MAIMON WORKING PAPERS No. 1 JANUARY 2023

NONSENSE ON STILTS: FUNDAMENTAL MEASUREMENT AND THE GENERIC EQ-HEALTH AND WELLBEING (EQ-HWB) INSTRUMENT

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

ABSTRACT

The illusory belief that composite multiattribute ordinal preference scores have a central role in health technology assessment to support assumption driven simulated cost-effectiveness claims has been a staple over the past 30 years. This is seen in the application of cost-per-QALY claims in assumption driven modeled simulations to support pricing and access recommendations and in the recent CHEERS 2022 guidance for creating imaginary claims. The standards of fundamental measurement are not recognized in health technology assessment. The result is that the widely used EQ-5D-3L and EQ-5D-5L multiattribute instruments fail the required standards in fundamental measurement. They fail to recognize the need to translate observations or raw scores by application of Rasch rules to justify disease specific, single attribute, unidimensional interval scales. This failure is set to continue with the commitment to creating a more comprehensive multiattribute instrument the EQ-Health and Wellbeing (EQ-HWB) questionnaire. Although destined to be both a complement to and a more acceptable successor to the EQ-5D-5L as a generic 25-item polytomous instrument, including claimed continuing support for preferences and QALYS, the EQ-HWB is a nonstarter; it fails completely the Rasch standards for measuring therapy response; it is nothing more than a subjective ordinal raw score. The purpose of this commentary is to deconstruct the EQ-HWB and demonstrate why, in failing to meet the Rasch rules as the necessary and sufficient means to translate ordinal counts to interval measures. Judged by the required standards for fundamental measurement it is a composite generic will o'the wisp; driven by the commitment to a single metric to capture therapy response that is an analytical dead end.

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

INTRODUCTION

There is only a limited awareness of the critical role played by Rasch measurement theory in translating subjective observations to true measurement. This is clearly apparent in health technology assessment where the focus on developing both generic and disease specific patient reported outcomes (PRO) instruments has proceed in an apparent disregard or denial of Rasch measurement. The result is that all generic instruments and the overwhelming majority of disease specific PRO instruments yield only ordinal scores [1] [2]. This is doubly unfortunate: first, because

ordinal scores, even if presented as ordered scales, say nothing about response to therapy for which you need ordered interval measures, but also (ii) because the quality adjusted life year (QALY) is in consequence a mathematically impossible construct as preference scores cannot support multiplication. As we cannot combine time spent in a disease stage by an ordinal preference score to create a claimed time spent in perfect health equivalent, one of the foundations of health technology assessment, the assumption driven modeled lifetime simulation to create incremental cost-per-QALY claims ceases to have any meaning [3 4]. The fact that simulation modeling also fails to mee the standards of normal science for credible, evaluable and replicable claims, driven by a belief in the future realism of assumptions, merely adds to the dismissal of this belief or meme in approximate information [5].

Concerns with the coverage and sensitivity of the two most widely used multiattribute generic instruments the EQ-5D-3L and Eq-5D-5L resulted, for the latter, an increase in the number of response levels for each of the five symptoms from three to five responses. Introduced in 2009 as a successor to the EQ-5D-3L, the EQ-5D-5L has had a checkered history due to the apparent incompatibility of the two instruments to produce consistent scores for the same patient population. Irrespective of the fact that both yield only ordinal preference scores, with a high proportion of negatively valued health states worse than death, the EuroQol Group continued to promote the EQ-5D-5L and it retained its acceptance as the key input to creating the mathematically impossible QALY for simulated model claims.

In the mid-2010s it was decided to seek a generic successor that could go beyond the health related quality of life (HRQoL) 5-symptom and 5-response level approach to defining and creating health states for subjective valuation and, as a presumed outcome with a ratio preference score, to create a QALY measure than was suitable for use across health and social care [6 7]. Funded jointly by the EuroQoL Research Foundation and the UK Medical Research Council the result is the EQ-Health and Wellbeing (EQ-HWB); a 25-item polytomous (Liker scale based) instrument as a standardized measure of aspects of health and wellbeing (with a 9-item short form) for patients and caregivers [8 9] . It is promoted as a separate instrument and not as a replacement for the EQ-5D duo.

