MAIMON WORKING PAPER 24 DECEMBER 2022

A CONTINUING DISASTER: THE EQ-HEALTH AND WELLBEING (EQ-HWB) INSTRUMENT

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

ABSTRACT

It is difficult not to underestimate the attraction of multiattribute patient reported outcome (PRO) instruments; they have been the mainstay for assumption driven cost-per-quality adjusted life year (QALY) modeled imaginary simulations to support blanket claims for cost-effectiveness. Concerns with the coverage and sensitivity of the two most widely used GENERIC multivariate instruments, the EQ-5D-3L and the EQ-5D-5L, has led to the development of the EQ-Health and Wellbeing (EQ-HWB) instrument., While different in design to the preceding instruments and supported by the Euro-QoL Foundation, the hope is that this will be a worthy successor to support QALY modeled claims for resource allocation and policy interventions in health care systems. This is an unrealizable ambition. Just as the preceding multiattribute instruments fail to meet the long accepted Rasch measurement standards for transforming ordinal counts to linear interval measure to evaluate response to therapy, the EQ-HWB fails on precisely the same criteria. The failure stems from the fixation with the need to support a gold standard QALY to capture and value health status as a single metric with ratio properties, bounded by zero and unity. Just as the commitment to assumption driven simulations is seen as a mirage of non-evaluable approximate information claims, the insistence on multiattribute preferences numbers and the impossible QALY are equally fallacious; a ratio generic scale is impossible. The purpose of this note is to make clear the importance of Rasch measurement for patient-centric claims in health technology assessment, demonstrating why the EQ-HWB is an analytical dead-end and a monumental waste of time and effort.

Keywords: EQ-HWB, Rasch rules, multiattribute failure,

INTRODUCTION

It may come as a surprise to those advocating multiattribute instruments such as the EQ-5D-5L as the vehicle for establishing ratio preference scores to create QALYs, but the required standards for interval (not ratio) measures were in place some 70 years ago. The contribution of Georg Rasch, a Danish statistician, established that if the object is to create an interval measure to capture response to therapy (or the response to ordered items in a test of mathematical; attainment) then we have to consider the interaction between the ability of the respondent and the difficulty of an item [1]. If not, as Wright and Linacre, eloquently expressed in a seminal 1989 paper, we have no option in patient reported outcomes (PROs) but to remain with raw observations, counting observed events or levels of performance [2]. Unless a PRO claim, generic or disease specific, meets Rasch measurement standards it must be, by default, an ordinal count or just a raw score.

The purpose of this note is to make abundantly clear that the proposed EQ-HWB instrument fails to meet the required measurement standards to assess response to therapy as an interval scale; a composite multiattribute instrument which fails measurement standards [3] [4]. This should come as no surprise as the EQ-HWB follows in the steps of equally invalid composite multiattribute instruments, the EQ-5D-3l and EQ-5D-5L. As it stands, the case for this new 'measure', as a raw score, is to extend benefits of treatments beyond health related quality of life (HRQoL) into the areas of social care and public health, to include independence or improved relationships with friends, families and caregivers [5] [6]. The EQ-HWB was jointly funded by the EuroQol Group Foundation and the UK Medical Research Council. If it is to be promoted as an 'improved' multiattribute foundation for preference scores and QALYs, then such a promotion is a waste of time; just as its predecessors failed to meet required Rasch measurement standards for patient reported outcomes (PRO) instruments, so has the EQ-HWB in both its long (25 item) and short (9 item) forms. The EWQ-HWB, however, is not intended to replace the EQ-5D-5L, but as a complementary instrument.

RASCH MEASUREMENT FOR THERAPY RESPONSE

The standards and rules for Rasch measurement are well established and have been recognized for over 60 years; except, unfortunately, by the many authors and agencies that have advocated generic and disease specific PRO instruments. Indeed, the belief system or meme that is represented by advocates of assumption driven modelled simulations, makes no mention of the critical role of Rasch measurement [7]. While the results of Rasch assessments of existing PRO instruments have not been neglected by leading textbooks and journals such as *Value in Health* and the *Journal of Medical Economics*, any discussion, let alone commitment, to the imperative of the Rasch framework is absent.

