MAIMON WORKING PAPER No. 9 JUNE 2023

AFTER THE QALY: MEASUREMENT AND THE ROAD NOT TAKEN (PART II: THE QLU-C10D INSTRUMENT)

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

ABSTRACT

Value claims for pharmaceutical products and devices must meet recognized standards to include both the standards for credibility, empirical evaluation, and replication that characterize the belief in progress and supports the commitment to normal science as well as those for fundamental measurement. In measurement terms, the objective is measure for value claims that is for a single, attribute, with unidimensional, linear, interval and invariant properties. This must apply to all value claims that based on subjective responses from patients and caregivers. Value claims that are not demonstrated to meet these Rasch measurement requirements must be rejected. The implications of this insistence or imperative for these standards have profound implications for health technology assessment. For the past 30 years, HTA has been locked into a meme that denies these standards. Instead, there is a commitment to developing assumptions-driven modeled simulations to create approximate information and non-evaluable claims for cost-effectiveness. This methodology is applied across the board, supported by the mathematically impossible quality-adjusted life year (QALY), where value claims in oncology and other disease areas are essentially a waste of time. The purpose of this commentary is to make the case that we need a new start, not just in oncology, to establish a meaningful framework for evaluating therapy options, setting the stage for the evolution of objective knowledge through lifetime disease area and therapeutic class reviews, supported by effective real-world outcomes-based contracting. Oncology is a significant starting point because of the extent to which the failure to meet these standards is institutionalized with groups such as the European Society for Medical Oncology (EORTC) and the required adoption of patient-centric disease-specific oncology measures that meet the standards for fundamental measurement has been accepted. There is, however, one outstanding exception: the development over the past 15 years of Rasch standard single attribute claims for a profile of patient centric claims in breast cancer (BREAST-Q), face aesthetics (FACE-Q) and obesity and weight loss (BODY-Q). This represents the way forward in oncology.

Keywords: EORTC, QLQ-C30, QLU-C10D, BREAST-Q, FACE-Q, BODY-Q, Rasch measurement

INTRODUCTION

Value claims for products and therapy interventions in medicine and health technology assessment (HTA) are only acceptable if they meet the standards of normal science and those of and fundamental or Rasch measurement. The critical departure point is that all value claims must meet measurement standards for unidimensionality with linear, interval and invariance properties. The unique role of Rasch or modern measurement theory in transforming quantitative observations to meaningful measurement was detailed in the Part I paper: Rasch measurement provides the necessary and sufficient means for this transformation. Applying these standards to the ongoing pursuit of generic single metric multiattribute preferences to support QALYs and simulated modeled claims makes clear that the effort is redundant; the effort put into the development of the EQ-Health and Wellbeing instrument over the past 5 or more years is essentially wasted effort. A multiattribute instrument fails Rasch requirements for single attribute measures; attempting through either time trade off (TTO) modeling of discrete choice modelling (DCM) is singularly inappropriate because in attempting to fit a model to subjective responses results only in ordinal scales. The Rasch model is unique in requiring data items of statements to fit the Rasch measurement model. The standards applied to select and fit items to the model ensures all requirement measurement standards are met.

Accepting these standards means that HTA conforms to accepted standards in the physical sciences and the more mature social sciences such as economics and education [1]. The hallmark of the standards in normal science is that all value claims must be credible, empirically evaluable and replicable; the commitment to set the stage for hypotheses testing, the discovery of provisional new facts, and ongoing disease area and therapeutic class reviews. Meeting the standards for fundamental measurement is also critical. Whether claims are for clinical endpoints, patient-reported outcomes, drug utilization or other resource utilization, all must meet standards as single attribute linear, interval or ratio measure with invariance properties. This is the essence of Rasch measurement and is a standard that must apply to any patient reported outcome in oncology or other disease intervention.

The purpose of this commentary is, first, to point to the failure to meet Rasch standards in the long-standing commitment by the European Organisation for Research and Treatment of Cancer (EORTC), notably the current endorsement of the development of the multiattribute QLU-C10D instrument and, second, to demonstrate the contribution of 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 and their application of Rasch scoring protocols [2] [3] [4].

THE CHALLENGE FOR ONCOLOGY

Judged by the standards of normal science and modern measurement or Rasch Measurement Theory (RMT), the last 30 years have witnessed a profound systemic failure to create instruments to evaluate therapy response and value claims in oncology. The failure is deep-seated, in particular with the role of the EORTC in its support for the generic multiattribute EORTC Core Quality of Life Questionnaire (QLQ-C30) together with the 50 bolt-on disease specific oncology modules by EORTC have any role in supplementary claims for therapy response and, most recently, the generic preference QLC-C10D instrument and the valuation of the instrument with general population normative data for 13 European countries, the US and Canada [5] [6] [7].