Unfortunately, despite the efforts devoted to creating the EQ-HWB, it is still a multiattribute instrument that generates only ordinal scores. As such, it cannot capture response to therapy or support QALY claims. The purpose of this review is to make clear the reason for this debacle: a failure to appreciate the difference between observations and measurement and the unique contribution of Rasch measurement to establishing well defined and accepted rules for creating the required ordered interval measure for disease specific therapy assessment; rules that have been recognized for the past 60 years. We have 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. 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. 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 PROs the authors of the EQ-HWB dutifully follow suit.

OBSERVATION VERSUS MEASUREMENT

Following Merbitz et al, Wright and Linacre stress the importance of recognizing that all observations start as ordinal, if not nominal data [10] [2]. The choice is to determine whether these observations are worth counting, whether they are dichotomous (presence, absence) or polytomous (ordered or Likert rating scale) while recognizing in the latter case that an order says nothing about the distance between the classification of the ordered categories. This is not measurement. To employ a measure means to transform these original observations to a calibrated system with a well-defined origin and unit; a linear scale which can be shown to be useful. Once this is achieved, we can apply statistical analysis. Simply adding integer values achieves nothing other than a raw score which can only support non-parametric statistics.

The recognition of the need to progress from counting observations to measurement can be traced back over a century with partial techniques, such as Thurstone's paired comparisons, with a complete solution presented by Rasch in the 1950s [11] [12] Rasch's solution is unique; *it is not only necessary but also sufficient for the construction of measures in any science* [2]. Rasch's insight was obvious yet profound: (i) a measure must maintain its linear calibrations irrespective of what it is measuring so that items are invariant in their level of difficulty and respondents to that item must retain their same level of ability regardless of the item encountered in that set of calibrated items that define the variable or trait being studied; and (ii) the interaction between the respondent and the attribute being measured must involve the likelihood of a successful response such that the more able the respondent the greater the likelihood of a success on any relevant item [2]. The Rasch model imposes on the data a single underlying unidimensional variable or attribute; measurement requires unidimensionality. In practical terms, the Rasch rules provide a basis for justifying an acceptable approximation to the ideal of a unidimensional linear scale [2]. Initially developed to explore dichotomous response instrument the extension to polytomous responses instruments was resolved by the early 1980s [1]. These applications are made readily available by software packages which analyzed the initial data set in terms of the potential creation of a single latent variable with unidimensional calibration, their reliability and the fit of observations to the Rasch model standards (e.g., RUMM2030, WINSTEPS [13] [14]).

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 [15]. 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.

Two points should be emphasized if we are to advocate a new start in PRO value claims: (i) all PRO claims must be for single, well defined and credible attributes and (ii) all PRO claims must be based on instruments that have been shown to be developed following Rasch rules for transforming subjective raw scores to unidimensional, interval measures [5] [16]. The EQ-HWB fails

on both criteria. There is no excuse for putting Rasch standards to one side, unless there is a belief, not articulated, that a multiattribute instrument does not require translation to an interval scale. The EQ-HWB is designed to capture, in 25-tems (9 items for a short form) 7 high level themes: feelings and emotions, cognition, self-identity, autonomy, relationships, physical sensations and relationships. An initial list of 687 candidate items was successively reduced to a final set of 25 items to constitute the generic health and wellbeing instrument. 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.

It must be emphasized that 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 of a distinction 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 potentially identified and hopefully measured for that entity [17]. 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 quantitative interval measure. We may, for example, have a polytomous item-based instrument; this generates integer scores and the opportunity or ability to provide a summation or a count of those scores; Rasch provides the rules by which we transform those ordered ordinal counts to interval measurement. Any notion of this need to transform to an interval scale with Rasch rules as the necessary and sufficient condition is absent in the EQ-HWB.

At the same time, the unique contribution of Rasch was to recognize that in patient-centric outcome assessment the observed response has to be seen in probabilistic predictive 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 their relative difficulty. Hence the application of Rasch rules to assess critically the property to be explored; to define that latent property as a credible single attribute.

The application of Rasch analysis provides a formal test of an outcome scale against a measurement model, operationalizing the formal axioms that underpin measurement: these axioms of additive conjoint measurement are the only rules and will determine whether ordinal or interval scales have been constructed [18]. The Rasch model takes precedence with the items selected determined by the model. There is no recognition in the description of the development of the EQ-HWB that there are axioms of fundamental measurement that have to be followed; simply adding

integer values from a polytomous instrument such as the EQ-HWB without any regard for the need to recognize the need for the interval measurement is never mentioned.

