MAIMON WORKING PAPER No. 8 JUNE 2023

AFTER THE QALY: MEASUREMENT AND THE ROAD NOT TAKEN (PART I: THE EQ-HWB)

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

Abstract

In 2020, the journal <u>Innovations in Pharmacy</u> issued a challenge for commentaries on what might be the next steps given the mounting criticisms of the application of generic multiattribute instruments to create preference scores to create claims based on quality adjusted life years (QALYs). In the years since a number of developments in multiattribute instrumentation and the creation of preference scores have come to, or are close to fruition, to provide what are viewed as successor multiattribute QALY constructs. The purpose of this first brief commentary is to make the case that these are essentially analytical dead ends; they perpetuate the essential flaw in the original multiattribute QALY and any multiattribute patient reported outcome instrument, that they fail to meet Rasch or modern measurement standards. The Rasch model is the necessary and sufficient means to transform ordinal observations or counts into the required single attribute linear, interval and invariant measure. Items from questions are fitted to the Rasch model; in classical test theory the opposite occurs. Generic multiattribute preference measures fail because they are multiattribute and overlook the required Rasch measurement model. Claims for the EQ-Health and Wellbeing (EQ-HWB) preference instrument fail for exactly the same reason the EQ-5D-3L/5L multiattribute instruments fail; they overlook Rasch measurement standards for a single attribute. The results are preference scores capped at zero that were not developed to meet linear, interval and invariant properties with a true zero. Instead, the preference scores have arbitrary floors for the worst health state, including negative values for so-called states worse than death. As a result, the QALY is a failed construct to support modeled and other claims for cost-effectiveness. The only successor to the QALY would be disease specific, single attribute measures developed by application of Rasch modeling.

INTRODUCTION

In early 2020, to celebrate the 10th anniversary of *Innovations in Pharmacy*, the Formulary Evaluations section of the journal issued a call for papers [1]. The objective was to consider and/or propose modern scientific methods for determining the evidence base for the fair pricing and accessibility of pharmaceutical products and devices. A number of questions were raised to focus on submissions and commentaries. The key question was:

- 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?

If we are to match the standards of normal science for single, credible, evaluable and replicable claims in health technology assessment (HTA), then the first step must be to address the question of measurement. We must address the fact that *statistical analysis has dominated social sciences to the almost complete exclusion of the concept of measurement* [2]. Importantly, it is not the intention of modern or Rasch measurement to replace classical statistical techniques but to ensure that analysts base their application of these techniques on variables with unidimensional, linear, interval and invariant properties. This can only be achieved with Rasch measurement where items are selected to fit the requirements of the Rasch model; a quality control requirement that distinguishes acceptable from unacceptable instruments and their measure of therapy response. This ensures that we meet the Rasch standard, which applies to the physical as well as the social sciences, that meaningful measurement has to be based on the arithmetical properties of interval scales.

Unfortunately, in HTA this requirement for interval calibration is typically ignored in evaluating subjective responses to instruments, where the instrument developers have failed to consider the imperative of Rasch quality assessment and the importance of interval scales as a prerequisite for meaningful statistical analysis. Given this neglect, the purpose of this first of two commentaries on recent applications and claims for measurement in the HTA, is to, first, make clear that all HTA profiles and claims must meet Rasch standards as the unique basis for acceptance of rejection before any statistical applications for impact or response are made and second, to also make clear that the continuing endeavors to create generic preference scores to support quality adjusted life years (QALYs) with the EQ-Health and Wellness (EQ-HWB) instrument once again fail at the conceptual level to meet Rasch quality standards and there is no option but to reject them.

The second commentary focuses on cancer and the contrast between, first, instruments proposed by the European Organisation for Research and Treatment of Cancer (EORTC) in particular including the generic Cancer Quality of Life QLQ-C30 instrument, developed in the mid-1990s, the fifty of so supplemental cancer type specific instruments and the more recent preference scored QLU-C30 instrument based on items from the QLQ-C30 [3] [4], and second the Rasch modeled instruments from the Memorial Sloane Kettering Cancer Center, the BREAST-Q suite of instruments, the FACE-Q Aesthetics instruments and the BODY-Q obesity and weight loss instruments [5] [6] [7].

