

Proceed Boldly Yet Cautiously — Psychometrics in the Patient-reported Outcomes (PRO) World

Dorota Staniewska, PhD, Psychometric Statistician, Outcomes Research, Evidera

INTRODUCTION

Advanced psychometric techniques have been gaining ground in recent years in evaluation of patient-reported outcome (PRO) instruments.^{1,2} Properly applied, psychometric modeling (whether from the IRT or Rasch families) can provide unparalleled power in detecting non-functioning items, help define disease-specific outcomes and specify responder behavior. Misused, these methods can lead to wrong inferences about the population and the selection of inappropriate items for analysis.

The advantages parametric modeling provides to instrument development and population behavior are reviewed here, together with words of caution regarding indiscriminate application of the measurement theory models.

PROCEED BOLDLY!

Item response theory (IRT) defines patient responses to each individual item as a function of the patient's characteristic (latent trait) and the characteristics of the item (generally called discrimination and difficulty following educational measurement conventions). IRT is a powerful technique allowing for more in-depth understanding of the underlying population and item characteristics.

Because IRT has been used extensively in educational testing over the last 40 years, robust analytic techniques have been developed for most of the estimation problems. Unlike the Classical Test Theory techniques that describe patient performance in terms of domain or total score, considering all items to be equal, the IRT approach examines each item's contribution to the construct measured by the whole instrument. With IRT, given acceptable item fit, more information can be gleaned about the quality of measurement and, because person latent traits and item difficulties are on the same scale, an immediate check of whether these two are compatible is possible. In particular, the following issues have strong theoretical underpinnings:

1. Construction of new instruments with strong measurement properties;
2. Evaluation of the fit of each individual item to the measurement model chosen;
3. Evaluation of the statistical consequences of choosing some items over others for the instrument;
4. Evaluation of the relative merits of different instruments measuring the same trait;

5. Detection of the presence of potentially biased items; and
6. Detection of changes in latent trait across different evaluation times for subpopulations of interest.

IRT methods allow for collecting items measuring the same latent trait for building robust and

PROPERLY APPLIED, PSYCHOMETRIC MODELING (WHETHER FROM THE IRT OR RASCH FAMILIES) CAN PROVIDE UNPARALLELED POWER IN DETECTING NON-FUNCTIONING ITEMS, HELP DEFINE DISEASE-SPECIFIC OUTCOMES AND SPECIFY RESPONDER BEHAVIOR.

statistically valid item banks. In addition, they naturally provide a measurable degree of precision at every latent trait and, through item and test information, describe the degree of precision of both the individual item and the whole instrument at each level of latent



trait. This is paramount in developing parallel forms of instruments, what is especially salient in Computer Adaptive Testing where in-depth information about each item is necessary in order to pick the one most appropriate to the current estimate of the latent trait of the patient. Applying IRT techniques can also be useful at the development stage of the instrument when psychometric item fit and distractor performance can be examined in order to select items that best fit the population.

Additional techniques readily available when using psychometrics are Differential Item Functioning (DIF) and equating. Differential Item Functioning was developed as part of Classical Test Theory and then expanded by application of IRT methods. DIF helps to identify potentially biased items, i.e., items for which one subgroup (for example, males when DIF due to gender is being examined) scores differently (lower or higher) on the item than the other subgroup when controlled for the latent trait. As the population can be partitioned in many ways (for example, by gender, race, education, disease group division), this is a very powerful technique alerting

the researcher to problems with certain items, but more importantly, having the potential to further inform the instrument development process. Thus, DIF is quite useful in PRO development to examine for subgroup differences in responses for particularly heterogeneous patient populations, but also to provide quantitative measure of variabilities discovered during the qualitative phase of development (provided adequate sample is available).