The challenge for EORTC and those utilizing these various instruments is that none of these were designed to meet the standards of fundamental measurement. Whether we are concerned with measurement in the physical or social sciences the requirement is for a single attribute, unidimensional scale with linear, interval and invariant properties.

The question of measurement in the social sciences, where subjective responses are the standard, presents a problem that Rasch measurement solved in the 1950s [8]. Observations in any science are not measurement unless they relate to a previously constructed and maintained calibrated unidimensional, linear and interval scale which is invariant to situations in which it is applied. This is, as emphasized in Part 1, is a prerequisite to the application of classical statistical analysis; otherwise, at best, the scale is ordinal and will only support non-parametric statistics. The trap that the EORTC instruments fall into; is that they cannot capture and support any claims for therapy response or status in oncology because they lack Rasch measurement standards.

The challenge that Rasch resolved was to transform observations into measurement; responses to the items selected for an instrument to a measure that met the required standards. Rasch addressed two issues: first, a measure must be invariant; it must maintain its quantitative status irrespective of the application context while, second, there is an unavoidable interaction between the instrument and the respondent which cannot be fully predetermined, but must involve a probabilistic component. In other words, the more difficult a questionnaire item, the lower the probability of a respondent with average ability, responding to it positively.

The ultimate challenge for anyone developing instruments to assess respondent status and response to therapy is to demonstrate that the various instruments meet Rasch standards; that there is a valid justification for the claim that instrument meets Rasch standards. This is not the question of *ex post facto*, pragmatic assessment of Rasch standards that, by happenstance, the particular instrument in the items selected (including discarding items) meets Rasch standards but in demonstrating, by an audit trail, that the instrument was developed following the application of Rasch standards for item selection for increasing difficulty, and the distribution of a random sample of respondent abilities to realize the particular item.

RMT does not support the creation of composite instruments. These are disallowed because of the need to ensure dimensionality and dimensional homogeneity [9]. Proposing composite measures such as the QLQ-C30 and its reduced item offshoot the QLU-C10D are illusory; chasing measures that are nothing more than a will o'the wisp. Ensuring the unidimensionality in an instrument that is designed to manifest a latent trait is mandatory; we have to operationalize the latent construct applying Rasch analysis. All the items in an instrument must support a single construct. Attempting to add together different latent constructs, such as bundled health state descriptions, will deny unidimensionality. Of course, once your composite score has been developed, there is the appeal of attempting to claim unidimensionality, or factors with unidimensional attributes. This is disallowed by the fact that in ignoring Rasch rules, you have only an ordinal score which supports only non-parametric statistics; *factor analysis is faulted by mistaking ordinally labeled stochastic observations for linear measures and failing to construct linear measures* [10]. If we are to judge the merits of a measured manifestation of a latent construct then, as with Rasch measurement, we require a coherent construct theory that orders observations and a specification equation. This allows scores to be predicted on a linear interval scale from responses to items. Add to these requirements the role of dimensional homogeneity: we can compare variables only if they have the same dimension and can be converted to each other (e.g., centigrade and fahrenheit). If there are different dimensions, as there are by definition in composite health related quality of life

(HRQoL) bundles (e.g., EQ-5D-5L symptom dimensions) then they all break the rules for dimensional homogeneity and hence construct validity.

THE EORTC QLQ-C30 INSTRUMENT (V3.0)

The QLQ-C30 is a 30-item multiattribute quality of life questionnaire designed to capture physical, psychological and social functions. It comprises booth multi-item scales and single item measures. There are five functional scales, three symptom scales a global health status or quality of life scale and six single items. No item appears in more than one scale. The scoring of the QLQ-C30 is by integer responses to a 4-level (3 threshold) Likert format (Not at all = 1 to Very much = 4); the Global Health/Quality of Life items are scored for 7 levels (6 thresholds). The scales, with item numbers in parentheses, are: (i) Global Health Status (2); (ii) physical functioning (5); Role Functioning (2); Emotional Functioning (4); Cognitive Functioning (2); Social Functioning (2) and (iii) Symptoms: Fatigue (3); Nausea and Vomiting (2); Pain (2); Dyspnoea (1); Insomnia (1); Appetite Loss (1); Constipation (1); Diarrhoea (1); Financial Difficulties (1). Scoring is by summation of the Likert integer value responses and standardizing to a score. This is accomplished by estimating the average of items that make up the scales and using a linear transformation to standardize the raw score to 0 -100. It is possible to represent scores for each functional.and symptom scale, as well as an overall scale.

