MAIMON WORKING PAPER 23 DECEMBER 2022

AFTER THE QALY? FAILING TO RECOGNIZE RASCH MEASUREMENT IN PROPOSING THE EQ-HEALTH AND WELLBEING (EQ-HWB) INSTRUMENT AS AN INTERVAL SCALE TO SUPPORT THERAPY RESPONSE CLAIMS AND QALYS

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

Abstract

The collective and continuing failure in health technology assessment (HTA) is the lack of awareness or unwillingness to come to terms with the requirements of fundamental measurement for evaluating therapy response in human subjects. This failure in HTA is one that, if not resolved, threatens to doom HTA to irrelevance. The failure can be attributed directly to a lack of understanding that measures have to be constructed empirically and experimentally, following well designed rules that transform an assessment of an attribute to construct a measure of that attribute. The unfortunate result is that if we set as our standard Rasch or modern measurement theory, we must reject all generic multiattribute preference measures and all disease specific patient reported outcome (PRO) measures unless the instrument can be demonstrated to have acceptable properties, in particular unidimensionality, order of item response and an interval scale. As detailed here, the Rasch model is the only framework to translate or capture structures in non-physical or latent attributes, and express these as interval measures. If not, we have numbers not measures. HTA has done its best to ignore Rasch measurement, it is never mentioned in best practice guidelines by groups such as the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) or in leading textbooks. The latest example of the unwillingness, or inadvertent failure, to address the issue of Rasch measurement as the only necessary and sufficient means to move from observations to interval measurement, is in the proposed successor to existing multiattribute instruments such as the EQ-5D-5L, the generic EQ-Health and Wellbeing (EQ-HWB) instrument. The purpose of this commentary is to make clear that in attempting to develop a more sensitive generic instrument, the EQ-HWB exemplifies the failure in instrument development that Rasch measurement addressed and resolved over 60 years ago. Once again considerable time and resource have been assigned to a patient reported outcome measures that is an analytical dead end.

INTRODUCTION

Assessments of therapy response, from the perspective of the patient or caregiver, have long neglected the standards of fundamental or Rasch measurement in health technology assessment [1] [2]. This is in evidence in both the continued focus and endorsement by academic groups, supported by journal editors, of generic multiattribute ordinal preference algorithms as well as in the misapplication of scoring algorithms in disease specific response claims; leading textbooks continue this false tradition [3] [4]. This failure to appreciate the limitations of fundamental measurement and the standards for Rasch measurement in value claims for competing pharmaceutical products and devices renders the majority of patient reported outcome (PRO) claims in HTA as both irrelevant and misleading. Unsurprisingly perhaps, this misbelief in measurement theory continues with the commitment of time and resources to creating a preference and hence QALY successor, the EQ-Health and Wellbeing (EQ-HWB) instrument, which fails Rasch measurement and the commitment to interval measures [5].

Criticisms of HTA in pointing to the failure to appreciate the need to subscribe to the standards of Rasch Measurement Theory (RMT) have been made by the present author, in a number of publications, many in this *Journal* [6 7]. This commitment is seen in the framework proposed and the commitment in this *Journal* to look beyond the quality adjusted life year (QALY) to a new start in evaluating response to therapy by patients and caregivers. This potential for a new start commitment is made clear in the Formulary Evaluation section of this *Journal* where it asks: *What are the standards of normal science, including fundamental measurement, that formulary committees should set and manufacturers should address in responding to requests for a formulary submission?* [8]. A pertinent question, given the dominant meme, not paradigm, in HTA is to endorse observations as measures (or failing to see the difference) [1]. Unfortunately for those advocating the role of HTA in therapy choice and resource allocation in health care, a mistake that was made clear in a seminal note by Wright and Linacre in 1989: *Quantitative observations are based on counting observed events or levels of performance. Meaningful measurement is based on the arithmetical properties of interval scales. The Rasch measurement model provides the necessary and sufficient means to transform ordinal counts into linear measures* [9]. It is salutary to recognize that the Wright and Linacre note not only summarized, at that time, 30 or more years of Rasch measurement, but preceded the development of both generic multiattribute instruments and the literally hundreds of disease specific instruments; and now as the *pièce de résistance*, the EQ-HWB ordinal scale.

