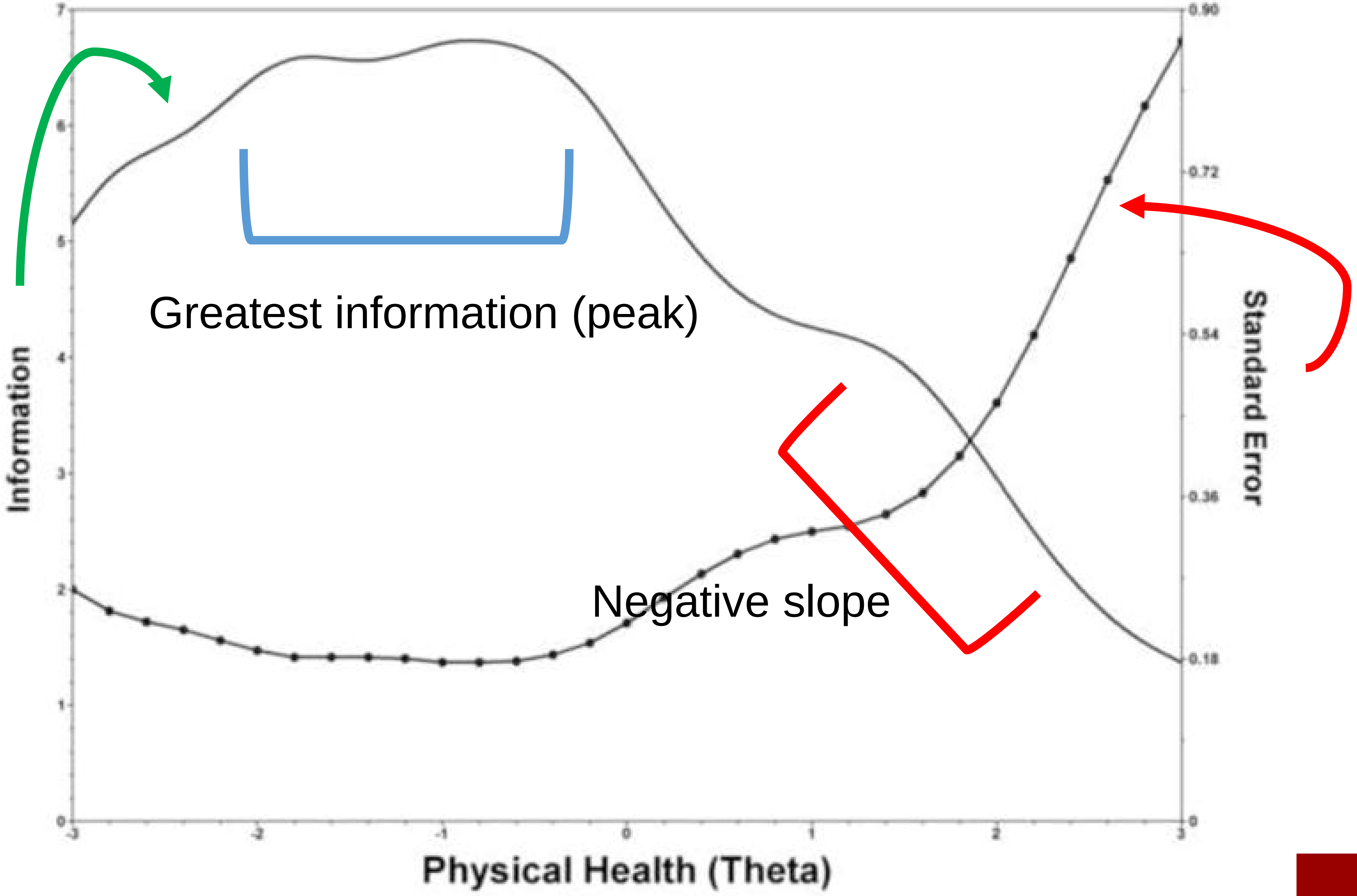
IRT Concepts: Item Information

- Sums together to form scale information

➕ Peak of curve shows where item yields greatest information

➕ Peaked curve = more information than flat (higher item discrimination parameter)

➖ Negative parameter (slope); should weed out item



Greatest information (peak)