ATTRIBUTES AND LATENT CONSTRUCTS

In the physical sciences and, by extension Rasch modelling, the starting point is a credible construct, trait or entity that cannot be directly observed; hence the term latent [8]. In the physical sciences a classic example is temperature, in the social sciences such as health technology assessment, the obvious example is quality of life. These are latent constructs because they are not directly measurable. If we are to capture them then we need to focus on which intrinsic properties of these latent constructs, properties of interest to the observer, not the entity itself, can be measured; where the precursor of measurement is assessment and the creation of an instrument that meets Rasch standards.

Properties are defined by single attributes. There is no attempt, indeed it would defeat the objectives of measurement, to attempt to capture and measure a bundle of attributes at the same time. Each property of interest, attribute or trait should be assessed and measured separately; the broad concept of quality of life can be manifested as an attribute of interest in a number of contexts. The question is to assess whether it is possible to measure, following Rasch standards, the particular attribute of interest to the analyst. If the aspect of quality of life is the extent to which the needs of patients are met, then this has to be assessed through a unique instrument that captures needs systematically. The instrument's item responses are taken to be evidence for or an observable manifestation of the amount of the latent construct or trait the operationalizes this assessment.

Rasch measurement supports the identification of items to support instrument development where, in the case of patient reported outcomes, the assessment focuses on both the ability of the respondent and the difficulty of the item. The result, if Rasch standards are met, is an instrument for a single attribute, such as needs fulfillment, that is the basis for value claims and the evaluation of the extent to which needs are met by therapy interventions. The distinctive feature of Rasch measurement is that the model itself is independent of any data; the requirement of invariance that is a feature of all measurement.

Rasch modelling allows the amount of a trait to be mapped onto a line and its reliability to be assessed. Items selected for the instrument must a capture a different yet unique aspect of the trait to be evaluated. If this standard is met then we can claim that our measure is unidimensional. If different items assess different traits and different combinations assess different aspects than we conclude that the instrument is multidimensional. Rasch measurement is concerned that we can claim following and meeting Rasch standards that the instrument has transformed subjective responses to a measure of a single attribute with unidimensional, linear, interval and invariant properties.

THE NECESSITY OF MEASUREMENT

It is critical to understand that *the Rasch model represents the structure that responses from assessments should have before they can provide measurement and how they can be transformed to provide measurement.* Meeting the standards of the Rasch model is, as emphasized above, a necessary precursor to measurement as understood in the physical and more mature social sciences. All instrumentation to support patient reported outcomes must meet Rasch standards if we are to make a claim for progress in the social sciences in assessing disease impacts and value claims for therapy response. This point is made clear when we consider the distinction between classical test theory (including item response theory) and Rasch measurement. In the former case the observed data have primacy and modeling attempts to describe those data and explore the application of variable choice to capture the dependent variable; the basis for exploration of models to capture multiattribute health status preference scales. Rasch is completely different; the Rasch model is paramount. Rasch requires the data to fit the model through the size and structure of residuals. It is only then that we can consider the application of classical statistical techniques.

The importance of Rasch can be illustrated by what is referred to as the pragmatic or ex post facto assessment of any instrument. There is a substantial literature on the application of Rasch standards for an existing instrument to evaluate the extent to which a claim can be made that it is approximately consistent with the required criteria. This is only a stopgap solution because data collection, the selection of items, must begin with the application of Rasch standards. Rasch must guide data assembly through item selection; there must be a theoretical framework to guide data assembly. Attempting to salvage a justification that an existing instrument meets Rasch standards means that we put to one side any theoretical justification for why the data were assembled and manipulated in the first place. The measurement of a valid construct must be the first goal; to operationalize a latent construct or trait to assess, as a credible hypothesis, that there is a meaningful manifestation of the construct of interest. This is not restricted to needs fulfillment, but could be latent constructs such as sleep, fatigue, pain, mobility, depression or cognition. Exploration of the credibility and manifestation of attributes to capture aspects of these is not accomplished by one-line items asking for a single question response; items must be selected as a sample of all possible items, with items ordered by their intrinsic difficulty of being responded to; this applies to both dichotomous and polytomous items where each item threshold has its own item difficulty estimate.

Contrasting classical modeling with Rasch standards, points to the inevitable failure in the former in their elusive hunt for the will o'the wisp of an impossible multiattribute algorithm that will produce a true zero. Whether the techniques employ time trade off (TTO) or discrete choice experimentation (DCE), the failure to consider Rasch standards and the fit of items to a unidimensional interval model, means either undershooting with a floor greater than zero in valuation (death) or negative scores (states worse than death). The approximation to a bounded preference scale is entirely co-incidental as there is no evidence that the approved algorithm has the ability to meet unidimensional standards with linear, interval and invariant properties. The algorithm only creates an ordinal scale.