The purpose of this brief commentary is to provide a non-technical overview of the contribution of the Rasch model to meeting the goal of fundamental measurement in patient reported outcomes (PROs), for both dichotomous and polytomous item response, and the failure to apply those standards in the EQ-HWB. The Rasch framework not only invalidates multiattribute generic preference or utility measures but the widespread application of Likert-based instruments for patient report outcomes (PROs) in disease areas and for target patient groups. The odd feature is that Rasch measurement has been accessible for almost 60 years, while the multiattribute generic instruments and the majority of disease specific instruments have been developed in the past 35 years; it is as though those subscribing to this failure in assessment and measurement have deliberately turned their backs on Rasch measurement; but without informing their audience of this challenge. It is perhaps indicative of this position that the leading textbook in HTA makes no mention of Rasch models; pursuing instead the pseudoscience of imaginary numbers [3].

A recent exchange in the *Journal of Medical Economics* makes clear not only the lack of competency in the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) checklist for instrument assessment but its claim to be a necessary gatekeeper for PRO outcomes 'seal of approval' [10]. This is seen in complete lack of appreciation in the concept of Rasch or simultaneous conjoint measurement; the limited understanding in HTA of the required measurement standards for PROs and the disregard demonstrated for the limitations imposed by the axioms of fundamental measurement as a necessary foundation for value claims.

Although COSMIN has been widely accepted for over a decade it fails for one reason: it takes the classical perspective of fitting the model to the data rather than selecting items, following conjoint simultaneous measurement and RMT, to fit the data to the model. In other words, if we follow RMT to create an interval scale for a unidimensional attribute or latent construct, then RMT fitting techniques determine the appropriateness or otherwise of the items to be selected to create the instrument. The COSMIN view is to ignore the limitations of fundamental measurement and use the PRO which is 'least worst' according to their checklist [11 12]. This is hardly a confident basis for selecting PROs and making claims for response to therapy, let alone the often noted casual and uninformed approach to developing instruments.

THE UNIQUE RASCH RULES

To date, the Rasch model provides the only framework for operationalizing latent constructs through assessment and the construction of single attribute measures with interval properties [13 14]. The Rasch framework aligns persons and items on the same scale; items are placed in order of difficulty and respondents in order of their ability to endorse an item. Multiattribute utility or preference scores must be rejected because they fail to identify specific attributes, they bundle them together, and then fail to apply a coherent model to yield an interval scale for evaluating response to therapy. Translating an interval to a ratio scale is problematic; which means that as interval scores cannot support multiplication, quality adjusted life years (QALYs) are mathematically impossible apart from the fact that preference or utility scores are multiattribute [15].

The Rasch rules for transforming dichotomous and polytomous ordinal observations and raw counts to measures interval measurement are well established with textbooks, video lectures, Excel spreadsheet demonstrations and software packages (RUMM2030, WINSTEPS, R). There is no excuse for failing to appreciate the rules for Rasch measurement and the importance of single attribute, unidimensional interval measures for patient (and caregiver) centric outcomes. Indeed, there are a large number of instruments developed over the past 20 or more years that explicitly follow Rasch rules for creating needs fulfillment quality of life measures. As well, there is an extensive literature that has evaluated existing disease specific instruments, typically with multiattribute domains and polytomous or Likert-based items, to assess whether or not the instrument is useful, given possible conformity with Rasch requirements, a real number line continuum while the raw scores are still ordinal. Unfortunately, while this is a stop-gap solution few instruments meet the required standards; in large part because their various sub-domains fail which render the application of the instrument invalid.

While ISPOR has shied away from endorsing a practice guideline to support the application of Rasch measurement standards in HTA, although recognizing in many papers in *Value in Health* that Rasch standards can be discussed and applied to assess patient reported outcomes instruments, the position is ambiguous. One explanation is that to whole-heartedly support Rasch as the required new measurement paradigm in HTA would cut the ground from under the belief system that supports HTA; the endorsement (including by journal editors) of the mathematically impossible QALY and the CHEERS 2022 guidance for constructing assumption driven modeled simulations to support imaginary (or approximate information) blanket claims for cost-effectiveness [16 17]. Apparently, where truth is consensus, you can have your cake and eat it.

THE FAILURE OF HEALTH TECHNOLOGY ASSESSMENT AND THE EQ-HWB

The failure of HTA to accept Rasch measurement can be easily stated: a failure to recognize that the only basis for evaluating response to therapy, given the standards for activities we label normal science, is the interval scale. Unless we can demonstrate that a measure has interval, or more problematically, ratio properties HTA fails. This has been made clear a in a number of commentaries in this *Journal*, notably is respect of the assumption driven modelled imaginary simulations to support ersatz cost-effectiveness claims by the Institute for Clinical and Economic Review (ICER) [18].