The Rasch approach to creating meaningful invariant interval scale measurement, a number that supports the range of arithmetical operations with a prior calibrated linear measurement system for parametric statistical analysis, rests on the necessity of rules to transform ordinal counts to interval measures. This applies equally to measurement in the physical sciences as it does to subjective responses with patient or caregiver centric outcomes; any calibration must be specific to defining the items that comprise an instrument or test for a single attribute and not a bundle of attributes that might, for example, be defining a health state. Hence the importance in the application of Rasch transformation rules distinguishing between measurement and assessment; the process by which we move from an entity, such as quality of life, to the selection of properties, attributes, constructs, variables or traits that are to be measured [8]. This is not a direct measure of, say, a latent trait, but the manifestation of the particular property of that latent trait we wish to measure in order to assess response to an external stimulus. These responses are typically qualitative; the Rasch framework transforms these to a qualitative interval measure. We may, for example, have a polytomous item-based instrument; this generates integer scores and the summation or a count of those scores; Rasch provides the rules by which we transform those ordered ordinal counts to interval measurement.

At the same time, the unique contribution of Georg Rasch was to recognize that in patient-centric outcome assessment the observed response has to be seen in probabilistic terms: what is the probability of a successful response by a respondent to a questionnaire item? The answer is to

establish rules to capture the fact that the response is due to the interaction between respondent ability to realize successful responses and the difficulty of the item [2]. There is, therefore, a distribution of respondent abilities interacting with items ordered in terms of the relative difficulty. Hence the critical assessment of the property to be explored; to define that latent property as a credible single attribute and assess its indirect manifestation by application of Rasch rules. This is achieved by an iterative process to generate an approximation to an interval scale (Rasch continuum); the criteria applied (which are accessible in a number of software packages that have been on-line for over 30 years include RUMM2030, WINSTEPS and R) are:

- 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 scores on items to locations on a Rasch continuum.

The Rasch rules (or model) are the only possible way for transforming counts to interval (and possibly) approximate ratio measures; as Wright and Linacre make clear: *The Rasch measurement model provides the necessary and sufficient means to transform ordinal counts into linear measures* [2].

THE DICHOTOMOUS RASCH MODEL

To illustrate the importance of Rasch modelling to create an interval measure, consider the dichotomous Rasch model, the first model developed in the 1950s. In this model the assignment of 1 to a correct response to an item and zero to an incorrect response is not the equivalent of assigning 0 or1 to a respondent gender, a nominal differentiation, but recognizes these as ordinal data where the meaning attached to 1 is greater than that attached to zero. The unit response is not only different from the zero response but is superior to it; a correct response to an item allows us to impute order where 1 represents more of an attribute than 0; the respondent is more able to realize the correct response. This gives us the Rasch rule that for any item the lowest code represents the lowest level of the of the latent variable, which holds for polytomous responses. This rule is not recognized in the responses to the EQ-HWB where, for example, the item ‘I feel exhausted’ proceeds from the lowest value ‘none of the time’ to the highest value ‘most or all of the time’. The latent construct ‘exhaustion’ is reverse ordered.

Order is important as the basis for the item pathway and the fit of the item to the Rasch model. The raises the key to differentiating the Rasch model from item response theory (IRT) and true score theory (TST). In the case of IRT and TST, the data have primacy, with psychometric claims

merely explore and describing the data (an unknowable true score and a random component) while in the Rasch model we confirm its measurement status and its predictive or probabilistic nature. IRT and TST must account for all the data while the Rasch model requires or confirms that the data fit the model. This ensures that we can claim that the measure is consistent with conjoint simultaneous measurement such that we can claim that the resulting scale has invariant, interval properties. As Wright expressed it: *Rasch models are the only laws of quantification that define objective measurement, determine what is measurable, decide which data are useful, and expose which data are not* [9].