To underline the importance of Rasch measurement we can consider programs that attempt to identify patient centered core-impact sets (PC-CIS), where impacts, not to be confused with outcomes, include any reported effect or ramification from a disease or treatment; described as capturing the patient voice without restriction grouped into disease related impacts, treatment related impacts, impacts related to financial considerations and the impact on the family.[9] [10] [11]. Presumably, at some time the quantitative assessment of impact claims has to be assessed; the transformation from qualitative observations from patient groups, a consensus claim, must meet Rasch ,measurement standards. An impact should only be considered as credible if it can be shown to be assessed in terms consistent with Rasch measurement.

The commitment has to be to interval measurement of the patient voice; if not, the exercise has possible sociological or political interest, but no meaning in Rasch measurement terms. We could never assess the extent to which unfulfilled impact requirements are being met for the target patient population if there is no basis for the discovery of new, yet provisional, facts. There is no alternative to Rasch. Attempting to measure impact by off-the-shelf instruments that fail Rasch standards in unacceptable; a requirement that should be met, which it is not, in the FDA Program for Clinical Outcome Assessment or the core outcome sets proposed by the International Consortium for Health Outcomes Assessment (ICHOM). More concerning, given the effort that has gone into the creation of core outcome sets far a range of disease states over the past 30 years is the absence of consideration of fundamental measurement; the application of Rasch quality control criteria as a filter for accepting outcome instruments to be included in core outcome sets; PC-CIS is not alone although it takes a broader perspective in defining core outcomes. The COMET Handbook, produced by the Institute of Public Health, University of Liverpool, which is focused on the selection of appropriate outcome measures for clinical trials makes no mention of fundamental measurement and the importance of Rasch measurement; there is no indication that Rasch is a quality control in the COMET database of published and ongoing studies in core outcomes [12] [13].

Although speculating, PC-CIS could make a substantive contribution to the literature on core outcome sets by advocating the role of Rasch measurement. Indeed, it is not just the application of a Rasch filter (and encouragement of Rasch instruments) but that every PC-CIS impact claim (and this applies across the board to core outcome sets) should be accompanied by a protocol indicating whether or not there is a Rasch standard instrument designed to capture that impact or a proposal for instrument development. Measurement is not downstream from impacts claims. The only impact claims that are relevant are those that can be assessed in Rasch terms with the required measurement properties.

Given the paucity of patient centric instruments that have been developed to meet Rasch standards and the limitations on proposing instruments that have been pragmatically evaluated against Rasch criteria, this is a significant task. Presumably, a PC-CIS will be specific, whether assessed by stakeholder interview or literature search, to a representative sample of a target patient population. Given the number of disease states and the likelihood that the PC-CIS may be specific to sub-groups within that population raises the stakes even higher; African-Americans and Asian-Americans may have an entirely different PC-CIS from others across disease states.

Although there is a widespread application in education and, more recently, psychology attempts to commit to Rasch in the evaluation of patient reported outcomes is conspicuously absent in health technology assessment. Certainly, there are many attempts reported to applying pragmatic Rasch assessment for existing instruments. After all, the requirement for a matrix of patient responses to assess ability and item difficulty are easily created and can be assessed against Rasch requirements. But there is no concerted effort to argue, as practice guidelines, for the application of Rasch to the creation of measurements for the manifestation of credible latent construct attributes. Indeed, among leaders in what has been described as the HTA meme, there is no mention of Rasch in leading textbooks, practice guidelines by groups such as the International Society for Pharmacoeconomics and Outcomes Research and degree programs in HTA; a failure, if that is the right term, that extends to basic concepts in measurement theory [14]. There is no practice guideline, for example, supported by ISPOR for Rasch measurement; indeed, the practice guidelines produced over the past 20 years overlook Rasch and the imperative of interval measurement completely. Instead, in ISPOR -HTA there is a continued endorsement of multiattribute ordinal preference scores and, with the latest CHEERS 2022 guidance, support for of assumption driven simulated modelled outcomes to create imaginary cost-effectiveness claims; the QALY is endorsed but with no mention of the constraints of fundamental measurement [15]. The commitment is to approximate information to support resource allocation and the allocation of resources; Rasch criteria play no role with continuing efforts to defend the QALY [16] [17].