The failure of the EQ-HWB to meet Rasch standards and the eschew any commitment to single attribute interval measurement, means it is an analytical dead end. It will, undoubtedly be pursued, claiming to carry forward the unique legacy of multiattribute generic instruments and the the support for the redoubtable QALY, but it is really a charade.

As presently constituted, the EQ-HWB is a polytomous multiattribute questionnaire as a departure from the structure of the EQ-5D-3L and EQ-5D-5L instruments. Rather than responding to a set of five symptoms and response levels which are input to a single equation preference scoring algorithm (which yields negative utilities), the EQ-HWB comprises 32 items scored from zero to 5 integers, summed to give an integer count out of 100. The items comprise 5 to capture difficulty (How difficult was it for you …,) and the balance for frequency (I had problems sleeping). The items are intended to me more relevant to capturing quality of life (or health and wellbeing) across disease states based on themes and sub-themes from a literature revie 7 themes). Once item was selected from each theme (or domain) as a minimum for the instrument.

By the standards of fundamental measurement, the EQ-HWB produces only ordinal counts. It falls into the Likert-trap where each Likert item response is ordinal and the items ignore issues such as difficulty and the ability of the respondent to realize the item response level. Integers can be summed, but the result is just an integer count ranging from zero to 100. The choice of 100 is, presumably, to allow the integer score to be taken at face value as a preference score. Treated as a composite multiattribute instrument, the EQ-HWB forgets we have been here before with the generic Nottingham Health Profile (never mentioned) which included sub-domains, but rejected an overall score; merely ordinal scores for each sub-domain [19]. Given the NHP was developed in the late 1970s, it is a recognized step in the development of Rasch standard needs-fulfillment disease specific instruments.

The mirage, if that is a correct term, is the belief that assumption driven modelled simulations to generate claims for cost-effectiveness, should be more appropriately labelled as non-science as they fail the demarcation test [20] . Integral to this modelling is the role of utility or preference scores to create, by disease stage for a hypothetical population, QALYs and the attendant incremental cost-per-QALY claims and the application of cost-per-QALY thresholds. In the ICER business model, none of the claims are empirically evaluable, and were not designed to be. The objective is to create ‘helpful’ (or unhelpful) approximate information to support formulary decisions [21]. The modelling drives the demand for inputs to support these non-evaluable claims; hence the perceived need for more applicable and more sensitive preference instruments such as the EQ-HWB. The need to create preference scores, even if they are only ordinal numbers, drives the search for improved ordinal multiattribute observations; the QALY tail wags the preference dog. This wagging mapping opportunity incudes, not only the EQ-HWB, but the many papers reporting mapping by regression models, from disease specific ordinal scales to ordinal preference scores, encouraged by ISPOR practice guidelines [22]. The fact that it is impossible to map from non-interval scales to create other scales is not a question that seems to have been addressed.

CONCLUSIONS

Judged by the standards for Rasch measurement, the phrase 'After the QALY' is a misnomer; it implies, unjustifiably, that there is some merit to the QALY, and the contribution of multiattribute preference score in its role as a useful construct. As such, it is viewed as an ultimate end product to the stepping stones, first the EQ-5D-3L, then the EQ-5D-5L and now the EQ-HWB as successively improved instruments to support the QALY. This misses the point: if measurement is to have any meaning to capture therapy response, then the measure must have single attribute, unidimensional, linear and interval properties. Certainly, as entities, 'quality of life' and 'quality adjusted life year' are of interest, but it is the properties of these that are to be measured not the entities themselves; properties variously described as traits, constructs or traits. If a trait is to be measured then, as Andrich points out, we have to have a controlled procedure, an indirect assessment of observations or responses, to manifest the property of interest. Importantly, these observations or responses must have an order to capture more or less of the property to be assessed [13]. This scoring, the application of rules, is the essence of Rasch measurement with the assignment of integers, initially to qualitative putatively ordered responses to assessments which can then be transformed into quantitative measurement.