Negative slope

IRT Concepts: Item Information

- Directly related to reliability

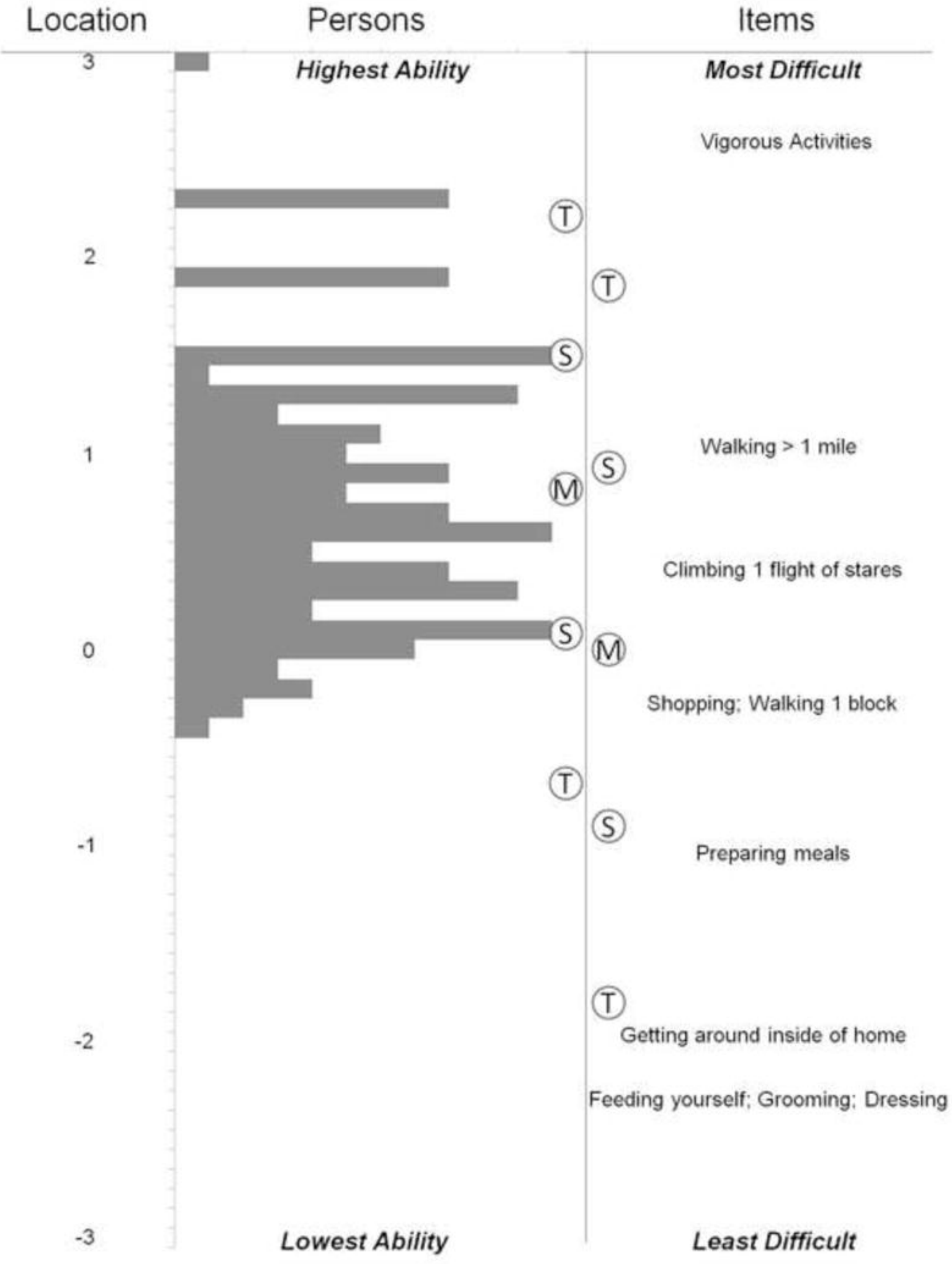
- Typically varies by location along the underlying continuum of the attribute (ie low, middle, high scores)

IRT Concepts: Person-Item Map

- Shows the relationship between item difficulty and person attribute
- Can show the extent of item coverage, redundancy, and range of the attribute in the sample

IRT Concepts: Sample Size Considerations

- Choice of IRT model
  - No. of parameters
- Type of response:
  - No. of categories
- Study purpose:
  - Trends vs precise measurements
- Sample distribution:
  - Even vs. skewed
- Number of items
- Item relationship with attribute



Example:

- Questionnaire contains more easy items than hard ones
- Redundant items; can be removed without sacrificing information
- Cluster at higher end of scale; need more challenging items

IRT Assumes:

- Monotonicity
  - Probability of endorsing each response category increases with person's location on the attribute
- Unidimensionality
  - Person's level on the construct accounts full for their responses

Item Response

✓Requires adequate sample size

✓Sample size considerations depend on several factors

✓Person-item map insights

Classical Test

✓Small qualitative data

✓Requires fewer items

✓Use as 1st step: get preliminary information on validity

References

Cappelleri JC, Lundy JJ, Hays RD. Overview of Classical Test Theory and Item Response Theory for Quantitative Assessment of Items in Developing Patient-Reported Outcome Measures. *Clin Ther* 2014;36(5): 648–662. doi:10.1016/j.clinthera.2014.04.006
https://www.clinicaltherapeutics.com/article/S0149-2918%2814%2900204-5/fulltext

Stover AM, McLeod LD, Langer MM Chen W-H, Reeve BB. State of the psychometric methods: patient-reported outcome measure development and refinement using item response theory *Journal of Patient-Reported Outcomes* 2019;3:50. https://doi.org/10.1186/s41687-019-0130-5
https://link.springer.com/article/10.1186/s41687-019-0130-5

Poku E, Duncan R, Keetharuth A, Essat M, Phillips P, Buckley Woods H, Palfreyman S, Jones G, Kaltenthaler E, Michaels J. Patient-reported outcome measures in patients with peripheral arterial disease: a systematic review of psychometric properties. *Health and Quality of Life Outcomes* 2016;14:161. DOI 10.1186/s12955-016-0563-y
https://hqlo.biomedcentral.com/articles/10.1186/s12955-016-0563-y