It would be unreasonable to argue that this neglect is deliberate but the fact remains that the acceptance of HTA and reference case simulation models that ignore measurement standards by formulary committees and single health system gatekeepers overlook an essential input to decision making. There is a fundamental belief in the need for a multiattribute composite single measure to support blanket claims for cost-effectiveness and resource allocation; the creation of a universal metric to driven resource allocation in health care. This belief is misplaced; there is no Holy Grail of a universal metric; one that is constructed by multiattribute algorithms combining a selected set of health dimensions and response levels. These multiattribute health state descriptions, or at least a sample of them, are valued subjectively to produce an ordinal score applying Time Trade Off (TTO) or discrete choice modelling (DCE) techniques which produce positive and negative scores, with different algorithms. The failure of multiattribute generic instruments to create preference scores with the required ratio property, capped at unity with a true zero, is due entirely to the lack of attention to the imperatives of Rasch measurement. Unless the agreed objective is to achieve a ratio measure the result will only be an ordinal scale. Rasch transformation, for single attributes, is the only basis for measurement.

HEALTH RELATED QUALITY OF LIFE

For some 40 years the key focus in health technology assessment, the creation of assumption driven modeled simulations to produce non-evaluable claims for cost effectiveness, has been to support claims for pricing, product access and the allocation of resources within health care systems. Once the unique contribution of Rasch modeling is seen as the necessary and sufficient standards for transforming ordinal subjective responses to single attribute measures is accepted, the concept of HRQoL as a measure, ceases to have any meaning. Belief in the ability of multiattribute algorithms to produce single attribute ratio measures for combining time spent in a disease state with a community preference score for that disease state is mathematically impossible

This does not mean that a credible commitment to quality of life defined in terms of a composite clinical entity cannot be subject to Rasch assessment. The task would have to be defined in terms of the manifestation and assessment of specific attributes considered relevant to evaluating the impact of therapy options. This would be defined as a profile of single attributes as manifestations of the latent construct and reported on separately; in each case the single attribute would have to meet Rasch measurement standards. This may form, with interval or ratio measures, the basis for a composite measure, although if the focus is on therapy response there would be more information gleaned from the separate measures rather than a composite measure.

Claims for a composite HRQoL QALY are misplaced; the QALY fails to meet any standard for fundamental measurement. Importantly, the failure of a QALY with an ordinal preference score leads to the questionable relevance of simulated modelled claims for long term cost-effectiveness to support formulary decisions [18]. If these simulated models are the primary justification for the acceptance of the QALY in HTA then the justification is misplaced. As these assumption driven simulations typically fail or a designed to produce empirically evaluable claims, it seems a pointless exercise to justify an ongoing commitment to the multiattribute EQ-HWB QALY as a needed input to such models.

If we are to transform subjective patient reported responses, whether in dichotomous or polytomous form, to measures that will support value claims for response to therapy then the necessary and sufficient condition is to apply the Rasch measurement model. This provides a unique, recognized and well established (over the past 70 years) framework for transforming subjective ordinal observations or counts to a single attribute, unidimensional, linear, interval and invariant measure; an approximate measure that can be transformed to a bounded interval or approximate ratio scale to meet the standards of fundamental measurement.

Although put to one side by mainstream HTA, there is a manifestation of a holistic quality of life latent construct: needs fulfillment. With its genesis in the Nottingham Health Profile developed in the late 1990s, there has been a commitment since the early 1990s to develop disease specific instruments applying Rasch modeling standards [19]. The needs fulfillment hypothesis is quite straightforward: the value of individual lives is dependent on the extent to which their human needs are fulfilled; value is low when few needs are met [20] [21] [22] . This does not mean that health related quality of life (HRQoL) measures of functions and symptoms are put aside. Given that the major influences on need fulfillment are the presence of disease and its treatment then the presence of disease and its treatment must play a key role; but this does not mean focusing only on purely clinical parameters. There are additional factors that need to be brought into play: social support, financial considerations, aids and support from others, education and the ability of the patient to respond, and other non-clinical influences. Rasch analysis is ideally suited to creating disease or target patient (including caregivers) instruments to create the required unidimensional, linear, interval and invariant measures of need fulfillment. Based on extensive qualitative interviews, the interviewer can probe the impact of the disease and the extent to which it adversely effects need fulfillment. Statements are identified concerning needs fulfillment and after testing for face and content validity a final item set can be developed to assess reliability and validity. Rasch assessment is applied and the needs fulfillment instrument constructed to meet the Rasch measurement standard. To date, some 30 instruments have been constructed and applied globally in clinical trials; these are available on line with the Galen measures database [23] . There is also the option of transforming the interval score to an approximate bounded ratio preference scale (fixed range 0 – 1) to support disease specific quality of life claims, including disease specific estimates of quality adjusted life years (I-QALYs) [24]. As a single attribute QALY, this provides a possibly key aspect of a profile of Rasch-standard attributes to assess therapy response.