The upshot is quite clear: Rasch rules are only applicable where single attributes are being assessed. We might focus on needs-fulfillment as a property of quality of life that is important; if so, we need to apply an established set of rules to transform a qualitative assessment to a quantitative measure, ensuring that the end product is an approximation to an interval scale. In Rasch terminology, the required data elements or items for questionnaire responses are fitted to the Rasch model for, in this case, a needs fulfillment instrument in target patient populations or disease areas [14]. Our hypothesis is, therefore, that in applying such an instrument the objective is to assess the extent to which, for a population with a known distribution of abilities, and items of increasing difficulty, the probability that a respondent will more successfully respond to items given the application of that therapy.

This is the essence of Rasch or modern measurement: the transformation of assessments. The probabilistic Rasch model is confirmatory in fitting the data and predictive of response. It has decades of application in education, psychology and, to a limited extent, PRO instrument development in HTA; it is not a successor but a well-established framework which is unique, necessary and sufficient for transforming observations to measurement that is antecedent to the pursuit of the analytical dead end of ordinal scores to support approximate information modeling that was in place before thought was given to ordinal multiattribute preferences and QALYs.

REFERENCES

[1] 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

[2] Langley PC and McKenna SP. Measurement, modeling and QALYs [version 1; peer review: 2 approved] *F1000Research* 2020, 9:1048

[3] Drummond M, Sculpher M, Claxton K et al. Methods for the Economic Evaluation of Health Care Programmes (4th Ed.). New York: Oxford University Press, 2015

[4] Facilitating bias in cost-effectiveness analysis: CHEERS 2022 and the creation of assumption-driven imaginary value claims in health technology assessment [version 1; peer review: 3 approved]. *F1000Research* 2022, 11:993

[5] Brazier J, Peasgood T, Mukuria C et al. The EQ-HWB: Overview of the development of a measure of health and wellbeing and key results. *ValueHealth*. 2022;25(4):482-491

[6] Langley P. Mapping Impossible Utilities: The ICER Report on Tezepelumab for Severe Asthma. *Inov Pharm*. 2022;13(2): No. 1

[7] Langley P. Concerns with Patient Reported Outcome Measurement and Value Claims for Therapy Response: The Case of Mavacamten and Symptomatic Hypertrophic Cardiomyopathy (SHCM). *InovPharm*. 2022;13(2): No. 16

[8] Innovations in Pharmacy. Formulary Evaluations https://pubs.lib.umn.edu/index.php/innovations/section/view/formularyevaluations

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

[10] McKenna S, Heaney A. Setting and maintaining standards for patient-reported outcome measures: can we rely on the COSMIN checklist? *J Med Econ*. 2021;24(1):502-11

[11] Mokkink L, Terwee C, Bouter L et al. Reply to concerns raised by McKenna and Heaney about COSMIN. *J Med Econ*, 2021;24(1):857-89.

[12] McKenna S, Heaney A. COSMIN reviews: the need to consider measurement theory, modern measurement and a prospective rather than retrospective approach to evaluating patient-based measures. *J Med Econ*. 2021;24(1):860-61

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

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

[15] Langley P. The Great I-QALY Disaster. *InovPharm*. 2020; 11(3): No 7

[16] Husereau D, Drummond M, Augustovski F et al. Consolidated Health Economic Evaluation Reporting Standards 2022 (CHEERS 22) Statement: Updated reporting guidance for health economic evaluations. *ValueHealth*. 2022;25(1):3-9

[17] Husereau D, Drummond M, Augustovski F et al. Consolidated Health Economic Evaluation Reporting Standards (CHEERS) 2022 Explanation and Elaboration: A Report of the ISPOR CHEERS II Good Practices Task Force. *ValueHealth*. 2022;25(1):10-31

[18] Langley P. Nonsense on Stilts – Part 1: The ICER 2020-2023 value assessment framework for constructing imaginary worlds. *Inov Pharm*. 2020;11(1):No. 12

[19] Hunt S, McKenna S, McEwen J et al. The Nottingham Health Profile: subjective health status and medical consultations. *Soc Sci Med A*. 1981;15(3 Pt. 1): 221-229

[20] Pigliucci M. Nonsense on Stilts: How to tell science from bunk. Chicago: University of Chicago Press, 2010

[21] Neumann P, Willke R, Garrison L: A Health Economics Approach to US Value Assessment Frameworks – Introduction: An ISPOR Special Task Force Report. *ValueHealth.* 2018; **21**: 119–123

[22] Wailoo A, Hernandez-Alava M, Manca A et al. Mapping to estimate health-state utility from non-preference based outcome measures: An ISPOR Good Practices for Outcomes Research Task Force Report. *ValueHealth*. 2017;20:18-27