If we accept the relevance of a latent construct that we can label quality of life, then from the perspective of fundamental measurement the QLQ-C30 falls short as a multiattribute instrument; a more useful framework would be to consider the symptoms/functions as attributes to be manifested as a profile of separate instruments that meet Rasch standards. As it stands, that is impossible with the QLQ-C30 as the number of items representing each attribute are too few to support a Rasch assessment of item difficulty and respondent ability to even get to first base for a Rasch evaluation; a one item response is hardly a meaningful basis for assessing the impact of cancer therapies. The insistence on a composite measure is simply a step too far.

The QLQ-C30 scale does not meet the standards for fundamental measurement; it fails to manifest a single attribute from a credible latent construct with unidimensional, linear, interval and invariant properties. If, as is commonly the case, a Likert response format that is used in the QLQ-C30 is captured by a summation and standardization of integer values then two assumptions are being made: first, that all items in the instrument are of equal difficulty and, second, that the thresholds between steps for response options are of equal distance. In other words, the interval nature of the data is assumed. This is not the case for the overall symptom and functional scales; summation is not possible to claim as a measure of response to therapy; claims for an overall scale also fail. We cannot fall back on the assumption that over certain ranges counts and measures are likely to be highly correlated; this has to be verified before any raw scores are subject to statistical analysis.

The presumption by the developers of the QLQ-C30 that these assumptions hold is surprising given that in the 30 years prior, Rasch modelling had provided solutions to the transformation of integer or Likert responses to fundamental measurement with, initially, the Rasch Rating Scale Model that assumed a common threshold structure for responses and then the Rasch Partial Credit Model where each item was assumed to have its own threshold structure.

It's instructive to consider that at the time the EORTC QLQ-C30 was first developed in the 1980s there were ample red flag warnings that pointed to the need to focus on fundamental measurement: manifesting single attribute linear interval measure of a latent trait or construct [11]. Yet, the focus on the development of a generic instrument that bundled together disparate HRQoL dimensions deemed relevant across the board for cancer patients, took priority. In terms of Rasch measurement, the effort failed at this first hurdle. This decision put to one side a more thoughtful question: if Rasch measurement disallows composite multidimension (or multiattribute) ordinal scores why not focus on single attributes that are common across cancer states: A coherent manifestation of the patient voice in therapy interventions as the patient (or caregiver) is the ultimate judge and possible beneficiary of therapy interventions.

The starting place must be a latent construct that is considered credible for patients and caregivers in disease states, and which can be manifested by the application of Rasch rules to capture and measure specific credible attributes. Since the early 1990s a proposed manifested construct for quality of life is needs-fulfillment (detailed in Part I); the framework for a construct or trait that the quality of life of a patient (or caregiver) is determined by the extent to which the needs that are identified from extensive subject interviews are met. In oncology and other chronic disease states, while health interventions may be expected to be the principal factors that impact needs, the needs of patients may not correlate with HRQoL clinical parameter considerations. Whether a therapy facilitates needs being fulfilled, the judgement belongs to the patient [6 7] .

Judged by the standard for fundamental measurement, therefore, the QLQ-C30 is a failure. All that has been produced is a summation of raw scores from the averaging of Likert items. There appears to be no concept of the need for a single attribute linear interval scale to capture measurement for specific cancer disease states. At best, we have a collection of functional items and symptoms which even by the standards for aggregating Likert responses fail because if we want tovaluate Likert scales we require a demonstrated *a priori* assumption that all items are of equal difficulty and that the thresholds between the steps are of equal distance. The Rasch model for polytomous responses makes no such assumptions. Indeed, we know, by application of IRT analysis that the QLQ-C30 items vary in terms of their difficulty [12] ; the problem of thresholds for different items has also been addressed in terms of thresholds for clinical symptoms [13]. While the QLQ-C30 has been categorized as an instrument to capture quality of life, it is better seen as including a one-item question asking about quality of life as simple integer values (c.f., Likert pain scales) with symptoms and functional status tagged on. As it stands, we have a composite or multiattribute instrument that lacks dimensionality and dimensional homogeneity and which fails the standards for simple aggregation of integer values to generate a raw score (where standardization is also disallowed). While we might categorize these raw scores as ordinal, they are no better than the multiattribute preference scores or utilities generated by the EQ-5D-3L/5L instruments. In terms of the standards for fundamental measure, the QLQ-C30 is not a generic quality-of-life instrument with acceptable measurement properties. A focus on specific items raises the question of why bother when there are many disease-specific instruments that could report, possibly more comprehensively, on the symptoms and functions identified in the QLQ-C30.