THE EQ-HEALTH AND WELLBEING (EQ-HWB) INSTRUMENT

As both a complement and successor to the existing multiattribute generic preference instruments, the EQ-HWB developed over the past 6 years has now reached the stage of being valued to support QALY and cost-effectiveness claims. Initially proposed and since modified as a ‘bolt-on’ solution to criticisms that instruments such as the EQ-5D-3L/5L were too restricted in their coverage of relevant health dimensions and response levels, the result is a 25-item instrument (with a 9-item short form EQ-HWB-S) that focuses on how a respondent’s life had been over the last 7 days. The instrument is in two parts (for the long 25 item version) with 5 items for difficulty (e.g., ability to see, hear and mobility) and 20 items for frequency (e.g., I had problems with my sleep). Each item response defined by a five level Likert scale to capture increasing difficulty or frequency. The claim is that the items selected cover the themes identified from a literature review of qualitative evidence on how health and healthcare, social care, and caring roles impact on health and wellbeing. The selection involved face validation and psychometric testing with views obtained from consultations and stakeholders to create a multiattribute instrument for both 25 and 9 item versions.

The development of the EQ-HWB is the complete opposite of the Rasch quality standards for instrument development. Items were selected by agreement among respondents; they were not selected to fit the Rasch measurement model. Item selection was driven by the preferences of consultees with at least one item for each subdomain There appeared to be no commitment or conception of the Rasch requirement for a measure that focused on a credible single attribute with unidimensional, interval, linear and invariant properties. This means that the EQ-HWB fails to meet standards for fundamental measurement and the possibility of the application, with multiple response items, of either the Rasch Rating Scale Model or the Partial Credit Rasch Model to create an instrument with the required properties. While the authors claim that the EQ-HWB is a complement and an improvement over the EQ-5D-3L/5L in the coverage of additional domains such as energy and cognition, together with domains for social relationships and control, as well as separating out anxiety and depression, it suffers from the same weaknesses that bedevil the earlier instruments; the false belief that domains can be bundled together to create a composite fundamental measure. Just as the EQ-5D-3L/5L instruments failed to meet Rasch standards, the EQ-HWB repeats this mistake. Algorithms to support valuation are the result of fitting both TTO and DCE in the case of the HWB to items and health states; there is no concept of fitting items to the Rasch model to ensure a linear, interval and invariant measure. Even this would be unacceptable as the EQ-HWB is multiattribute.

It is of interest to note a recent criticism and exchange over whether or not the EQ-HWB represents an improvement over previous multiattribute generic instruments. In this exchange two criticisms were raised: (i) there was insufficient patient and public engagement in the development of the EQ-HWB and (ii) the lack of clarity in terms of what the EQ-HWB measures [25]. In the former case, the critique pointed to the lack of representativeness of the patient pool with interactive sessions characterized as reactive in item selection rather than a basis in direct elicitation from patients; this limitation was recognized in the response with the rather weak defense that the items, given possible response burden were considered representation of functional symptom domains [26]. In the latter case the focus was on poor content validity, a failure to identify and define the individual domains and sub-domains. The authors' state: …. there is no attempt to ensure that the symptoms and function items are exhaustive in terms of health nor do they appear to associate symptoms or functions with wellbeing. In support of the EQ-HWB the response was that: *Our instrument does not aim to capture only wellbeing but aims to capture both subjective feelings and more objective functioning because these are domains of life that (most) patients and service users consider to be important and meaningful to their lives.*

Unfortunately, this exchange misses the essential point: the EQ-HWB fails the standards for fundamental measurement. Rasch modeling focuses on the properties of entities not the entity itself. These properties, variously referred to a constructs, attributes or traits are assessed indirectly through their manifestation as a set of observations. questions or statements. These responses are qualitative, with the order of these responses the first step to measurement. The key step is the scoring of these responses through the ordered assignment of integers and, through the application of Rasch measurement, the mapping of these responses onto a line with linear and interval properties. This points to the importance of measuring a single trait, construct or attribute; unidimensionality requires all questionnaire items to assess a single or common trait.