Application of the axioms of conjoint measurement is achieved by an iterative process to generate an approximation to an interval scale (Rasch continuum); the criteria applied in a number of software packages (RUMM2030, WINSTEPS) that have been available (on-line at low cost) for over 35 years are [18]:

- 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 (polytomous instruments)
- 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.

There is no excuse, apart from a fixation on creating a multiattribute instrument to create a single generic metric, for ignoring Rasch rules. Unfortunately, the contribution of multiattribute instruments to support a composite metric for claims for therapy response is illusory [9]. Two factors are critical in measurement: (i) accurate measurement of a latent construct requires unidimensionality, where all items measure that construct; and (ii) dimensional homogeneity where comparison is possible only for variables that have the same dimension.

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 anyway for health states, of the application of rules to transform raw scores or ordinal counts to an Rasch continuum single attribute unidimensional 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 which is still focused on the subjective or ordinal valuation of health states, defined by bundles of symptoms and response levels.

UNIQUE STATUS OF RASCH MEASUREMENT

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 or 1, a nominal differentiation. The Rasch model recognizes these as ordered 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.

Order is important as the basis for the item pathway and the fit of items to the Rasch model. This 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 exploring 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 or items proposed fit the model [1]. The EQ-HWB is clearly in the former camp. There is no concept of selecting polytomous items that meet the Rasch model order requirements; rather, items were selected by voting and affirmation (involving even colored score cards) while ensuring that at least one item was elected for each domain in the 25-item version, with 1 or 2 items retained. This procedure is the antithesis of the application of Rasch rules as the necessary and sufficient basis for creating items where responses yield an approximation to interval measures.

It is important, therefore, to differentiate Rasch measurement from IRT which focuses on the probability of expected response as a function of the ability of a person and parameters characterizing the item; an item characteristics curve as a function of a latent trait. This certainly has similarities with the Rasch model given the notion of latent trait modelling with some describing the Rasch model as a one-parameter IRT model. This misses the point of Rasch measurement; Rasch rules address the strict standards of scientific measurement with the application of conditional maximum likelihood estimation to construct fundamental measures [1]. A unique paradigm, as Andrich describes it, where a mismatch between a model and the data is not viewed as a problem with the model but with the data [19] . The Rasch model does not try to fit or describe the observed data (maximizing variance explained) but requires the data to fit the Rasch model to create a unidimensional linear or interval measure. The fit of the data to the Rasch model provides justification for claiming that we have created a required interval single attribute measure [1] ; 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* [20].

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 with the unique Rasch requirements.

THE QALY FIXATION

This willing choice of nonstarter 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, it is difficult to imagine a greater waste of time and resources in a PRO instrument development. 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 [15]. 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. Empirical evaluation of claims, the test for demarcation between science and non-science as such is a distraction; the fact that, on this criterion, assumption driven simulated modelled claims are best characterized as non-science or pseudoscience is irrelevant [21].

For those who believe in the mystery of the QALY as a gold standard with global application to drive resource utilization within health systems, the rejection amounts, in effect, to apostasy. From a relativist perspective, truth is consensus supported by rhetoric, persuasion and authority; the multiattribute QALY is an essential construct, irrespective of the absence of meaningful measurement properties [22]. Hence the willingness of analysts to engage in developing and endorsing the EQ-HWB. Endorsement that extends to groups such as ISPOR and single payer health gatekeepers such as the National Institute for Health and Care Excellence (NICE) in the UK. The QALY is an indispensable construct; without the QALY reference models and cost-per-QALY thresholds are misapplied and irrelevant. In the US, the standard bearer for the QALY is the Institute for Clinical and Economic Review (ICER) which holds to the odd, yet false belief, that health economists have confidence that preference scores supporting the QALY have ratio properties [23] . ICER's position is understandable as the QALY model is their principal business case; their position on the EQ-HWB is unclear as it offers an alternative ordinal preference measure and presumably more impossible QALYs.