MULTIATTRIBUTE POLYTOMOUS INSTRUMENTS

Once the unique contribution of Rasch measurement to transform ordinal observations to interval measures is recognized, the failure of multiattribute instruments, both generic and disease specific, becomes apparent. They fall at the first hurdle: their ordinal counts are not measures. Add to this is fact that, by design, they are composite instruments. There was no intent, possibly by design, to recognize the necessity of defining response in terms of single attributes.

The proverbial cat escapes from the bag by the use of the term 'multiattribute', in the case of generic instruments such as the EQ-5D-3L, EQ-5D-5L and EQ-HWB, but also the single score instruments such as the time trade off (TTO) and the standard-gamble (SG). All are based on the belief that bundles of symptoms and response levels can be assigned, as raw counts or integer responses, to the category of an interval or ratio measure. There is no concept, which would be inapplicable, of the application of rules to transform raw scores or ordinal counts to an interval scale; a transformation that recognizes the interaction between item difficulty (for a single attribute) and the ability of the respondent. The Rasch model provides the required rules, but these play no part in the development of the EQ-HWB; an application which is impossible given the needed commitment to a multiattribute scale as a successor offering. The Rasch rules apply only to the transformation of raw scores to an interval measure where the intent is to develop a measure for a single attribute.

This failure of multiattribute instruments is difficult to comprehend given the recognized and applied Rasch rules at the time, in the 1980s, when these instruments were being proposed and, in the decades since, the creation of multiattribute disease specific instruments. Judged from the standards of Rasch measurement, this is a difficult to imagine waste of time and resources. The answer is, again unfortunately, not difficult to find: the need for quality adjusted life years (QALYs) to justify blanket claims for cost-effectiveness. Claims which still resonate as the gold standard in health technology assessment; epitomized in the fixation on assumption driven modelled lifetime simulations to create incremental cost-per-QALY claims which are, by design again, non-evaluable as demonstrated in leading textbooks [10] . The QALY rests on the perceived need to 'value' health states; single attribute evaluable clinical, PRO, drug utilization and resource utilization claims are just not acceptable. In short, Rasch has no role; the standards of modern measurement, for the current meme of health technology assessment, are an unfortunate distraction which is best ignored. Empirical evaluation of claims, the test for demarcation between science and non-science is a distraction; the fact that, on this criterion, assumption driven simulated modelled claims are best characterized as non-science or pseudoscience is irrelevant [11]

If the objective is to create QALYs, even in the guise of modelled approximate information, there is the issue of the application of the multiattribute algorithm and the resulting so-called preference score. This ordinal scale is characterized by those who believe in QALY models, such as the Institute for Clinical and Economic Review (ICER) where the QALY model is their principal business case, not just as an interval measure but as a ratio measure; a measure with a true zero although the various algorithms produce negative values or states worse than death. The ratio measure property is essential: to create a QALY you need to multiply time spent in a disease state by a ratio score with a bounded 0 – 1 property to discount to the equivalent of time with perfect health. This is no possible if the 'preference' score is ordinal; the QALY is impossible [12]. Yet, groups such as ICER persist because they 'have confidence' that health economists have faith that the preference score has ratio properties; no proof is provided [13]. It is just a question of unsupported belief.

THE EQ-HWB DISASTER

The genesis for the EQ-HWB was the increased dissatisfaction with the limited symptom and response coverage of the EQ-5D-3L and EQ-5D-5L; applied to specific disease states there was seen to be a need to 'bolt-on' additional symptoms and response levels. The caveat, of course, is that the new mix of symptoms and response levels had to represent a compromise between their application relevance across a variety of disease states and the need to minimize respondent burden; the same concern that had plagued the developers of the EQ-5D-3L and the later increase to 5 response levels with the EQ-5D-5L. The result is a 25-item polytomous instrument (with a 9-item short form) which, like its progenitors, was again multiattribute with a disregard, by design or ignorance, of Rasch measurement. In other words, we are still locked into defining quality of life in composite health symptom terms (HRQoL) where health states are defined in terms of the response level for 25 items.