The EQ-HWB, designed as a multiattribute scale, fails to meet Rasch standards. Certainly, it might prove possible to apply a pragmatic Rasch assessment to assess the extent to which the EQ-HWB is consistent with Rasch requirements, but this misses the point. The focus of Rasch is on transforming, as a single trait, subjective responses to interval scores. It is the only approach; but it rests on a credible entity and the ability to define the attributes of interest for separate assessment and measurement. With patient centric responses Rasch measurement is founded on a sample of items drawn from in depth patient interviews to assess needs. This engagement with patients involving items that are selected to accommodate patient ability and item difficulty has to be the starting point. This sets the basis for a probabilistic framework for response to therapy: the likelihood of a successful response is a function, in dichotomous terns, of the difference between item difficulty and patient ability. This has been recognized for some 30 years in the development of Rasch needs fulfillment disease specific instruments; and for 70 years in the general application of the Rasch model.

Certainly, if we consider needs fulfillment as a manifestation of quality of life, then we can and have developed measures for disease states and target patient population to assess needs and the extent to which those needs are met and the impact of therapy [27] [28] [29]. The focus, if we are to meet fundamental measurement standards, must be on the assessed manifestation of a latent construct; to measure the selected property of that construct Consider, as an example the latent construct 'sleep'. In the EQ-HWB this is assessed by a single item "I had problems with my sleep" with responses, for the past 7 days, in the range none of the time to most or all of the time. It is not clear how a response to this question is to be interpreted. Trying to capture in one item the contribution of problems (undefined) to health and wellbeing across disease states seems an inadequate attempt to capture sleep experience. In common with other items there are numerous PRO instruments that have reported on aspects of not only sleep, but fatigue, memory, anxiety, depression, anxiety and pain; all of which are represented in the EQ-HWB by one or two items. In the case of sleep, assessing the quality of sleep, the extent to which sleep is non-restorative and its association with fatigue are well established. Illustrative of a pragmatic Rasch assessment of the Greek version of the Pittsburgh Sleep Quality Index (PSQI) pointed to the need to modify the instrument to meet Rasch standards [30]. If sleep is considered a critical manifestation of quality of life or health and wellbeing then it should be evaluated as a separate manifestation or as a latent construct in its own right, not summarized in terms of one item which might be better phrased to emphasize the quality of sleep experience from the patient's perspective. At least, the end product would be, as demonstrated for the revised PSQI, a measure that meets Rasch measurement standards.

If the focus for patient reported outcomes is considered critical in evaluating the impact of therapy interventions and value claims for new therapies, then it seems doubtful that a single item for each of a range of health-related domains is adequate, particularly when patients are not involved in item selection. This points, of course, to the fundamental weakness of multiattribute instruments; however hard we try to bundle and value health states for the chimera of a single metric the resulting instrument scores fail measurement standards. We might apply time trade off (TTO) or discrete choice models (DCE) to create preferences, but the result will be, inevitably, a measure that fails the standards of fundamental measurement for an interval or ratio scale. There is no recognition of the need, not to fit a model to the data, but to fit the data, the items of a questionnaire, to the required Rasch calibrated measurement model. If there is no conception of the need for a single attribute ratio scale, then the exercise is fated from the start. Certainly, we can consider credible manifestations of a latent construct we might label health and wellbeing, but this has to be evaluated and data items selected to model credible single attributes, based on patient engagement, to produce a profile of specific measures.

The common feature of multiattribute instruments is the impossibility of anchoring the algorithmic preference or utility scores in the range from zero to unity. While the scores are defined in terms of decrements from ‘perfect health and wellbeing’ or a cap of unity, the scores inevitably fail to conform to a true bounded ratio scale from unity to zero. The result that is that valuation of health states can exceed unity (perfect health and wellbeing) and with the more adverse health states yielding negative scores. which are labelled ‘states worse than death’. A recent pilot study of the short form EQ-HWB is no exception [31]. A comparison of TTO and DCE attempts to fit alternative valuation models to the raw data health state descriptions produced utility scores for the worst health state for the 9-item short form EQ-HWB a range of utilities for the TTO model from -0.335 to 0.990 and for the DCE -0.335 to 1.021. A further hybrid rescaled assessment applying a Tobit model to improve model fit to the data produced a similar range of utility scores ranging from -0.368 to 0.0996 and -0.384 to 0.997 respectively. This is far from a ratio scale with a true zero. There is, as far as can be ascertained, no effort to specify the Rasch measurement standards required for this successor to the QALY and ensure that the EQ-HWB short form was capable of meeting these requirements to support single attribute claims. Instead, there are continuing reassessments to produce a better fit or fudge of the valuation to come as close as possible to the ‘ideal’. Even so, in the absence of Rasch requirements we have no basis on which to assume that the preference scores have the required measurement properties; the fact that the EQ-HWB is multiattribute ensures that this is the case. All we are left with for the EQ-HWB is a subjective ordinal score; not interval measurement.