It must always be remembered that focusing on symptoms and functions may suggest prospective instrument items but ones that may be of little interest to patients as elements in their quality of life; we might infer that there is an impact but we need a direct measure. This is where the needs of patients and caregivers become pivotal [14] [15] [16]. If quality of life takes its cue from the ability of patients to meet their needs, where chronic disease may have a major role, then we have to identify those needs; not a collection of Likert scores for categories of symptoms and functional status ordinal levels. That is, by items selected from interviews that are intended to assess an underlying trait or latent construct. If the latent construct is needs fulfillment, then we have a firm basis for developing cancer disease-specific instruments aiming for an accurate repose for perceived direct patent benefit. This is an imperative: a measure with over 60 years of experience in its application.

The case against the QLQ-C30 applies equally to the 50 EORTC proposed cancer specific type modules, which are viewed as supplemental to the QLQ-C30 with the additional data supporting more intensive assessments of cancer experience. Once again, total scores are simply integer aggregations, transformed to a standard scale. None of the resulting scales meet the required fundamental measurement standards and are no guide to therapy responses. While none of the modules was developed in respect of Rasch standards it is worth noting that few of these instruments have been pragmatically assessed against Rasch standards; typically, in circumstances where there is a focus on particular functions and where selected items are captured from complementary instruments [17] .

A recent study of breast cancer in Saudi Arabia is indicative of the misapplication of both integer scores from the QLQ-C30 and the QLQ-BR-23, in reporting on values for each functional and symptom scale, and overall. The mean scores for the QLQ-C30 functional scales range from 79.6 for social functioning to only to 63.6 physical functioning and for symptoms 42.73 for insomnia to 15.2 for diarrhea. As ordinal scales, the QLQ-C30 and QLQ-BR-23 cannot support basic arithmetical operations; at best they support non-parametric statistics which means attempts to apply standard techniques to support mean value and dispersion claims are disallowed. This oversight is all too common for multiattribute instruments where preference or scores are, judged by the imperatives of Rasch measurement, only ordinal while reported in literally thousands of HTA publications either as preferences or, by extension, QALYs.

THE EORTC QLQ-C10D INSTRUMENT

Over the past 10 years considerable effort has gone into creating a multiattribute health state classification system utility instrument from the QLQ-C30 items, the EORTC QLU-C10D. Supported by the Multi-attribute Utility in Cancer (MAUCa) Consortium the QLU-C10D is intended to support the application of composite scores in cost utility analysis with country specific value sets for the QLQ-C10D, where the QLQ-C30 is utilized in clinical trials or observational studies. The QLu-C10D is based on the established domain structure of the QLQ-C30 to select a subset of items and dimensions. Twelve items were selected representing 10 dimensions: physical functioning (mobility), role functioning , social functioning, emotional functioning, pain, fatigue, sleep, appetite, nausea and bowel problems. Two problems should be noted before going any further: first, as a composite multidimensional or multiattribute instrument the QLU-C10D fails to meet the standards for fundamental measurement where the starting point is the manifestation of a single attribute to apply Rasch analysis to create a linear, interval and invariant measure to capture the underlying latent trait; and, second, while the item selection process involved Rasch criteria to assess the item most representative of those items in multi-item dimensions, followed by a range of psychometric assessments including confirmatory factor analysis, it was not recognized than in order to apply classical statistical techniques you must ensure that you are basing the analysis on linear interval measures. The Rasch measurement standard is the prerequisite to statistical analysis; original observations are not a calibrated measurement system, a score measuring a theoretical trait that has construct validity with all objects ordered on the measurement dimension. Rasch standards must be the quality control to determine which data sets are acceptable and which should be rejected.

Having avoided the scientific measurement imperatives of Rasch modelling, the next step in the creation of the composite QLU-C10D is to value bundles of health dimensions. If we consider the Australian discrete choice valuation of the QLU-C10D as the exemplar there are, for the QLU-C10D generic instrument, 4 utility weights attached to each of the 10 dimensions, ranging from 0 = not at all to 4 = very much. All are qualitative responses; no measurement properties other than order. Given the number of possible bundled health state valuations by a representative sample of the Australian population (4^{10} = 1,048,576) the decision was to apply discrete choice was reduced to 1920 health states in 960 choice sets. Discrete choice does not create measurement requirements that meet Rasch standards.