There are two parts to the EQ-HWB: (i) a 5 item symptoms difficulty section (how difficult was it to see, to hear, to get around; to engage in activities and to care for yourself and (ii) a 20-item symptom frequency response section; all responses are on a 5-response Likert scale with four thresholds for each item. All items refer to experience over the past 7 days.

Considerable effort went into item options with the final selection with qualitative reviews of the literature and focus groups. The result was the classification of 32 subthemes grouped into 7 high level themes: feelings and emotions, cognition, self-identity, autonomy, relationships, physical sensations and activity. It is not clear if these should be considered attributes; if so, then following Rasch they should be separately transformed from raw scores to interval measures. Even so, the EQ-HWB has clinical symptoms as the focus, following from the five difficulty symptoms identified in the EQ-5D-3L/5L: mobility, self-care, usual activities, pain/discomfort and anxiety/depression. An HRQoL mindset that is apparently impossible to escape from if driven by generic considerations for a gold standard to support resource allocation within health systems, even if the QALY is an impossible mathematical construct.

But this last point is not necessarily a constraint; the reference here is to the Nottingham Health Profile (NHP) instrument [14]. Developed in the mid-1970s. Responses are in a dichotomous format as opposed to the polytomous EQ-HWB instrument, but item selection following the assembly of

2200 statements from 700 people, eventually reduced initially to 138 statements and then to 82. The final questionnaire comprised 38 statements with problems with health categorized into: sleep, physical mobility, energy, pain, emotional reactions and social isolation. Within each category the items were weighted by severity using Thurstone's methods method of paired comparisons from a sample of the general public. The weights reflect the severity of the item from the patients' perspective. Each section is scored out of 100. But this is where the NHP parts company from the EQ-HWB: the focus is on the health profile as a population measure; there is no aggregate score. These six sections comprised Part 1 of the instrument; Part 2 comprises scores on seven outcome scores each defined by a statement. While the NHP did not apply the Rasch framework (it was early days before Rasch packaged software modelling), there are two points to emphasize: first, the object was to create a score for each section to create a profile (which does not ask directly about symptoms); and second, it was recognized that any overall or aggregate score was incompatible with a profile that provides a measure of patient perception 'as a direct reflection of need and possible demand'. As a profile it made no sense to consider an overall score, whether sections were weighted or unweighted. The EQ-HWB puts this aside, with no intention of providing profiles, the question is then how these various item responses are to be aggregated to a single, weighted or unweighted, raw score.

Even though the NHP does not apply Rasch rules, there are too few items in each section for even a retrospective assessment, this is clearly a precursor to the focus on single attributes, properties of the more abstract concept of the quality of life, in disease specific needs fulfillment Rasch models by authors associated with the NHP development to capture patient value in PRO instruments [15 16 17]. The focus is on the single attribute by disease area from the view that life gets its quality from the extent to which needs are met; a single attribute capturing through intensive interviews the needs appropriate to that disease or target patient group and the transformation to an interval scale.

AN ANALYTICAL DEAD END

Irrespective of claims that the EQ-HWB has obvious merits, to its developers, as a successor to the multiattribute EQ-5D-3L and EQ-5D-5L, the exercise is clearly one which fails to appreciate the imperative of Rasch rules to support the transformation from raw scores to an interval scale. There is nothing which, at this stage, might be recovered. The failure of the EQ-HWB was determined as soon as the decisions was made to create a multiattribute instrument with no thought given to the required measurement properties: an interval or ratio scale. Nor was any consideration apparently given to the recognized limitation on adding integer values from Likert scales; the need to assume that all items are of equal difficulty and that the thresholds between the assigned integer values are of equal value or distance. If not, then all that is achieved is an ordinal summation of ordered integer values; a raw score. This is all that the EQ-HWB has achieved; the failure to recognize the importance of Rasch rules to transform raw scores to interval measures defining response for single attributes.

