Published in final edited form as:

Contemp Clin Trials. 2010 January ; 31(1): 1–3. doi:10.1016/S1551-7144(09)00212-2.

Classical and modern measurement theories, patient reports, and clinical outcomes

Rochelle E. Tractenberg

Building D, Suite 207 Georgetown University Medical Center, 4000 Reservoir Rd. NW
Washington, DC 20057

Classical test theory (CTT) has been widely used in the development, characterization, and sometimes selection of outcome measures in clinical trials. That is, qualities of outcomes, whether administered by clinicians or representing patient reports, are often describe in terms of “validity” and “reliability”, two features that are derived from, and dependent upon the assumptions in, classical test theory.

There are many different types of “validity”, and while there are many different methods for estimating reliability, it is defined, within classical test theory, as the fidelity of the observed score to the true score. The fundamental feature of classical test theory is the formulation of every observed score (X) as a function of the individual’s true score (T) and random measurement error (e):

$$X = T + e$$

CTT focuses on total test score – classical test theoretic constructs operate on the summary (sum of responses, average response, or other quantification of ‘overall level’) of items, individual items are not considered. An exception could be the item-total correlation (or split-half versions of this). The total-score emphasis of classical test theoretic constructs means that when an outcome measure is established, characterized or selected on the basis of its reliability (however estimated), tailoring the assessment is not possible, and in fact, the items in the assessment must be considered exchangeable. Every score of 10 is assumed to be the same. Another feature of CTT-based characterizations is that they are ‘best’ when a single factor underlies the total score. This can be addressed, in multifactorial assessments, with “testlet” reliability (i.e., the breaking up of the whole assessment into unidimensional bits, each of which has some reliability estimate). Wherever CTT is used, constant error (for all examinees) is assumed, that is, the measurement error of the instrument must be independent of true score. This means that an outcome that is less reliable for individuals with lower or higher overall performance does *not* meet the assumptions required for the interpretation of CTT-derived formulae.

CTT offers several ways to estimate reliability, and assumptions for CTT may frequently be met – but all estimations make assumptions that *cannot be tested* within the CTT framework. If CTT assumptions are *not* met, then reliability may be estimated, but the result is not meaningful. The formulae themselves *will* work; it is the *interpretation* of these values that cannot be supported.

IRT is a probabilistic (statistical, logistic) model of how examinees respond to any given item(s). Item response theory (IRT) can be contrasted with classical test theory in several ways; often IRT is referred to as “modern” test theory, which contrasts it with “classical” test theory. IRT is *NOT* psychometrics. The impetus of psychometrics (& limitations of CTT) led to the development of IRT. CTT is not a probabilistic model of response. Both the classical and modern theoretical approaches to test development are useful in understanding, and possibly “measuring”, psychological phenomena and constructs (i.e., both are subsumed under “psychometrics”). IRT has potential for the development and characterization of outcomes for clinical trials because it provides a statistical model of how/why individuals respond as they do to an item – and independently, about the items themselves. CTT-derived characterizations pertain only to total tests and are specific to the sample from which they are derived, while IRT-derived characterizations of tests, their constituent items, and individuals are general for the entire population of items *or* individuals. This is another feature of modern methods that is highly attractive in clinical settings. Further, under IRT, the reliability of an outcome measure has a different meaning than for CTT: if and only if the IRT model fits, then the items always measure the same thing the same way – essentially like inches on a ruler. This invariance property of IRT is its key feature.

Under IRT, the items themselves are characterized; test or outcome characteristics are simply derived from those of the items. Unlike CTT, if and only if the model fits then item parameters (and test characteristics derived from them) are invariant across *any* population, and the reverse is also true. Also unlike CTT, if the IRT model fits, then item characteristics *can* depend on your ability level (i.e., easier/harder items can have less/more variability).

Within IRT, unlike in CTT, items can be targeted, or improved, with respect to the amount of information they provide about the construct level(s) of interest. This has great implications for the utility and generalizability of clinical trial results when an IRT-derived outcome is used; and computerized adaptive testing (CAT) obtains responses to only those items focusing increasingly on a given individual’s construct (or ability) level. CAT has the potential to precisely estimate what the outcome seeks to assess while minimizing the number of responses required by any study participant. With IRT, tests can be tailored, or ‘global’ tests can be developed with precision in the target range of the underlying construct that the inclusion criteria emphasize or for which FDA labeling is approved.