AMAZING GRACE

Indicative of the widespread failure to recognize the failure of multiattribute instruments to produce other than ordinal scores have been attempts to modify these scores to adjust the mathematically impossible cost-per-QALY thresholds to account for disease severity with disability and provide adjustments to capture diminishing returns to health. A recent example is the Generalized Risk-Adjusted Cost-Effectiveness (GRACE) framework which attempts to challenge the assumption that returns to HRQoL never diminish. In economics the concept of diminishing returns is well established and has been for some 200 years; typically applied in the analysis of consumption and production. In the latter, we can consider a production function which combines factor inputs (e.g., units of capital and labor): diminishing returns are the decrease in marginal units of output as a single factor is increased with other inputs held constant that will occur at a particular point. Certainly, we can consider the creation of health as similar to a

production function, where increased application of a particular unit of health input, other inputs held constant, yield a diminished marginal contribution to overall health.

The problem lies in the units of measurement of overall health. In economics units of input and output are measured in interval or ratio terms. Health is another matter because the units to capture health benefit, subjective multivariate preference scores, QALYS and cost-per-QALY thresholds fail to meet standards of fundamental measurement. Attempting, as the GRACE framework proposes, to generalize existing assumption driven simulation model cost-effectiveness claims to incorporate diminishing returns to health improvements as disease severity increases is impossible. Not only do these cost-effectiveness models, as noted above, fail the standards of normal science but they also fail to recognize the imperative of Rasch measurement. It seems pointless, given their manifest deficiencies, to propose adjustments to such a framework. We are not concerned and must avoid bundling assumptions, adjusted for disease severity and attitudes to risk, into a single meaningless health state metric. The focus must be on transforming observations to interval measurement defined for single, unidimensional attributes within defined target patient populations to support single attribute and empirically evaluable value claims.

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 [24]. The caveat, of course, is that the new mix of symptoms and response levels had to represent a compromise between their application or 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. 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 based on qualitative reviews of the literature and with focus groups. The result was the classification of items by 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 evaluated or transformation 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.

There appears to have been a lack of awareness of the question of fundamental measurement in the development of the EQ-HWB where item selection had to follow Rasch rules. If there had been then the exercise could have been put to one side in favor of a focus on single attribute,

unidimensional value claims that could have been deemed to be relevant as a 'common core' of value claims where each met the required measurement standard [25]. As it is, irrespective of any justification for the choice of individual items, bundling them together to create a 'new' multiattribute (or multi-item) scale or raw score, means that such considerations (e.g., as a 'bolt on supplementary dimension') are irrelevant [9].

It is of interest to consider a generic instrument, not mentioned in the EQ-HWB paper that recognizes the need to single attributes, eschewing a composite multiattribute scoring; the Nottingham Health Profile (NHP), developed in the late 1970s (and, of interest, also funded by the UK Medical Research Council) [26]. 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 [27] [28] [29]. 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.

A NEW START IN HEALTH TECHNOLOGY ASSESSMENT

The failure of the EQ-HWB was determined as soon as the decisions was made to create a re-badged 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 [1]. 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.

If we are to assess response to therapy then we have no alternative but to follow the Rasch rules and attempt to capture the attribute of interest as a unidimensional linear interval measure. There is no need (or sense) to try and combine attributes into a single metric that lacks unidimensionality and construct validity. Where there are a range of attributes associated with therapy response then we have to present a profile for clinical, PRO, drug utilization and other resource utilization empirically evaluable value claims for review by a formulary committee each of which, if applicable, meets Rasch standards; a new start in health technology assessment [5] . This provides, supported by attribute assessment protocols, the framework for the empirical evaluation of claims, ongoing disease area and therapeutic class reviews to track value claims and even outcomes-based contracting.

Protocols, it must be emphasized, are as essential part of the new start proposal with the focus on evaluable value claims. In this respect it is worth noting that there have been a number of proposals for protocols to support real world evidence claims assessment. The most recent of these is the Harmonized Protocol Template to Enhance Reproducibility (HARPER) as a good practices task force report by ISPOR and International Society for Pharmacoepidemiology (ISPE) to assess treatment effects by enhanced transparency and reproducibility [30] . Unfortunately, as with all ISPOR good practice reports, there is no mention of fundamental measurement and the need to demonstrate that the value claim for the treatment effect object to be assessed must be based on a single attribute, unidimensional interval or ratio scale. This, of course, would open the question of the basis on which the value claims were developed; certainly not as any element of the ISPOR-favored assumption driven, simulated, modeled cost-effectiveness CHEERS guidance framework which lacks empirically evaluable claims [5]. HARPER is, presumably, to be limited to clinically evaluable interval or ratio measured claims that by-passes or skirts around blanket claims for cost-effectiveness. Surprisingly, there are no questions raised to demonstrate how HARPER might support outcomes-based contracting and its relevance for ongoing disease area and therapeutic class reviews, let alone the imperative of the standards for fundamental measurement.