Equating allows for patient latent traits (i.e., scores) obtained across different administrations of the instrument to be put on the same scale. In particular, while the follow-up version of an instrument might differ from the baseline version (through, for instance, the addition of new items), as long as the number of overlapping items is sufficient (30 to 70 percent, depending on the construct³), the IRT-based scale score from the two instruments can be directly compared with equating. This in turn allows for valid interpretations of any observed improvements in score. Another application of equating scale scores would be equating two different populations (e.g., the pediatric and adult cancer patients) so that they can also be directly compared.

PROCEED CAUTIOUSLY!

While software for analysis of instrument responses has been developed (e.g., RUMM, IRTPRO, Multilog, and even an experimental SAS procedure), both the setting up of the models and the interpretation of the output are not always as straightforward as they might seem and should be approached with care. In particular, standard normal distributions for the latent trait and the difficulty parameter are generally assumed and will be generally estimated; however, if this is not the case with the PRO (if, for example, the behavior is unipolar, like alcohol abuse disorder, or bimodal, like spinal muscular atrophy) care should be taken to set reasonable initial estimates of population statistics.

One should never forget that item response theory models come with strong parametric assumptions; all models have the assumption of unidimensionality (only one trait is measured by a collection of items), monotonicity (probability of a higher response increases with increased latent trait) and local independence (only the latent trait explains the performance on the item conditioned on it; the

responses are independent). While small deviations from the three assumptions are permissible^{4,5}, conspicuous violations of any of these assumptions result in faulty inferences about model fit and the violation of construct validity (i.e., what is measured by the instrument is no longer what was intended, and may, in fact, be impossible to ascertain). In the context of PRO instruments, this directly translates into the impossibility of interpretation of the significance of improvements in the score. If violations are suspected, either IRT is not appropriate for the scale or more advanced IRT approaches need to be employed (such as ones developed by Mark Reckase⁶ or Howard Wainer⁷).

Furthermore, a much larger sample size than for nonparametric analysis is needed in order to provide reliable estimates of thresholds. While the recommendation of the sample sizes varies^{8,9} and has not been systematically studied in the high-reliability PRO realm, generally, at least 300 patients per item is recommended.¹⁰ However, some authors¹¹ indicate that sample sizes exceeding 100 are sufficient for Rasch modeling of PRO data, while others¹² point to the number of items and variances of scores as being more consequential for estimation.

The Food and Drug Administration (FDA) encourages, but does not require, the use of IRT or Rasch analysis as part of the PRO instrument development and evaluation process for those PRO endpoints that will be used for product labeling.¹³ To the best of our knowledge, these psychometric analyses have generally resulted in more focused conversations and in the development of instruments more grounded in measurement theory. However, misusing the IRT or Rasch analysis can lead to inappropriate



inferences about both the items and the population of interest, especially if the local independence, unidimensionality and monotonicity assumptions are violated.

The Rasch-only approach to instrument analysis does have an immediately observable disadvantage. Because the same discrimination parameter is assumed and estimated for all items, item dependencies might not be immediately apparent, especially if residuals and residual correlations are not examined carefully. In more parameter-heavy models, item dependencies can be immediately assessed by unusual behavior of each item's discrimination parameter. In fact, the validity assumption of the same value of the discrimination parameter for every item should be carefully considered. As it is presumably somewhat impossible to ascertain the validity of this assumption from the content perspective, it is probably safer to check if discriminations are similar in the Generalized Partial Credit Model and the Graded Response Model.

Despite all the above caveats regarding the Rasch model, it needs to be stated that if the Partial Credit

Model fit is found to be comparable to any other psychometric model, it should be favored over other models because of its simplicity and relative ease of interpretation of output.

CONCLUSIONS

We are by no means claiming that this is a complete list of advantages of IRT and warnings about misapplying the models. We are hoping, however, that this article will give the reader both insight and pause about this exciting direction that PRO research has been taking over the last 10 years.