IRT is powerful and offers options as clinical outcomes that CTT does not provide. However, IRT modeling is complex. The Patient Reported Outcome Measurement Information System (PROMIS, <http://www.nihpromis.org>) is an example of clinical trial outcomes that are being characterized using IRT. All items (for content area) are pooled together for evaluation. Content experts identify the “best” representation of their area – supporting test face and content validity. IRT models are fit by expert IRT modeling teams using all existing data, so that large enough sample sizes are used in the estimation of item parameters. Items that don’t fit the content, or statistical, models are dropped. The purpose of PROMIS is “To create valid, reliable & generalizable measures of clinical outcomes of interest to patients.” (<http://www.nihpromis.org/default.aspx>). Unevaluated in PROMIS – and many other- protocols is the direction of causality, as shown in Fig. 1. Using the construct “quality of life” (QOL), Fig. 1 shows that causality flows *from* the items (qol 1, qol 2, qol 3) to the construct (QOL). That is, in this example QOL is a construct that arises from the responses that individuals give on QOL inventory items (3 are shown in Fig. 1 for clarity/simplicity). The level of QOL is *not* causing those responses to vary, variability in the responses is causing the construct of QOL to vary. This type of construct is called “emergent” and is common. The problem for PROMIS (and similar applications of IRT models) arises from the fact that IRT models *require a causal factor* underlying observed responses, because conditioning on the cause must yield conditional independence in the

items. This conditional independence (i.e., when the underlying cause is held constant, the previously-correlated variables become statistically independent) is a critical assumption of IRT. QOL and PROMIS are only exemplars of when this causal directionality is an impediment to interpretability.

If one finds that an IRT model does fit the items (qol 1–3 in Fig. 1), then the conditional independence in those observed items must be coming from a causal factor; this is represented in Fig. 1 by the latent factor F; conditioning on the factor that *emerges* from observed items induces *dependence*, not independence. Therefore, if conditional independence *is* obtained, which is required for an IRT model to fit, and if the construct (QOL in Fig. 1) is *not* causal, then there must be another –causal – factor in the system (F in Fig. 1). The implication is that the factor of interest (e.g., QOL) is *not* the construct being measured in an IRT model such as that shown in Fig. 1 <in fact, it is F>. This problem exists – acknowledged or not – for *any* emergent construct such as QOL is shown to be in Fig. 1. Many investigations into factor structure assume a causal model, *all* IRT analyses assume this. Fig. 1 shows that, if the construct is not causal, then that which the IRT model is measuring is not only not the construct of interest, it will also mislead the investigator into believing that the IRT model is describing the construct of interest. Efforts such as PROMIS, if inadvertently directed at constructs like F rather than QOL, waste time and valuable resources and give a false sense of propriety, reliability, and generalizability for their results.

CTT and IRT differ in many respects. A crucial similarity is that both are models of performance; if the model assumptions are not met, conclusions and interpretations will not be supportable *and* the investigator will not necessarily be able to test the assumptions. In the case of IRT, however, there *are* statistical tests to help determine whether the construct is causal or emergent. Whether tested from a theoretical or a statistical perspective, IRT modeling should include the careful consideration of whether the construct is causal or emergent.

Further Reading

1. Bock, RD.; Moustaki, I. Item response theory in a general framework. In: Rao, CR.; Sinharay, S., editors. Handbook of Statistics. Vol. Vol. 26. The Netherlands: Elsevier; 2007. p. 469-513. Psychometrics.
2. Bollen, KA. Structural equations with latent variables. New York: Wiley; 1989.
3. Bollen KA, Ting K. A tetrad test for causal indicators. Psychol Methods. 2000; 5:605–634.
4. DeWalt DA, Rothrock N, Yount S, Stone AA. Evaluation of item candidates: the PROMIS qualitative item review. Med Care. 2007; 45(5, suppl 1):S12–S21. [PubMed: 17443114]
5. Embretsen SE, Reise SP. Item response theory for psychologists. LEA. 2000
6. Fayers PM, Hand DJ. Factor analysis, causal indicators, and quality of life. Qual Life Res. 1997; 6(2):139–150. [PubMed: 9161114]
7. Haertel, EH. Reliability. In: Brennan, RL., editor. Educational Measurement, 4E. Washington, DC: American Council on Education and Praeger Publishers; 2006. p. 65-110.
8. Jones, LV.; Thissen, D. A history and overview of psychometrics. In: Rao, CR.; Sinharay, S., editors. Handbook of Statistics. Vol. Vol. 26. The Netherlands: Elsevier; 2007. p. 1-27. Psychometrics.
9. Kane, MT. Validation. In: Brennan, RL., editor. Educational Measurement, 4E. Washington, DC: American Council on Education and Praeger Publishers; 2006. p. 17-64.
10. Kline, RB. Formative measurement and feedback loops. In: Hancock, GR.; Mueller, RO., editors. Structural equation modeling: a second course. Charlotte, NC: Information Age Publishing; 2006. p. 43-68.

11. Pearl, J. Causality: Models, reasoning and inference. Cambridge, UK: Cambridge University Press; 2000.
12. Sechrest L. Validity of measures is no simple matter. Health Serv Res. 2005; 40(5):1584–1604. part II. [PubMed: 16178997]
13. Wainer, H.; Bradlow, ET.; Wang, X. Testlet response theory and its applications. Cambridge, UK: Cambridge University Press; 2007.

Fig. 1.