MAIMON WORKING PAPERS No. 22 NOVEMBER 2022

EVALUATING DISEASE SPECIFIC PATIENT REPORTED OUTCOMES MEASURES FOR RASCH INTERVAL MEASURE APPLICATION

Paul C Langley, Ph.D., Adjunct Professor, College of Pharmacy, University of Minnesota, Minneapolis MN

ABSTRACT

Previous commentaries in this Journal have made the case that the majority of polytomous patient reported outcomes instruments fail to meet the required measurement standard to support respondent profiling and response to therapy. Aggregating raw, integer scores from a bundle of Likert response items and presenting the raw score from each item response as an acceptable measurement metric fails because they ignore the subjective characteristic of the data, presuming also that the item responses have ratio, or at least, interval properties. Unfortunately, to achieve this with Likert thresholds for integer response, the assumption is that the relative value of each item response is the same and unit increases across thresholds have equal value for all items representing the same dimension for the theoretical concept. These are arbitrary assumptions that require careful analysis and verification. We need to take explicit account of relative item difficulty and the potential variation in thresholds of each item. In Rasch measurement theory these assumptions are recognized and can be verified using in the Rating Scale Model and the Partial Credit Model for construction and evaluation of polytomous patient reported outcome instrument to capture a unidimensional single attribute or latent construct. The fundamental point is that observations are always ordinal while measurements must be interval. If we are to claim that an instrument is producing a unidimensional interval measurement and not an ordinal score or raw number for an attribute, then we have to demonstrate that this is the case. The question for polytomous single attributes is whether or not they approximate in their raw scores to a unidimensional interval measure. Unless a case is presented to support the statement that the tool is generating unidimensional interval scale, the default assumption for polytomous PROs that present raw scores is that they are ordinal given they are representing the same concept. The purpose of this brief commentary is to set out challenge for polytomous instrument development and application: do they meet Rasch measurement standards for an interval score? A challenge that has to be addressed for the overwhelming majority of existing instruments if they are to be accepted to support value claims for therapy response and criteria that must be an integral part of instrument development.

INTRODUCTION: THE RASCH MODEL

Two propositions are key to recognizing the limitations inherent in disease specific patient reported outcomes (PRO) claims: (i) meaningful measurement to support value claims is based on the arithmetical properties of interval scales; and (ii) that Rasch measurement is the necessary and sufficient means to accept transformation of ordinal counts into linear measures [1] . Quantitative observations or count of events are not measures, they are merely nominal or ordinal numbers. If the observations or events are considered appropriate for quantification as counts, whether dichotomous or polytomous, then they must be justified as interval scores to support claims as measures. Counts are not measurement; to achieve measurement we have to create linear scales from these counts to support arithmetic operations. The Rasch model, first proposed in 1953, provides a complete, necessary and sufficient, solution for constructing PRO measures which retain their quantitative or calibration status irrespective of their application, while recognizing that that the measure must accommodate the interaction between the object to be measured and the measuring instrument [2]. The inherent unpredictability of this interaction led to a probabilistic interpretation when an

individual responds to an item in a questionnaire; the probability of positively responding to an item is a mathematical function of the difference between the item's relative difficulty and the ability of the respondent to realize that difficulty for a dichotomous response.

Where the response is polytomous, the Rasch model establishes the relative difficulty of each item stem recording the development of difficulty within that item as the rating scale has a number of thresholds and we need to model the likelihood of failure and success within each threshold [3] [4]. The polytomous Rasch Rating Scale Model estimates progression from one category to the next by estimating a single set of thresholds for each item; the Rasch Partial Credit Model drops this assumption as well as the assumption of the same number of response categories for each item; different numbers of response levels for different items for the same instrument with an increasing score representing an increase in ability or the value claim being assessed. Response patterns from a set of ordered items are tested against what is expected of the model as a valid or interval summed raw score [5] [6].

Although the Rasch model has been conceived and used widely in education, it has been noticed that it can be applied to establishing patient reported outcomes (PRO) instruments. Although in recent decades the number of applications of Rasch modeling in healthcare has been growing, it is still limited, due in part to the lack of awareness of the rules of fundamental measurement and the barrier to application due to the variety and the difficulty of accessing and using the various Rasch software packages. This is unfortunate as there are a few papers which have endeavored to provide an overview of Rasch standards, but with little in the way of a formal statement structure illustrating how to present and assess the various Rasch measures [5]. This is important because there is no single metric to receive a seal of approval. Rather the acceptance is holistic with a range of assessment criteria to be judged individually and in relation to each other for each PRO instrument to assess the underlying hypothesis.

The purpose of this brief commentary is to provide an overview of the role of Rasch assessment in polytomous PRO instruments where item responses are based on Likert scales with integer responses. These are the most common form of disease specific instrument, typically reported as a simple sum of integer values to yield ordinal scores. Summing integer values does not create a measure; it is an ordinal score and cannot be claimed as the basis for reporting therapy response is if it were an interval scale. The Rasch model provides the way in which raw scores can be transformed into unidimensional measures. This is achieved by an estimation process that iterates item estimates against person estimates, to converge to a preset convergence that maps discrete data (dichotomous responses) onto a real number line value as inferential measures based on probabilistic functions [3].

VALUE CLAIMS

In respect of PRO subjective value claims, the only acceptable claims are for unidimensional, single attribute claims that are consistent with the standards of fundamental measurement. In practical terms, this means that all disease specific instruments can only be accepted if they meet these requirements. In a recent commentary it was proposed that for a new start in health technology assessment the focus should be on value claims that accept two premises: [7]

- All value claims must refer to single attributes that meet the demarcation standards for normal science: they must be credible, evaluable and replicable
- All value claims must be consistent with the limitations imposed by the axioms of fundamental measurement: they must meet interval or ratio standards

Where a submission is to be made the relevant value claims that meet required measurement standards should be at the discretion of the formulary committee and appropriate to the target patient population in

the disease area. For the proposed new start, a critical feature is that all value claims must be accompanied by a protocol detailing how that claim is to be assessed and the results reported in a meaningful time frame; even if it is based on a previous pivotal clinical trial). Where a disease specific PRO value claim is presented, the assessment should include reporting the Rasch criteria for approximation to an interval scale.