At no stage in this process of determining subjective utility weights was any consideration given to Rasch standards. This would, of course, have been irrelevant as the commitment was to a composite multiattribute health status preference or utility score, rather than a single attribute linear, interval measure and invariant measure. The main considerations were to manipulate the data, fitting a model to the data, to constrain the aggregate of health state valuation responses to be capped at zero with death, presumably the worst health state, at zero. While it is possible to model a cap of unity (perfect QLC-C10D) there is no perception of the need to create a bounded metric which yields interval, linear and invariant properties. There is no evidence to suggest that discrete choice modelling can support such claims; we are no further ahead that the community scores proposed to support the EQ-5D-3L/5L instruments.

One feature of the QLU-C10D that is worth emphasizing is the lack of consideration given to invariance which is a requirement of measurement in the physical sciences. To achieve invariance, it must be demonstrated that the measure of any variable by any measurement system should be independent of any particular measurement instrument that is appropriate to the task. Invariance is directly related to unidimensionality; to achieve invariance we measure one construct at a time to achieve a linear, interval measure. In PRO measures, invariance means that all person and item parameters must be the same irrespective of the context in which the instrument is applied. The QLU-C10D fails to meet this requirement. Rather than one valuation set that has application globally, the QLU-C10D utility algorithm is country-specific. The result is not unexpected, applied as decrements in utility score from the 'no problem' health state of unity the values for the worst health state for nine countries range from -0.159 (Netherlands) to 0.15 (Canada) and with six countries reporting negative scores both for the worst health state and for health states approach [18] [19] [20] [21] [22] [23]. If invariance as a recognized objective had been achieved the worst health state would be zero in all instances, with no need for separate country valuations. Extending the hypothetical argument, would mean that the application of a global or common set of utility decrements would allow cancer QALY claims to be compared between countries and not country specific with no basis for comparison. Cost-utility and modelled simulated cost-effectiveness claims for the same product will, although imaginary, vary between countries.

The failure to achieve invariance, or at least its equivalence for a multiattribute scale which guarantees its absence along with the absence of linear and interval properties, repeats the experience with the EQ-5D-3L/5L and other earlier generic instruments with their health states worse than death for the various country-specific valuations. The latest valuation of the EQ-5D-5L in the US, for example, yields values ranging from -0.573 to 1, and with 20% of heath states worse than death [24]. Employing DCE instead of TTO, which was employed in the earlier generic instruments, makes no difference to the value range and states worse than death. with preliminary findings reported for the short-form EQ Health and Wellbeing (EQ-HWB) instrument (as detailed in Part I) with TTO values ranging from -0.335 to 0.990 and the DCE values ranging from -0.335 to 1.021 [25].

THE BREAST-Q PACKAGE

The BREAST-Q is not a single instrument; it is a package of scales to evaluate patient reported outcomes among women with different types of breast surgery [26]. The conceptual framework focuses on breast conserving therapy, mastectomy and reconstruction. The common construct for all interventions is quality of life and satisfaction, with reconstruction adding two further dimensions of expectations and breast sensation. The first version of BREAST-Q was published in 2009 and version 2.0 in 2017.All modules were developed applying Rasch measurement to create single attribute, unidimensional, linear, interval and invariant measures to ensure high content validity and accurate tracking of clinical change [27]. Each scale is transformed to score out of 100. The scales can be used to support research but also to inform clinical care at the individual level. A recent review of BREAST-Q content validity to ensure continuing relevance and need for new scales concluded that while BREAST-Q remained comprehensive new scales for upper extremity lymphedema, breast sensation, fatigue, cancer worry and work impact were developed [28].

SCORING THE BREAST-Q SCALES

Each of the BREAST-Q scales (together with checklists and stand alone items) is scored independently for both dichotomous and polytomous responses scored as YES/NO. There are five scale groupings, each matched to the French/Kincaid reading level:

- BREAST-Q: Breast Cancer Core Scales (5 scales)
- BREAST-Q: Mastectomy Scales (11 scales including core scales)
- BREAST-Q: Breast Conserving Therapy Scales (13 scales including core scales)
- BREAST-Q: Reconstruction Scales (22 scales including core scales)
- BREAST-Q: Reconstruction Expectations Scales (6 scales)

Quality of life is defined by eight scales, including the core scales. These are:

- Adverse effects of radiation (6-items)
- Animation deformity (12-items)
- Cancer worry (10-tems core scale)
- Fatigue (10 items core scale)
- Impact on work (8-items core scale)

