

Key terms, abbreviations, and definitions

Term	Definition
Category response curve (CRC)	See item characteristic curve (ICC)
Ceiling effect ¹	Occurs when a score distribution is skewed so that a large proportion of respondents score at the highest level on the measure. When ceiling effect occur, the measure cannot detect changes in a higher direction.
Classical test theory (CTT)	Traditional psychometric methods such as total score, point-biserial and biserial correlations, item discrimination as item-total score correlation, item difficulty as proportion correct, and Cronbach's α , in contrast to IRT
Computer-Adaptive Testing (CAT)	CAT integrates the advances in measurement theory and the power of computer technology to administer a PRO instrument that selects questions on the basis of a patient's response to previously administered questions (or possibly other prior information). Highly informative questions are carefully selected so that we may determine a person's standing on a domain (e.g., physical functioning, depression) with the minimal number of questions without a loss in measurement precision.
Culturally sensitive ¹	In the context of culturally sensitive measures, measures that consider the perspective of a cultural subgroup in framing a concept or writing an item. Thus, concepts and measures account for the experiences, beliefs, values, expectations, perceptions, and language of a cultural group.
Dichotomous (binary) response categories	An item having two response categories such as yes/no, true/false, or agree/disagree.
Differential Item Functioning (DIF)	One group responds differently to an item than another group despite controlling for differences on the measured construct. When the data for the two groups are modelled separately, DIF will result in different IRT model parameter estimates for the item. Scales containing DIF items may have reduced validity for between-group comparisons, because their scores may be indicative of a variety of attributes other than those the scale is intended to measure.
Difficulty (as in item difficulty)	See definition for <i>Threshold</i> Parameter. Describes the location of the item along theta (the construct continuum). The term "difficulty" reflects IRT's educational upbringing where a "difficult" math item would require a person with high math skills to answer the question correctly. In health outcomes measurement, the term difficulty may apply when measuring physical functioning. An item with high difficulty may ask if a person

	can perform vigorous activities like running and an item with low difficulty may ask if a person can walk one block.
Discrimination parameter (a , α) – slope	IRT model item parameter that indicates the strength of the relationship between an item (or item response) and the measured construct. Also, the parameter indicates how well an item discriminates between respondents below and above the item threshold parameter, as indicated by the slope of the ICCs.
Floor effect ¹	When a score distribution is skewed so that a large proportion of respondents score at the lowest level on the measure. When floor effects occur, the measure cannot detect changes in a lower direction.
Health-Related Quality of Life (HRQOL)	Quality of life domains affected by one's health.
Information function/curve	Indicates the range over θ for which an item, item response, or scale is most useful (precise) for measuring persons' levels. When applied to an item category response, it's called the category information function; these can be summed for an item to obtain the item information function; these in turn can be summed for a set of items to obtain the test or scale information function. When item information is summed for all the items in a bank (such as might be used for a CAT) the result is the bank information function.
Item	A question (including its response choices) in a scale.
Item Bank	A large collection of questions that are organized, calibrated, and matched to a given construct. An item bank can be used as the foundation for CAT-measurement or deriving short fixed instruments.
Item Banking	The data-driven positioning of items within a set of other items measuring the same underlying latent trait or construct.
Item characteristic curve (ICC)	Models the probabilistic relationship between a person's response to each category for an item and their level on the underlying construct (θ). Also called category response curve (CRC) for polytomous items, and item response functions for dichotomous items. Some literature refers to ICCs as trace lines.
Item Response Theory (IRT) Modeling	IRT refers to a set of mathematical models that describe, in probabilistic terms, the relationship between a person's response to a survey question and his or her standing on the construct being measured by the scale.
Item-total correlation	Measure of the relationship between an item and the total score from the set of items within the scale. Higher correlations indicate a stronger relationship between the item

	and scale score. Item-total correlation is a CTT concept (measuring an item's discrimination ability) that is not used in IRT.
Latent trait	A latent trait is an unobservable latent dimension, like depression, fatigue, or pain, which is thought to give rise to a set of observed item responses. In IRT, the latent trait being measured by a scale is denoted as theta (θ).
Local independence assumption	Once you control for the dominant factor influencing a person's response to an item, there should be no significant association among the item responses. Usually occurs in items for which there is some common stimulus or word stem.
Minimally important difference (MID) ¹	Also known as Minimal Clinically Important Difference. That difference in score on a scale that corresponds to the smallest change in health status that stakeholders (patients, significant others, clinicians, and other decision makers) consider important. The MID is the smallest difference that would lead to the initiation of a change in a course of action, such as prescription of medication or change in behavioural regimen. Assessments of MID need not and do not always agree among different stakeholders in the health decision.
Patient-Reported Outcomes (PRO)	Health outcomes that are optimally measured by asking the consumer (i.e., patient). This includes satisfaction with care, health-related quality of life, and symptoms experienced by a condition or its treatment.
Person fit ¹	The extent to which a particular individual's pattern of item responses is consistent with that predicted by the IRT models. Inconsistencies between the response patterns of particular individuals and the IRT model predictions may suggest careless or cognitive problems on the part of the particular respondents.
Polytomous response categories	An item having more than two response categories. For example, a 5-point Likert type scale.
Rasch Models	Refers to a family of IRT models that constrain the discrimination parameter to be equal across all items. Rasch models have several attractive features including the ability to compare two person's summed scores independent of the subset of items they endorsed.
Reliability ¹	The extent to which a measure is free of measurement error; the ration of true score variance to observed score variance. Two basic forms of reliability are reproducibility and item homogeneity or consistency. Internal consistency reliability is a method for estimating reliability from the correlations among the items in a multi-item scale.
Scale	Consists of multiple items that measure a single domain such as fatigue.

Standard error of measurement (SEM)	Describes an expected observed score fluctuation due to error in the measurement tool. Standard deviation of error about an estimated score. In CTT, the SEM is the same for all score levels; in IRT it can vary from score to score, and therefore can be used as a termination criterion in CAT when the number of items is allowed to vary.
Theta (θ)	Unobservable construct (or latent variable) being measured by a scale. It is estimated from the responses people give to test items that have been previously calibrated by an IRT model.
Threshold parameter (b, β) – difficulty, location, severity	IRT model item parameter that indicates the severity or difficulty of an item response, or the location along the θ -continuum of the item response categories. See discussion under <i>Difficulty</i> . The term “severity” could reflect, for example, in pain measurement (and for other relevant constructs) how severe the pain must be to indicate “quite a bit...over the past 4 weeks pain has interfered with normal work”.
Unidimensionality assumption	Assumes that one underlying (or dominant) factor (variable or trait) accounts for a person’s response to a question within a scale.

Notes:

¹ Definition taken from Medical Care Special Supplement (Volume 38, Number 9, Supplement II, pp. II-7-II-13)