Jacob Kean, PhD and Jamie Reilly, PhD

49

Item Response Theory

DEFINITIONS AND DESCRIPTIONS

- Item response theory (IRT): A statistical approach used to evaluate the quality of measures, such as patient-reported outcomes (PROs), clinical rating scales, questionnaires, surveys, and achievement tests (see Table 48.1 in Chapter 48 for contrast with classical test theory [CTT]).
- Latent variable: Constructs that in principle are "hidden" and cannot be measured directly.

INTRODUCTION

- Many variables of interest in medical and other clinical and behavioral research are latent variables.
- IRT methods aid in developing psychometrically robust latent-trait measures.
- This chapter describes the basic premise of IRT and highlights the key advantages of using IRT-based measures in research and clinical practice.

IMPLICATIONS

- Measures developed using IRT methods are quickly becoming the standard in rehabilitation research.
- The transition to IRT-based measurement in other areas of clinical research (eg, epidemiology) has been slower but has recently accelerated due to two developments.
    - Patient-Reported Outcomes Measurement Information System (PROMIS) effort, funded by the National Institutes of Health, to develop an accurate and efficient IRT-based measurement system for patient-reported health-related quality of life.
    - The publication of guidelines by the Food and Drug Administration regarding minimum quality standards for PROs that are used to support prescription labeling claims, which are arguably better met using IRT rather than CTT methods.
- IRT-based measurement offers significant advantages in statistical precision and power to detect effects.
    - IRT methods often require fewer subjects to detect treatment effects with comparable or even superior statistical precision relative to many other existing methods.
    - IRT can accordingly offer many advantages (eg, cost, statistical power) in the design and execution of clinical trials.

BACKGROUND

- IRT is often called a "modern" psychometric approach in contrast to CTT but its novelty is relative.
    - IRT originated decades ago with the work of Rasch in the 1960s and Lord in the 1950s.
- IRT models treat each encounter between an item and a respondent as a sort of a competition.
    - If the respondent answers correctly, then he or she wins the competition. If the respondent answers incorrectly, then he or she fails the competition.
    - After hundreds of respondents compete against dozens of items, the difficulty of the items is readily estimable, and all future encounters with new

respondents can be modeled as the probability of success on a given item.
- Importantly, IRT models can also be used to model rating scales where there is no correct response, only more or less of a given trait (eg, Likert scale ratings, such as, "never," "sometimes," "always").

■ The item-level (micro) focus of IRT models contrasts with the test-level (macro) focus of CTT. This difference in perspective has important implications for validity and reliability estimates derived from the two models.
- CTT establishes test-level validity that is compromised with modifications to the measure (eg, when a participant abandons early).
- In contrast, IRT uniquely validates each item, so items remain valid when only a subset of items from the validated pool is administered.
- Likewise, reliability in a CTT model is based on the entire test, whereas IRT reliability varies across the continuum with more precision at the center of the performance continuum.

■ Performance estimates produced by IRT models quantify the latent trait being assessed on an interval scale; CTT scales latent ability on an ordinal scale.
- Both interval and ordinal scaling allow comparison of different levels of a latent trait.
- However, the comparison of performance on an interval scale is a comparison of distance rather than the comparison of rank order using ordinal scaling.
  - For example, the results of a horse race can be reported as 1¼ mile times (ie, interval), as horse lengths of victory (ie, interval; approximately 8 feet), or as ordinal finish position listed nominally—win, place, show (ie, ordinal), or as a numerical order—1st, 2nd, 3rd (ie, ordinal).
  - Ordinally scaled nominal or numerical finish position tell you which horse was best, whereas interval scales such as finish times and lengths of victory tell you *how much better* the winner was compared with the other horses.

■ The difference between ordinal and interval scaling also applies to clinical and outcome measures.
- Measures developed with CTT provide a relative ranking of the latent trait being measured and tell you, for instance, which patient has more pain.
- In contrast, measures developed with IRT provide a "yardstick" that reveals, for instance, how much more pain a given patient has than another.
- Just as a yardstick is unaffected by the height of the persons being measured, measures developed with IRT are sample independent.
  - Their performance is unaffected by the group being measured.
- In contrast, the scale properties of measures developed with CTT vary across populations, which can be particularly problematic for the sometimes small and often heterogeneous populations encountered in many areas of clinical research.

STRATEGIES

■ One of the most important qualities of measures developed using IRT is the ability to deliver the measures *adaptively*.
- A computer adaptive test (CAT) takes advantage of the probabilistic IRT model to deliver only the items necessary to ascertain the level of the latent trait in a particular respondent, adaptively selecting the next item to administer based on the response to the previous item.
  - For example, if a respondent indicates on a physical functioning measure that he or she has difficulty standing from a seated position, the likelihood that this respondent can complete a marathon is exceedingly small.
  - Because only a few items (5–8) are needed to precisely target the latent trait of the respondent, there are no practical constraints on the number of items that can comprise a measure.
  - As a result, a CAT item bank may include hundreds of items to cover the full continuum of performance but administer only 5–8 items to precisely assess any given individual.
- The extensive coverage of the continuum and precise targeting of latent ability results in measures that are both more efficient and more precise than legacy measures developed with CTT.

■ IRT models make stronger assumptions of the data than do CTT models. Violation of those assumptions can lead to difficulties in the interpretation of results.
- For instance, the latent ability estimates generated by calibrating data to an IRT model assume the model actually fits the data being modeled.
  - IRT models incorporate a variety of robust indices at the item and model level to ensure fit of the data to the model that are more comprehensive than those available with CTT-based methods.

■ Measures delivered as CATs are custom individualized assessments that are often well-suited for clinical and research use.
- Increased precision achieved using CATs reduces the sample size necessary to detect an effect in clinical research.
- Accordingly, it is cost effective to incorporate CATs in clinical research whenever feasible.

■ IRT-based measures are typically amenable to parametric statistical approaches because the data are interval level.
- In contrast, nonparametric statistical approaches should be used to analyze CTT-based measures based on the ordinal nature of the data.

PITFALLS

■ A relatively large sample size is necessary to develop or refine measures using IRT models.
- Rough estimates of the sample size vary, depending on the IRT model chosen, but approximate minimum sample sizes for commonly used IRT models are as follows:
  - Parameter logistic (1PL; Rasch) models: 150 participants
  - Parameter (2PL) model: 500 participants
  - Parameter (3PL) model: 1000+ participants

HELPFUL HINTS

■ Though IRT has both practical and methodological advantages over CTT, many excellent measures have been created using CTT.

■ Perhaps the most powerful approach for latent-trait measure development is to use methods from both traditions, which may be seen as complementary rather than antagonistic.

SUGGESTED READINGS

DeVillis, RF. Classical test theory. *Med Care.* 2006; 44(11):S50–59.

Hambleton RK, Jones RW. Comparison of classical test theory and item response theory and their applications to test development. *Educ Meas Issues Pract.*1993;12(3): 38–47.

Hobart J. Rating scales for neurologists. *J Neurol Neurosurg Psychiatry.* 2003;74(Suppl IV):iv22–26.

RESOURCE

The Rehabilitation Measures Database (http://www.rehabmeasures.org) (see Chapter 48 for more details) Assessment Center℠ (http://www.assessmentcenter.net) is a free, online research management tool, which includes IRT-based PROMIS measures as well as IRT-based disease specific quality of life measures, such as the traumatic brain injury quality of life (TBI-QOL) and spinal cord injury quality of life (SCI-QOL) measures.