- Physical well-being

What is often overlooked is the false implication that the Rasch model converts or transforms ordinal scale data to an interval scale; a claim that overlooks the difference between observations and measurement [29]. The fit of the data or items to the Rasch model is all that is required; the item category scores only have to be ordered. The fit of data to the Rasch model assures us that we have successfully measured a quantitative variable as manifestation of a credible latent construct. In the case of an instrument with dichotomous responses the raw score is the number of items that are answered correctly; but a count of item responses is not measurement. In Rasch terms measurement is defined as the discovery of the structure of quantity in the data; not the assignment of numbers to objects. To achieve measurement the raw score is the input to the analysis. As a count, the raw score has a point of origin for all items being incorrect, capped at the point where all items are correct. Having established that a quantitative latent variable can be inferred from the data by meeting Rasch fit criteria, the final step is to transform these non-linear raw data to a linear interval scaled measure of the latent variable. This is achieved by a linear transformation of logit values for each raw score. The Transformation is {linear value = (logit value*slope) + intercept}. This yields the required Rasch single attribute, unidimensional, linear, interval and invariant scale capturing the manifested latent construct.

The BREAST-Q scales follow this required linearizing process, where scale anchor point is zero successfully completed responses (in a dichotomous scale) or thresholds passed (polytomous scale). The anchor point is arbitrary and is not a true zero that would be the key characteristic of a ratio scale. Care has to be taken in understanding the transformation from the raw score to a linear scale. The term 'raw score' should not be confused with the term ordinal scale and the presumption that the Rasch model automatically creates an interval scale. The primary role of Rasch modelling is to estimate the likelihood that a respondent will successfully answer or respond to an item; items that are ordered in terms of the interaction between respondent ability and item difficulty. As expressed in terms of a Wright map, the iterative maximum likelihood procedure required of the Rasch model maps the distribution of abilities against the distribution of items in terms of logits. While the required Rasch model is linear and interval , the logits mapped to the Wright model do not have interval properties. We might claim a linear conversion to percentages and then to logits, but the iteration does not produce an interval; it merely fits selected items measured in logits to the required Rasch continuum or interval scale. It is also worth noting that a key role potentially played by the Wright map is to question the number of items with the same logit value and the importance of gaps in the logit sequence. However, if we are seeking a 'raw score equivalent' linear scale, then we cannot assume that the raw scores of items are on an interval scale; The Rasch model assumes that the relationship between a latent trait and the observed response is logistic. This means that the logit values do not have equal intervals along a measurement continuum. To achieve this requires the application of a linear scaling re-calibration is required to transform to a linear scale with equal intervals; each logit value is transformed to obtain a corresponding linear scale value.

MISSING VALUES IN RASCH MODELLING

Once a instrument or scale in the case of the BREAST-Q has been developed, presumably with the imposed benefits of a complete person ability and item difficulty matrix, further application of the instrument in a target patient or respondent group may result in missing values. The pertinent question is the point at which the instrument or the respondent population has to be discarded if missing cell responses are considered overwhelming. Fortunately, Rasch analysis can be quite robust given missing data, as Rasch iterative calculations do not require a completely ordered matrix. All that is required is a sufficient density of data. It is nor proposed in Rasch analysis that missing cell values can be accommodated by interpolation of an inauthentic datum.

The question of missing data is built into the Rasch model and the application of iterative maximum likelihood to establish required item and ability logits. It is not a question, as with the BREAST-Q transformations of arbitrarily trying to plug in missing values, but to ask whether or not the analysis is affected by missing data or whether, for practical purposes, we can ignore the missing data elements. A recent simulation to assess the 'affect' of missing data considered three situations: MCAR (missing data completely at random); MAR (missing at random) where an incorrect response to an item led to the next item being missed and MNAR (missing not at random) where 'missingness' is related to the values of the variable that would have been reported but are not [30] . The simulation found that that when responses were missing completely at random or missing at random item parameters for the Rasch model were unbiased. Against this, where responses were missing but not at random, all item parameters were severely biased, particularly when the number of missing responses was high. The implication is, given problematic assessment of prior MNAR, the focus should be on ensuring maximum competed responses. The impact of missing values is not resolved by the BREAST-Q solutions; instead, should be a red flag for revising the scale or facilitating a complete response.

ABANDONING RASCH: THE PROPOSED BREAST-Q UTILITY MODULE

After a long-term, and continuing, commitment, to Rasch measurement to support a range of single attribute invariant unidimensional scales to capture outcomes and symptoms in breast cancer interventions, it is a surprise to see a recent commitment to put Rasch measurement standards aside in a proposal to develop a preference-based measures, the BREAST-Q Utility module [31] . The principal reason put forward for this endeavor was the failure of existing generic PROs to capture the unique concerns of women with breast cancer. The authors maintained that the development of this multiattribute instrument followed established procedures for instrument development, involving both extensive subjective or qualitative phases, supported by a quantitative assessment for the to five health related quality of life concerns (HRQoL) and item ranking for these to be included in the instrument. The final version of the instrument captured the top five HRQoL concerns across all stages of breast cancer: appearance of the breast, fatigue, cancer worry, impact on usual activities and anxiety.

