48

Jacob Kean, PhD and Jamie Reilly, PhD

Classical Test Theory

DEFINITIONS AND DESCRIPTIONS

- Classical test theory (CTT): 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 for comparison with item response theory, IRT)
  - Also known as true score theory
- Latent variable: Constructs that in principle are "hidden" and cannot be measured directly
  - Examples: Pain, depression, cognitive status, and rehabilitation outcome

INTRODUCTION

- Many variables of interest in rehabilitation and other medical and behavioral research are latent variables.
  - To empirically evaluate a latent variable, researchers make an oblique inference about the underlying construct.
- CTT methods have commonly been used for the development of latent-trait measures but are largely being supplanted or complemented by IRT methods.
- This chapter describes the basic premise of CTT and points out both the advantages and limitations of using CTT-based measures in research and clinical practice.

IMPLICATIONS

- There is new and increased attention on the importance of PRO as a key element of all aspects of clinical care and research.
- As the emphasis on the use of PROs and clinical rating scales has increased, so have expectations of their precision, refinement, and efficiency.
- Though CTT methods hold some advantages over IRT methods, the latter are generally regarded as superior and are the predominant approach to the development and refinement of measures in medical and other clinical and behavioral sciences.
- Knowledge of CTT remains important because the preponderance of legacy measures were developed with these methods, which have important statistical and practical implications.
- CTT methods can be used to complement IRT methods in the development and refinement of latent trait measures.

BACKGROUND

- CTT originated with the work of Spearman in 1904.
  - "Classical" in contrast to the more modern tradition of IRT, though both have been around for decades.
- At the heart of CTT is this idea:
  - A respondent's observed score (X) on an item or a set of items that comprise a measure is made up of his or her true score (T) and error (E).
    - Observed score: The estimation of a respondent's true score obtained by using a measure.
    - True score: The average of the observed scores obtained over an infinite number of repeated testing with the same test.

TABLE 48.1 Comparison of Classical Test Theory and Item Response Theory

	Classical Test Theory	Item Response Theory
Items	Characteristics not examined in detail	Primary focus of the analysis
Validity	Validity is based upon the total test and is invalidated with any changes to the measure	Validity is assessed for each item in the bank and remains valid with deletion of a subset of items
Reliability	Reliability is based upon the total test and is the same, regardless of ability	Reliability is calculated for each patient's "ability" and varies across the continuum, with more precision at the center of the performance distribution
Level of measurement	Ordinal	Interval
Ability estimate	Observed score is sample dependent	Ability estimate is sample independent
Assumptions	Weak, easy to meet	Strong, more difficult to meet

- The repeated tests are considered to be independent (to have no influence on subsequent tests), which is impossible in practice, so true score is a theoretical construct.
- Error: The discrepancy between an examinee's observed test score and his or her true score
  - Error comes from individual variability in examiner administration, idiosyncratic subject-level factors, such as fatigue and motivation at test time, and other factors.

- An advantage of CTT over IRT is the familiarity of CTT methods to a wide scientific audience.
  - Another is that some CTT statistical tests are commonly available in many popular statistical packages.
  - The weak assumptions made of the data by CTT are often cited as an advantage that makes CTT more widely applicable, as is the conceptual simplicity of the model (ie, $X = T + E$), though weak assumptions and conceptual simplicity have contributed to the lack of refinement in many legacy measures.
    - For example, the conceptually simple CTT model scales latent traits on an ordinal scale, whereas IRT scaling is typically interval (see Chapter 49).
- The focus of psychometric analysis in the CTT tradition is typically at the "test" (measure) level, in contrast with the item-level focus of IRT.
- The test-level focus of CTT has important consequences for measures developed in this tradition:
  - Measures must be used as they were validated because of measurement properties (ie, validity, reliability) of the scale are not imparted to the items themselves.
    - A subset of test items is a new, unvalidated measure.
  - Scale reliability, as computed by Cronbach's coefficient alpha, increases as more items are added.
    - Accordingly, measures developed with CTT can be unnecessarily long and cannot be shortened without validating the subset as a new form.
    - Another dominant reason that IRT is supplanting CTT is the decreased response burden (ie, shorter measures) afforded by IRT methods and computer-adaptive administration.

STRATEGIES

- Arguably the biggest factor that has contributed to the relative lack of refinement in some legacy measures developed using CTT methods is not the limitations of CTT methods but the failure of scale developers to carry out a comprehensive set of instrument development and evaluation approaches.
  - For example:
    - The construct and purpose of measures are often undefined.
    - The decisions to be made using measures are rarely explained.
    - The scoring of measures often ignores the dimensional structure of the construct.
    - The difficulty, discrimination, and differential functioning of items are rarely evaluated.
- The conceptual simplicity of CTT belies the difficulty of creating good measures.

PITFALLS

- The nature of ordinal scales produced by CTT limits their analysis and interpretation.
  - Ordinal scales are less powerful than interval scales because they provide only a rank-order of the strength of the attribute being assessed rather than a scalable number of interval units between points.
  - An example of the contrast between ordinal and interval measures is the result of a horse race: Ranking horses by finish position (eg, 1st, 2nd, and 3rd) is an ordinal measure, whereas finish time is an interval measure, providing both a rank order and the number of units (eg, seconds) between finish positions.
  - Ordinal scales do not, as a general rule, meet the assumptions of parametric statistical tests (eg, homogeneity of variance, normality of distributions).
    - Nonparametric, rank-order statistical approaches should be used to analyze CTT-based measures unless the assumptions of parametric statistical analyses are confirmed in the data set of interest.
  - Arithmetic operations (eg, addition, division, and square root) and descriptive statistics that depend on arithmetic operations (eg, mean, standard deviation) conducted with ordinal data cannot be interpreted because the numbers on ordinal scales are ordered labels, not numerical units (eg, it is meaningless to average places in a horse race).
  - Psychometric properties of CTT-derived measures, such as item difficulty, reliability, and standard error of measurement, vary across samples, which can be particularly problematic for the often small and heterogeneous samples encountered in many areas of clinical research.
  - Although ordinal data are informative, interval measures typically provide more valuable detail. Thus, the scaling of CTT typically yields less refined information relative to IRT models, and such specificity has acted as a driver toward IRT as a dominant measurement development paradigm.

HELPFUL HINTS

- Though IRT is gaining favor over CTT for both practical and methodological reasons, 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. *Edu. Meas.: Issues Pract.* 1993;*12(3)*;38–47.

Hobart J. Rating scales for neurologists. *J. Neurol Neurosurg Psychiatry.* 2003;*74(Suppl IV)*:iv22–6.

RESOURCE

The Rehabilitation Measures Database (www.rehabmeasures.org) is a resource designed to help clinicians and researchers identify reliable and valid instruments used to assess patient outcomes during all phases of rehabilitation. The database provides evidence-based summaries, administration and scoring instructions, and a representative bibliography with citations linked to PubMed abstracts, and includes the measure itself when possible.