CONCLUSIONS

The continued obsession in health technology assessment to create a broadly based multiattribute instrument to capture functions and symptoms is at an analytical dead end; the EQ-HWB is the latest casualty. It is not just the selection of items which ignores inputs from patients but the attempt to create an algorithm which transforms subjective responses to interval and ratio measures for multiple attributes. Given the focus of Rasch measurement on the creation of single attribute, unidimensional, linear, interval and invariant measure, a measurement model in place for some 70 years, the focus in HTA represents a commitment to instrumentation that fails both the standards of normal science and fundamental measurement. Just as the notion of a util was rejected in the mid-19th century, so the notion of a QALY metric must be abandoned, at long last, now in the early part of the 21st century. The only framework for transforming subjective responses to a single attribute, unidimensional, linear, interval and invariant measure is through Rasch modeling. Attempting to fit a model to subjective data, such as preferences for health states defined in multiattribute terms will fail. Certainly, an algorithm can be applied to generate a score, but that score will have only ordinal properties; it will not be linear, interval and invariant.

Developing the EQ-HWB falls into the same trap as the development of the EQ-5D-3L and EQ-5D-5L: the assumption that the standards of fundamental measurement can be put to one side in favor of a multiattribute , health state preference scale that is, presumably, designed to have ratio properties; bounded by unity and zero. Attempting to apply a model to the data, in the case of generic multiattribute instruments, not only ensures the absence of a true zero but also a measure that fails to demonstrate linear, interval and invariant properties. Indeed, reviewing the literature, the assumption appears to be that the scale has linear and interval properties, but no proof is provided. The Rasch model provides a basis for ensuring those properties are present while, at the same time, eschewing community preferences for health states in favor of the interaction between patient ability and item difficulty.

It might appear odd after some 70 years of Rasch modelling and the widespread adoption of Rasch measurement to capture subjective responses, supported by a range of readily available software packages and internet tutorials, that HTA persists in inventing and reinventing ordinal health status scales that are deemed to have ersatz ratio properties. The answer is clear cut: with the focus on assumption driven, modeled simulations to support non-evaluable claims for cost-effectiveness, the commitment to approximate information, the importance of employing a single multiattribute metric to create QALYs as inputs to the model is paramount. The embrace of these models to support gatekeeper threshold cost-per-QALY rules, makes the commitment to impossible multiattribute QALYS difficult to overcome.

In answer, therefore, to the question raised in the 2020 call for papers: the only successor to the QALY is another QALY construct that makes the same mistakes. A commitment to creating preference scores that may claim, as with the EQ-HWB, to capture a greater range of symptoms and function but which achieves nothing in terms of the standards of fundamental measurement. The QALYs are still impossible mathematical constructs and the preference algorithms continue to produce ordinal scores. The true successor, recognized for some 40 years, is to focus on single attributes, defined in terms of target patient populations in disease states, applying Rasch models to create measures that match the standards of the physical sciences.

REFERENCES

[1] Langley P, McKenna S. Post QALY: Scientific Approaches to Real World Formulary Decision Making. *InovPharm*. 2020;11:No 2 https://pubs.lib.umn.edu/index.php/innovations/article/view/3375

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

[3] Aaronson N, Ahmedzai S, Bergman et al. The European Organisation Research and Treatment of Cancer (QLQ-C30): a quality of life instrument for use in international clinical trials in oncology. *J Natl Cancer Inst*. 19893;85(5):365-76

[4] King M, Costa D, Aaronson et al. QLU-C10D: A health state classification system for a multiattribute utility measure based on the EORTC QLQ-C30. *Qual Life Res*. 2016;25(3):625-36

[5] Pusic A, Klassen A, Scott A et al. Development of a new Patient-Reported outcome measure for breast surgery: The BREAST-Q. *J Plastic Recon Surg*. 2009;124(2):345-53

[6] Klassen A, Cano S, Schwitzer J et al. FACE-Q scales for health-related quality of life, early life impact, satisfaction with outcomes, and decision to have treatment: development and validation. *Plast Reconstr Surg*. 2015;135(2):375-386