The first draft of the BREAST-Q Utility (Version 1.0) module captured nine unique items or dimensions of patient experience with 4 – 5 response options. This was subsequently amended (Version 2) to 12 dimensions with 14 items, with 4 response levels each, with further amendments following patient and expert opinion input, to 10 unique dimensions with 21 items, captured by four or five response levels. The next stages in instrument development are to examine the pattern of responses and psychometric properties of the module followed by a valuation survey to elicit utility weights for each dimension included in the module. Apparently, a primary objective of the final utility preferences is to support cost-effectiveness claims, presumably with assumption driven modelled simulations to create imaginary outputs. This will require, rather than generic QALYs as the key input, breast cancer QALYs created from the application of the BREAST-Q utility scores.

The problem, as detailed in the first paper which assessed the EQ-Health and Wellbeing (EQ-HWB)O instrument in the first of these two commentaries and in the first part o the present paper with the assessment of the EORTC QLU-C10D instrument, is the lack of credibility for the entire exercise. If the objective is to create a multiattribute preference utility scale with ratio properties, then the fact that the authors aim for a multiattribute scale means this is impossible; the only basis for creating a preference score is Rasch measurement theory. This, of course, refers only to a ratio measure for a single invariant measure. As noted previously, in both the present and preceding commentary, ratio measures for composite subjective variables are mathematically impossible. The first step must be to start from developing., as a manifestation of a credible latent construct, a single attribute, unidimensional, linear, interval and invariant scale. This can then be transformed to a bounded interval or approximate ratio scale to create preferences. These can be applied to create the equivalent of QALYS, but for single attributes. This process could be easily applied to any of the existing BREAST-Q scales to create a preference and utility score for that scale.

Instead, the proposed BREAST-Q utility score will, if capped at unity with utility decrements for each of the items or dimensions of the scale, fail to meet the ratio scale requirement of a true zero. Applied to different breast cancer populations there is the likelihood of either a floor vale > 0 or a floor vale < 1 (a state worse than the most adverse outcomes across the various dimensions). The resulting QALY equivalent score would have no basis for cost-effectiveness imaginary modelled claims; a result that is all too apparent with the EQ-HWB and the QLU-C30D. At the same time, if the principal driver is to support cost-effectiveness claims, then reference case models developed by NICE in the UK and ICER in the US have no merit whatsoever and they fail both the standards of normal science with credible, evaluable and replicable claims and those for fundamental measurement.

CONCLUSIONS

Measurement in oncology, as defined by the widely used EORTC modules and the BREAST-Q scales, represents two competing paradigms: the traditional True Score Theory (TST) or Classical Test Theory (CST) and the application of Rasch Measurement Theory. THE EORTC approach is essentially descriptive and exploratory where the models are developed to fit the data; Rasch is confirmatory and predictive where, in probabilistic terms, the data are designed or selected to fit the model. It is only with Rasch that we have a clear cut objective for measurement: all claims for therapy interventions and responses must be in terms of single attribute, unidimensional, linear, interval and invariant measures. If we are concerned with progress and the discovery of new yet provisional facts in oncology response and capture the patient voice, then we must start with measurement where the Rasch model is not an option, but the sufficient and necessary imperative for transforming observations to measurement. If we forget this imperative, oncology is poorly served. In these terms the difference between the commitment to Rasch, single attribute measures in BREAST-Q standards in marked contrast to the long-discredited attempts to capture the patient voice with TST/CST.

A fundamental error is to purse the Holy Grail of a single metric to capture the quality of life of patents. The BREAST-Q offers a number of scales which it labels under the quality of life umbrella; wisely the authors made no attempt to aggregate these to a single metric. Certainly, this results on a possible portfolio of scales which can be reported on separately; but this is a strength and not a weakness. Unfortunately, judged by the available evidence, those supporting the BREAST-Q seem attracted by the will o'the wisp of a single preference metric for quality of life. This is best a path not taken as the resulting composite measure, if the historical application of TST/CST is a guide, is that the end product is nothing more than an ordinal score. This is evidenced not only by the latest empirical results for the EQ-HWB but the valuation attempts for country specific versions of the QLU-C10D.