If we seek a guidance framework for value claims that meet Rasch measurement standards then we have a candidate in the recently released Rasch Reporting Guidelines for Rehabilitation Research (RULER) [31 32]. These are not a recent innovations; studies reporting Rasch measurement in the context of rehabilitation outcomes were first reported in 1988 (with some 109 studies reported by 2019). The purpose of RULER is to provide peer-reviewed, evidence-based and consistent guidance for reporting studies that apply Rasch measurement theory in a rehabilitation context so that there are uniform expectations on how to write and evaluate research on rehabilitation outcomes assessments. RULER stands in market contrast to CHEERS 2022 and, although not referenced by ISPOR or CHEERS 2022, it is a template that should be applied across disease areas if we are to evaluate Rasch-based claims for therapy response. Such a commitment should be seen as an essential part of a new start in health technology assessment.

RULER is not alone in the need for guidance in the recognition of Rasch standards in reporting model claims. Although essentially ignored in mainstream HTA, it is worth noting an earlier paper that considered the role of the Rasch measurement model in rheumatology for resolving issues of

when Rasch analysis should be used and what should be reported in any Rasch analysis [18]. These criteria can, of course, be applied if there is interest in whether an existing PRO instrument is sufficiently robust to meet Rasch standards [33]. This is, again of course, a stopgap measure; but the contrast with the CHEERS 2022 guidance is salutary [34]. It is absurd to attempt to claim, *ex post facto*, and attempt to demonstrate by application of Rasch criteria, that an instrument defined by a multiattribute set of items, such as the EQ-HWB, has unsought Rasch measurement standards; it is pointless to even try to undertake such an exercise.

If we remain committed to the importance of a generic multiattribute framework with application across disease states then the focus could be on an NHP -type framework where there are 'core' attributes that comprise the PRO attributes of interest, with a claimed profile of interval measures. It is doubtful if this would gain much traction as the attributes chosen would have to be reassessed following Rasch criteria for each new disease area or target patient population application; a comparison of profiles without the option of a single composite metric. Once the impossibility of a single composite HRQoL type metric is recognized, then the solution is to focus on each disease state, with attributes selected relevant to the treatment outcomes defined as clinical measures, PROs and resource utilization terms. As it stands with the present focus on composite or multiattribute claims is competition between the EQ-5D3L, EQ-5D-5L and the EQ-HWB as the 'preferred' metric for PRO claims and putative resource allocation.

While not considered in the development of the EQ-HWB is the fact that we have ample experience with the application of Rasch rules in applied measurement. There are a range of single attribute, Rasch modeled disease specific instruments quality of life defined in needs fulfillment terms that have been developed over the past 25 years for a range of disease states by Galen Research [35] which have been widely applied and with of the latest contributions the Alzheimer's Patient Partners Life Impact Questionnaire (APPLIQue) [36 37]. These instruments provide the basis with their interval scores for translating an interval scale into a 'bounded' ratio scale to assess response to therapy and, if required, quality adjusted need fulfillment life year estimates analogous to the ordinal preference scores of multiattribute instruments [38].

CONCLUSIONS

The will o'the wisp of multiattribute models supports the belief that there must be, or there has to be, a single metric that collapses those multiattribute items of interest to a single preference score with ratio properties. Thus, the QALY is proposed as collapsing mortality and morbidity into a single measure to support a unique single claim for cost-effectiveness, created by assumption driven modeled simulations with non-evaluable outcome claims. This is clearly at odds with the Rasch model and its support for single attribute, unidimensional measurement; the only acceptable form of measurement. The application of Rasch rules to transform observations to a linear unidimensional measure makes clear that the Rasch model represents a unique and unassailable required measurement paradigm.

Once the Rasch requirements are recognized, the case can be made that the EQ-HWB fails as a measure; it is nothing more than a raw or ordinal score with item responses that, however aggregated, fail to support claims for therapy response. This failure is quite unsurprising; at no

time apparently in the process of developing the EQ-HWB was there any recognition of the Rasch model and its application for PRO instruments with interval measurement properties.