It is worth noting that the EQ-5D-3L and EQ-5D-3L multiattribute instruments must also be judged failures in terms of the absence of required measurement properties. While they fail by reason of their attempt to combine bundled symptoms and responses to create health states which, by judicious choice (or fishing for a best fit) of a modeled algorithm, raw scores are collapsed to

create a so-called preference score, this score lacks meaning. It certainly falls at the first hurdle to meet the required Rasch standards, focusing on the chimera of combing what are described as clinical attributes, but the resultant preference score lacks any coherence as a basis for transforming to an interval, let alone a ratio scale.

CONCLUSION

If we are concerned to evaluate response to therapy then we have to meet the standards of Rasch measurement; it is the only measurement model thar provides the necessary and sufficient means to transform ordinal counts or raw scores to unidimensional, interval linear measures for credible properties of a latent construct. If we are to combine the difficulty of an item with the ability of the respondent to predict the probability of a successful response then there is no option but to focus on disease or target patient populations where response is defined in terms of specific attributes. The assessment must conform to Rasch requirements and reported as part of any defense of the instrument with acceptable approximation to an interval score for single attribute.

Multiattribute measures are not acceptable. Just as we could, presumably, see a role for a gold standard QALY, constructed from a defensible (by Rasch standards) preference score with ratio properties, we are asking for the impossible. The EQ-HWB is like stepping back in time to a medieval world of non-science or pseudoscience. If the developers had wanted to make every possible mistake in modern measurement theory, they have certainly succeeded. In the case of the EQ-HWB, there is a well-worn phrase*: If you are in a hole, stop digging.*

REFERENCES

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

[2] Wright B, Linacre J. Observations are always ordinal; measurements, however, must be interval. *Arch Phys Med Rehabil*. 1989; 70(12):857-60 https://www.researchgate.net/publication/20338407_Observations_are_always_ordinal_measurements_however_must_be_interval

[3] 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

[4] McKenna S, Heaney A. Composite outcome measurement in clinical research: the triumph of illusion over reality. *J Med Econ*. 2020;23)10):1196-1204

[5] EuroQoL Group. EuroQol is developing a new instrument- The EQ-HWB. February 16, 2021 https://euroqol.org/eq-5d-instruments/eq-hwb/

[6] National Institute for Health and Care Excellence. A new instrument for consideration of a broader range of benefits for people, their families and carers. 17 February 2021 https://www.nice.org.uk/news/blog/a-new-instrument-for-consideration-of-a-broader-range-of-benefits-for-people-their-families-and-carers

[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

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

[9] Wright B. Fundamental measurement for psychology. In Embretson S, Hershberger S (Eds). The new rules of measurement: What every educator and psychologist should know. Mahwah, NJ: Lawrence Erlbaum Associates, 1999

[10] 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

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

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

[13] Langley P. To Dream the Impossible Dream: The Commitment by the Institute for Clinical and Economic Review to Rewrite the Axioms of Fundamental Measurement for Hemophilia A and Bladder Cancer Value Claims. *InovPharm*. 2020;11(4): No. 22

[14] Hunt S, McEwen J, McKenna S. Measuring health status: A new tool for clinicians and epidemiologists. *J Royal College General Practitioners*. 1985, 25:185-88

[15] McKenna S, Wilburn J. Patient value: its nature, measurement, and role in real world evidence studies and outcomes-based reimbursement. *J Med Econ*. 2018;21(5):474-80

[16] McKenna S, Heaney A, Wilburn J et al. Measurement of patient reported outcomes. 1:The search for the Holy Grail. *J Med Econ* 2019;22(6):516-22

[17] McKenna S, Heaney A, Wilburn J. Measurement of patient-reported outcomes. 2: Are current measures failing us? J *Med Econ*. 2019;22(6):523-30