[7] Klassen A, Cano S, Alderman A et al. The BODY-Q: A patient-reported outcome instrument for weight loss and body contouring treatments. *Plast Reconstr Glob Open*. 2016;4(4)

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

[9] Perfetto E, Oehrlein E, Love T et al. Patient-Centered Core Impact Sets: What they are and why we need them. *The Patient – Patient Centered Outcomes Research*. 2022;15:619-27

[10] Langley P. Evidentiary Standards for Patient-Centered Core Impact (PC-CIS) Value Claims. *InovPharm*. 2022;13(3):No. 15 https://pubs.lib.umn.edu/index.php/innovations/article/view/5016/3241

[11] Perfetto E, Schrandt S, Escontrias et al. Patient-Centered Core Impact Sets (PC-C!S): What they are and what they are not. *InovPharm*. 2023; 14(1):No. 3 https://pubs.lib.umn.edu/index.php/innovations/article/view/5264/3348

[12] Williamson P, Altman, D, Bagley H et al. The COMET Handbook: version 1.0. *Trials*. 2017;18 (Supp 3): 280 file:///C:/Users/Paul/Downloads/s13063-017-1978-4.pdf

[13] Institute of Public Health, University of Liverpool, COMET Database. https://www.liverpool.ac.uk/population-health/research/groups/comet-initiative/?

[14] 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

[15] 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)

[16] Neumann PJ, Willke R, Garrison LP. A Health Economics Approach to US Value Assessment Frameworks – Introduction: An ISPOR Special Task Force Report. *Value Health*. 2018; 21:119-123 https://www.valueinhealthjournal.com/article/S1098-3015(17)33891-3/pdf

[17] Rand L, Raymakers A, Rome B. Congress' Misguided Plan to Ban QALYs. *JAMA June 2023* https://jamanetwork.com/journals/jama/fullarticle/2806096?guestAccessKey=237f9f5e-4ba1-46e1-82ea-bf0df490d966&utm_source=silverchair&utm_medium=email&utm_campaign=article_alert-jama&utm_content=olf&utm_term=060823

[18] 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)

[19] Hunt S, McKenna S, McEwen J et al. The Nottingham Health Profile; Subjective health status and medical consultations., *Soc Sci Med*;15(3):221-9

[20] McKenna S, Wilburn J. Patient value: Its nature, measurement and role in real world evidence studies. J Med Econ. 2018;21(5): 474-80 https://www.tandfonline.com/doi/epdf/10.1080/13696998.2018.1450260?needAccess=true&role=button

[21] 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 https://www.tandfonline.com/doi/epdf/10.1080/13696998.2018.1560303?needAccess=true&role=button

[22] 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 https://www.tandfonline.com/doi/epdf/10.1080/13696998.2018.1560304?needAccess=true&role=button

[23] Measures Database. Galen Research Ltd, Manchester UK https://www.galen-research.com/measures-database/

[24] Langley P. The Great I-QALY Disaster. *InovPharm*. 2020; 11(3): No 7 https://pubs.lib.umn.edu/index.php/innovations/article/view/3359/2517

[25] Perfetto E, Burke L, Love R et al. Measuring Health and Well-Being: We need to get it right for patients, with patients. *ValueHealth*. 2023;26(3):435-37

[26] Brazier J, Peasgood T, Mukuria C et al. Author Reply. *Value Health*. 2023; 26(3):435-37

[27] McKenna S, Wilburn J. Patient value: Its nature, measurement and role in real world evidence studies. J Med Econ. 2018;21(5): 474-80 https://www.tandfonline.com/doi/epdf/10.1080/13696998.2018.1450260?needAccess=true&role=button

[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 https://www.tandfonline.com/doi/epdf/10.1080/13696998.2018.1560303?needAccess=true&role=button

[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 https://www.tandfonline.com/doi/epdf/10.1080/13696998.2018.1560304?needAccess=true&role=button

[30] Panayides P, Gavrielides M, Galatopoulos C et al. Using Rasch measurement to create a quality pf sleep scale for a non-clinical sample based on the Pittsburgh Sleep Quality Index (PSQI). *Eur J Psych*. 2013;9(1):113-135

[31] Mukuria C, Peasgood T, McDool E et al. Valuing the EQ Health and Wellbeing Short Using Time Trade-Off and a Discrete Choice Experiment: A Feasibility Study. *ValueHealth*. 2023; in press https://www.valueinhealthjournal.com/action/showPdf?pii=S1098-3015%2823%2900057-8