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 a single attribute. Multiattribute measures are not acceptable. The EQ-HWB is like stepping back in time to a medieval measurement world of non-science. If the developers had wanted to make every possible mistake in modern measurement theory, they have certainly succeeded, repeating the same lack of awareness as exemplified by the measurement failures of the EQ-5D-3L and EQ-5D-5L. As the EQ-HWB is owned by the EuroQoL group, they now have three strikes against them

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

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

[4] Langley PC and McKenna SP. Measurement, modeling and QALYs [version 1; peer review: 2 approved]. *F1000Research* 2020, **9**:1048 (https://doi.org/10.12688/f1000research.25039.1)

[5] 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)

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

[7] 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

[8] 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

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

[10] Merbitz C, Morris J, Grip J. Ordinal scales and the foundations of misinference. *Arch Phys Med Rehab*. 1989;70:308-32

[11] Thurstone L. A method for scaling psychological and educational data. *J Educ Psychol*. 1925;15:433-51

[12] Rasch G. Probabilistic Models for some Intelligence and Attainment Tests. Copenhagen: Danish Institute for Educational Research, 1960.

[13] RUMM2030: Rasch Measurement Tools for Research and Education https://www.rummlab.com.au/

[14] WINSTEPS: Software for Rasch Measurement and Rasch Analysis https://www.winsteps.com/index.htm

[15] 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

[16] McKenna S, Heaney A, Langley P. Fundamental Outcome Measurement: Selecting Patient Reported Outcome Instruments and Interpreting the Data they Produce. *InovPharm*. 2021; 12(2): No. 17

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

[18] 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 (Arthritis Care & Research)*. 2007;57(8):1358-62s \
No. 13

[19] Andrich D. Understanding resistance to the data-model relationship in Rasch's paradigm: a reflection for the next generation. *J Appl Meas*. 2002;3(3):325-59

[20] 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

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

[22] Wootton D. The Invention of Science: A New History of the Scientific Revolution. New York: Harper Collins, 2015

[23]Langley P. Peter Rabbit is a Badger in Disguise: Deconstructing the Belief System of the Institute for Clinical and Economic Review in Health Technology Assessment. *InovPharm*. 2021; 12(2): No.20

[24] Finch A, Brazier J, Mukuria C. Selecting bolt-on dimensions for the EQ-5D: Examining their contribution to health related quality of life. *ValueHealth*. 2019;22:50-61

[25] Langley P. Evidentiary Standards for Patient-Centered Core Impact (PC-CIS) Value Claims. *InovPharm.* 2022;13(3): No. 15

[26] 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

[27] 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

[28] 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

[29] 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

[30] Wang S, Pottegard A, Crown W. HARmonized Protocol Template to Enhance Reproducibility of Hypothesis Evaluating Real-World Evidence Studies in Treatment Effects: A Good Practices Report of a joint ISPE/SPOR task force. *ValueHealth.* 2022;25(10):1663-1672

[31] Mallinson T, Kozlowski A, Johnston M et al. Rasch Reporting Guidelines for Rehabilitation Research (RULER): the Ruler statement. *Arch Phys Med Rehab*. 2022;103:1477-86

[32] Van de Winckel A, Kozlowski A, Johnston M et al. Reporting Guideline for RULER: Rasch Reporting Guideline for Rehabilitation Research: Explanation and Elaboration. *Arch Phys Med Rehab*. 2022;103:1487-98

[33] Combrinck C. Is this a useful instrument? An introduction to Rasch measurement models in Kramer S, Laher A, Fynn et al (Eds.) Online readings in Research Methods. Psychological Society of South Africa. Johannesburg 2020

[34] Langley P. Rasch Measurement and Patient Reported Value Claims: A Primer for Hemophilia. *InnovPharm*. 2023; 13(4): No. 13

[35] Galen Research, Measures Database https://www.galen-research.com/measures-database/

[36] Hagell P, Rouse M, McKenna S. Measuring the impact of caring for a spouse with Alzheimer's disease: Validation of the Alzheimer's Patient Partners Life Impact Questionnaire (APPLIQue). *J App Measurement*. 2018;19(3):271-282

[37] McKenna S, Rouse M, Heaney A et al. International development of the Alzheimer's Patient Partners Life Impact Questionnaire (APPLIQue). *Am J Alzheimer's Disease and Other Dementias*. 2020;35:1-11

[38] Langley P. McKenna S. Fundamental Measurement: The Need Fulfilment Quality of Life (N-QOL) Measure. *InovPharm*.2021;12(2):No. 6