REFERENCES

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

[2] 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

[3] 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

[4] 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)

[5] Kaasa S, Bjordal K, Aaronson N et al. The EORTC core quality of life questionnaire (QLQ-C30): validity and reliability when analysed with patients treated with palliative radiotherapy. *Eur J Cancer*. 1995;31A(13-14):2260-63

[6] Wheelwright S, Bjordal K, Bottomley A et al. EORTC Quality of Life Group guidelines for developing questionnaire modules. 5th. EORTC, 2021

[7] Nolte S, Liegl G, Petersen M et al. General population normative data for the EORTC QLQ-C30 health quality of life questionnaire based on 15,386 persons across 13 countries, Canada and the United States. *Eur J Cancer.* 2019;107:153-63

[8] Rasch G. Probabilistic Models for some Intelligence and Attainment tests. Copenhagen: Danish institute for Educational Research, 1960

[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] Wright B. Comparing Rasch measurement and factor analysis, *Structural Equation Modeling: A Multidisciplinary Journal*, 1996; 3:1, 3-24

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

[12] Shih C-L, Chen C-H, Sheu C-F et al. Validating and improving the reliability of the EORTC QLQ-C30 using a multidimensional Rasch model. *ValueHealth*. 2013;16:848-54

[13] Giesinger J, Kuijpers W, Young T et al. Thresholds for clinical importance for four key domains for the EORTC QLQ-C30: physical functioning, emotional functioning, fatigue and pain. *Health Qual Life Outcomes*. 2016; 14:87

[14] 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

[15] 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

[16] 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

[17] Regnault AQ, Pompilus F, Ciesluk A et al. Measuring patient-reported physical functioning and fatigue in myelodysplastic syndromes using a modular approach based on EORTC QLQ-C30. *J Patient Rep Outcomes*. 2021;5(1): 60

[18] King M, Viney R, Pickard A et al. Australian utility weights for the EORTC QLU-C10D, a multiattribute utility instrument derived from the cancer-specific quality of life questionnaire, EORTC QLQ-C30. *PharmacoEconomics.* 2018;36:225-38

[19] Kemmler G, Gamper E, Nerich V, Norman R, Viney R, Holzner B, King M. German value sets for the EORTC QLU-C10D, a cancer-specific utility instrument based on the EORTC QLQ-C30. *Qual Life Res.* 2019; 28(12):3197-3211.

[20] Nerich V, Gamper EM, Norman R et al French Value-Set of the QLU-C10D, a cancer-specific utility measure derived from the QLQ-C30. *Appl Health Econ Health Policy*. 2020. doi: 10.1007/s40258-020-00598-1. PMID: 32537694

[21] Gamper E, King M, Norman R et al. EORTC QLU-C10D value sets for Austria Italy and Poland. *Qual Life Res*. 2020;29: 2485-95

[22] Finch AP et al. Estimation of an EORTC QLU-C10D value set for Spain using a discrete choice experiment. *PharmacoEconomics*. 2021; 39(9): 1085-109

[23] Jansen F, Verdonck-de Leeuw I, Gamper E et al. Dutch utility weights for the EORTC cancer-specific utility instrument: The Dutch EORTC QLU-C10D. *Qual Life Res*. 2021;30:2009-19

[24] Pickard A, Law E, Jiang R et al. United States Valuation of EQ-5D-5L Health States Using an International Protocol. *ValueHealth*. 2019;22(8): 931-41

[25] 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

[26] BREAST-Q Version 2.0. Users Guide January 2023 https://qportfolio.org/wp-content/uploads/2023/01/BREAST-Q-BREAST-CANCER-USER-GUIDE.pdf

[27] Pusic A, Klassen A, Scott A et al. Development of a new Patient Reported Outcome Measure for Breast Surgery: The BREAST-Q. *Plast Reconstr Surg*. 2009;124(2):345-43

[28] Kaur M, Chan S, Bordeleau L et al. Re-examining content validity of the BREAST-Q more than a decade later to determine relevance and comprehensiveness. *J Patient Rep Outcomes*. 2023;7(1):37

[29] Salzberger T. Does the Rasch Model convert an ordinal scale to an interval scale? *Rasch Measurement Trans. 2010*;24(2):1273-75 https://www.rasch.org/rmt/rmt242a.htm

[30] Waterbury, G. ,Missing data and the Rasch model: The effects of missing data mechanisms on item parameter estimation. *J App Measurement*. 2019;20(2):154-66

[31] Kaur M, Klassen A, Xie F et al. An international mixed methods study to develop a new preference-based measure for women with breast cancer: the BREAST-Q Utility module. *BMC Women's Health*. 2021;21:8