RASCH MEASUREMENT: EVALUATING PATIENT REPORTED OUTCOME CLAIMS

The foundation for Rasch measurement for PROs is, in retrospect, the unsurprising claim that meaningful measurement requires interval scales and, where feasible, ratio scales with a true zero and interval properties. The Rasch measurement model provides the basis for creating dichotomous and polytomous instruments with interval scales and, for existing patient centric instruments, the application of performance criteria that allow the properties of an integer ordinal scale to be evaluated for potential interval standards.

This is a major task if we are to move from numbers to interval or even ratio measures [1]. Each PRO instrument should be evaluated to ensure that they refer to single attributes for a defined and credible latent trait or construct, meeting Rasch measurement standards, for both dichotomous and polytomous designs as single attribute constructs; that is, they have approximate interval measurement properties. Rasch statistics, in other words, assess the degree of accurate measurement. In disease specific instruments, the polytomous designs predominate. This means that they must be evaluated by either the Rating Scale Model or the Partial Credit Model. This is not a question of proposing a simple algorithm to translate raw ordinal scores to interval scores, but of a detailed assessment of the extent to which a proposed polytomous instrument meets fundamental measurement criteria for a single attribute interval scale. In other words, while the raw scores remain ordinal the instrument (either in its entirety or for sub-domains) can be interpreted in terms that it is consistent with Rasch requirements where each value of has a corresponding point on a Rasch continuum.

It is important to recognize what the assessment, utilizing one of the polytomous software tools (e.g., R, RUMM2030, WINSTEPS) is intended to achieve. The purpose is to evaluate the properties of the instrument; to test the hypothesis that accurate and useful measurement is evident in an instrument[6]. That is, the raw ordinal scale may be considered sufficiently approximate to an interval scale. This would not only apply to an existing instrument but also in developing a new instrument to ensure it approximates to an interval scale; which is the same objective in creating or assessing a dichotomous instrument. This is achieved by applying a number of criteria to the instrument to assess its internal functioning [5] [6]. These assessments should capture Rasch statistics (e.g., RUMM2030) for:

- Overall instrument and item functioning (reliability, individual item fit statistics, global model fit)
- Unidimensionality of underlying construct
- Local independence of items
- Categories and thresholds ordering
- Differential item functioning
- Person and item alignment

The judgement is holistic; which means it is important that a full range of statistical assessments for each of the criteria are presented and reasons for acceptance detailed. Presenting these assessment criteria is important because it presents third parties with the option of agreeing or disagreeing with the holistic or overall assessment that the hypothesis is reasonable in claiming approximation to an interval scale.; there is no magic transformation but a maximum likelihood estimation taking us from summary scores on terms to locations on a Rasch continuum.

As a final point: there is no necessary overall evaluation for a PRO with claimed sub-domains. Each sub-domain should be assessed separately, with possible rejection of some and acceptance of others; however,

even if sub-domains meet the criteria, there is no basis for attempting to add these together to create a multiattribute instrument. Each domain must be reported on separately; composite indices are not unidimensional [8].

CONCLUSIONS

After so many years of ignoring or failing to apply the established standards for Rasch measurement, the challenge facing disease specific polytomous PROs and their developers is to justify value claims for therapy response. If evidence to test the hypothesis that the raw ordinal score is a meaningful then there has to be a clear path from raw scores, to Rasch measures to Rasch derived utilities bounded within 0 - 1 limits, but which is still an interval score. Of course, there will be pushback; analysts are always more comfortable with known quantities. But they will face increasing rejection as formulary committees and other health system decision makers will insist on claims with PRO instruments exhibiting credible attributes from abstract entities such as quality of life and providing a coherent case for their acceptance as meaningful interval scores.

REFERENCES

[1] Wright B, Linacre J. Observations are always ordinal; Measurements, however, must be interval. *Arch Phys Med Rehab*. 1989; 70(12):857-60

[2] Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research, 1960

[3] Bond T, Yan Z, Heene M. Applying the Rasch Model: Fundamental Measurement in the Human Sciences. 4th Ed. New York: Routledge, 2021

[4] Andrich D, Marais I. A Course in Rasch Measurement Theory: Measuring in the Educational, Social and Health Sciences. Singapore: Springer, 2019

[5] Tennant, A, Conaghan P.The Rasch Measurement Model in Rheumatology: What Is It and Why Use It? When Should It Be Applied and What Should One Look for in a Rasch Paper? *Arthritis Rheumatism*. 2007;57, 1358-1362.

[6] Combrinck C. *Is this a useful instrument? An introduction to Rasch measurement models* in Kramer S et al. (Eds). Online Readings in Research Methods. *Psychological Soc South Africa*. Johannesberg, 2020. https://www.psyssa.com/wp-content/uploads/2020/08/Chapter-6_Is-this-a-useful-instrument_An-Introduction-to-Rasch-measurement-models.pdf

[7] Langley P. Nothing to Cheer About: Endorsing Imaginary Economic Evaluations and Value Claims with CHEERS 22 [version 1; peer review: 2 approved]. *F1000Research* 2022, 11:248 (https://doi.org/10.12688/f1000research.109389.1)

[8] McKenna S, Heaney A. Composite measurement in clinical research: the triumph of illusion over reality? *J Med Econ*. 2020;23(10):1106-1204