It is true that careful application of psychometric techniques will greatly inform the instrument development process and provide incredible insight into patient responses as a function of their disease severity. The blind application of these techniques, however, could result in faulty inferences and thus substantive misjudgments in the validity of the resulting instrument, and therefore potentially fatal conclusions regarding the trait measured and improvements in score.

We cannot stress enough that the presence of reliable estimates



for psychometric methods is not a guarantee of either content or construct validity of the instrument and will not compensate for failures in data collection, item phrasing, population misspecifications, etc.

The validity of instrument development still needs to hold, and methods to ensure this validity have been discussed in this forum before.^{14,15}

A careful examination of the data and its assumptions will ensure success with applying any model and lead to reliable and valid conclusions, resulting in a more powerful instrument being developed. 

For more information, please contact Dorota.Staniewska@evidera.com.

References

- ¹ Reeve BB, Hays RD, Bjorner JB, et al. Psychometric Evaluation and Calibration of Health-related Quality of Life Item Banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Med Care*. 2007 May; 45(5 Suppl 1):S22-S31.
- ² Cook KF, Keefe F, Jensen MP, et al. Development and Validation of a New Self-report Measure of Pain Behaviors. *Pain*. 2013 Dec; 154(12):2867-2876.
- ³ Kolen MJ, Brennan RL. Test Equating, Scaling, and Linking: Methods and Practices (Statistics for Social and Public Policy). Springer; 2004.
- ⁴ Harrison DA. Robustness of IRT Parameter Estimation to Violations of the Unidimensionality Assumption. *J Educ Behav Stat*. 1986; 11(2):91-115.
- ⁵ Ackerman TA. The Robustness of LOGIST and BILOG IRT Estimation Programs to Violations of Local Independence. Paper presented at the annual meeting of the American Educational Research Association; April 1987; Washington, DC.
- ⁶ Reckase MD. The Difficulty of Test Items that Measure More than One Ability. *Appl Psychol Meas*. 1985 Dec; 9(4):401-412.
- ⁷ Wainer H, Lewis C. Toward a Psychometrics for Testlets. *J Educ Meas*. 1990 Mar; 27(1):1-14.
- ⁸ Reise SP, Yu J. Parameter Recovery in the Graded Response Model Using MULTILOG. *J Educ Meas*. 1990 Jun; 27(2):133-144.
- ⁹ Forero CG, Maydeu-Olivares A. Estimation of IRT Graded Response Models: Limited Versus Full Information Methods. *Psychol Methods*. 2009 Sep; 14(3):275-299.
- ¹⁰ Kim S-H, Cohen AS. A Comparison of Linking and Concurrent Calibration Under the Graded Response Model. *Appl Psychol Meas*. 2002 Mar; 26(1):25-41.
- ¹¹ Chen WH, Lenderking W, Jin Y, Wyrwich KW, Gelhorn H, Revicki DA. Is Rasch Model Analysis Applicable in Small Sample Size Pilot Studies for Assessing Item Characteristics? An Example Using PROMIS Pain Behavior Item Bank Data. *Qual Life Res*. 2014 Mar; 23(2):485-493.
- ¹² Sébille V, Blanchin M, Guillemin F, et al. A Simple Ratio-based Approach for Power and Sample Size Determination for 2-group Comparison Using Rasch Models. *BMC Med Res Methodol*. 2014 Jul 5; 14(1):87.
- ¹³ U.S. Department of Health and Human Services. Food and Drug Administration; Guidance for Industry on Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. *Federal Register*; 2009 Dec; 74(235):65132-65133.
- ¹⁴ Vernon M, Wyrwich K. Methods for Selecting and Measuring Endpoints that are Meaningful to Patients in Rare Disease Clinical Development Programs. *The Evidence Forum*. Bethesda, MD, USA: Evidera; 2014 Mar:10-12.
- ¹⁵ Lenderking W, Revicki D. Clinician-reported Outcomes (ClinROs), Concepts and Development. *The Evidence Forum*. Bethesda, MD, USA: Evidera; 2013 Oct:36-40.