MAIMON WORKING PAPERS No. 9 FEBRUARY 2022

HEALTH STATES WORSE THAN DEATH: MODERN MEASUREMENT THEORY AND THE REJECTION OF ORDINAL PATIENT REPORTED OUTCOMES

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

Abstract

The fact that multiattribute preference score algorithms should produce negative values or states worse than death should come as no surprise. It reflects a long standing, but now resolved, debate over fitting a model to the data versus fitting the data to a model: classical test theory (CTT) and item response theory (IRT) versus modern measurement theory with Rasch measurement theory (RMT) to create patient reported outcomes (PROs). In the first instance the effort is directed towards maximizing the fit while the second selects data elements that are consistent with the model. The implications are profound: in the former case the resulting composite preference scores, defined by fitting the scoring algorithm to the data, cannot support anything but non-parametric statistics; they defy the standards of fundamental evidence and modern measurement theory. In the case of RMT the commitment is to fundamental evidence with the creation of interval and, if possible, ratio measures. Seen from this perspective those committed to PROs face a critical decision: do they accept modern measurement theory and the standards of normal science or do they persist with a demonstrably false measurement as part of their technology assessment meme. The most egregious example of rejecting normal science is with generic ordinal preference scores created by algorithms that ensure that there are negative values which invalidate, not only the score itself but applications in creating quality adjusted life years (QALYs). Other examples abound in disease specific PROs where the overwhelming majority create ordinal scores; ignoring the limitations imposed by the axioms of fundamental evidence. The implications are unfortunate: for over 30 years the focus in health technology assessment has been on creating generic and disease specific scores to capture response to therapy that are, from the perspective of modern measurement theory, false. Given they are in fact all ordinal scores, any claim to measure therapy response is misplaced. The purpose of this commentary is to demonstrate why modern measurement theory must be accepted to support PRO claims. For too long those developing PROs and attempting to map PROs have been following a measurement will o'the wisp belief system. It is time to abandon ordinal scores and reject 30 years of PRO misinformation.

INTRODUCTION

Health technology assessment occupies a unique position among the physical and more advanced social sciences: the belief that therapy decision can be based on invented evidence and non-evaluable claims for cost-effectiveness. This belief system can be traced back over 30 years with the decision by 'leaders' in the field to reject hypothesis testing in favor of the creation of approximate evidence;

rejecting science to embrace non-science or metaphysics and pseudoscience [1 2 3]. The manifest failings of this belief system, or meme, have been extensively documented to include the rejection of the standards for credible, evaluable and replicable value claims and, notably, the rejection or lack of awareness of the axioms of fundamental measurement and modern measurement theory [4 5 6].

The purpose of this commentary is to consider the implications of patient reported outcomes, both generic multiattribute scores and disease specific instruments in terms of modern measurement theory. Advocates of this belief system have assumed, or wish to believe, that multiattribute preference scores from instruments such as the EuroQoL EQ-5D-3L and EQ-5D-5L have (or must have) ratio measurement properties to support quality of life (QALY) based assumption driven imaginary simulations. The Institute for Clinical and Economic Review (ICER) in the US, for example, is adamant that health economists have the upmost confidence that preference scores have ratio properties and can thus support imaginary QALY modeling; which is their core business model [7]. In fact, the generic multiattribute preference scores, among other failings, are nothing more than composite ordinal scores which have no place in supporting value claims for therapy response. It is not just the generic preference scores that fail the criteria for fundamental measurement but the overwhelming majority of disease specific PRO value claims also fail these standards for response to therapy; they create only ordinal scores.

STANDARDS FOR VALUE CLAIMS

Previous commentaries have made the case for meaningful value claims and the processes that should be in place for formulary committees and other health system decision makers to accept those claims for empirical evaluation that meet modern measurement standards. There are two requirements:

- **All value claims must refer to single attributes that meet the demarcation standards for normal science: they must be credible, evaluable and replicable**
- **All value claims must be consistent with the limitations imposed by the axioms of fundamental measurement and must meet interval or ratio standards**

These requirements are unexceptional, at least if viewed from the perspective of modern measurement theory and the standards common in the physical sciences education and economics. In health technology assessment there appears to be a lack of awareness of the axioms of fundamental measurement, in particular as they apply to PROs. Briefly, an interval scale is one with is invariance of comparisons, the scale units are equal in an ordered relationship with an arbitrary zero; that is, the scale can support addition and subtraction. A ratio scale has a true zero or universal constant where no value can take negative scores. An interval or a ratio scale, which captures interval properties, are ideal and these measures are critical if any value claims involve latent attributes such as quality of life (QoL).

If a scale has only ordinal properties then values on the scale have an ordered relationship with each other but the distance between each is unknown. Importantly, as this is typically overlooked, an ordinal scale cannot support arithmetic operations; it can only support non-parametric statistics, medians and modes, with minimal application in value claims for response to therapy. Unfortunately, the overwhelming majority of PRO instruments, both generic and disease specific have only ordinal properties.

Once these standards are introduced we enter the realm of modern measurement theory for PROs and the application of Rasch Measurement Theory (RMT) with its critical place in supporting interval and, in special cases, ratio value claims. Surprisingly (or perhaps not) RMT is essentially absent in health technology assessment in the development of PROs to support value claims. While there are examples of analysts attempting to re-jig an established PRO to select items that meet Rasch measurement standards *ex post facto*; the authors of the various multiattribute generic health related quality of life (HRQoL) scales and those concerned with developing disease specific PRO scales appear unaware of RMT. If they had been they would have recognized the application of conjoint simultaneous measurement, introduced in the 1960s, to assess in probabilistic terms, expected response, to an instrument that accommodates respondent ability and item difficulty [8] [9]. As detailed by Rasch:

> ***….a person having a greater ability than another person should have the greater probability of solving any item of the type in question and similarly, one item being more difficult than another means that for any person the probability of solving the second item is the greater one [10] .***

Since the early 1990s RMT has set the standard for a measure of the latent construct need-fulfillment quality of life (QoL) which, in meeting the standards for interval measurement (and more recently ratio measurement) relegates PRO instruments claiming health related quality of life (HRQoL) to an historical measurement curiosity or relic. This is seen in the distinction between classical test theory (CTT) and item response theory (ITT) on the one hand and RMT on the other. The distinction is critical: CTT/IRT ignore fundamental measurement in focusing on the primacy of capturing all items in a data set where the results are descriptive of that data set (exploratory and descriptive) while RMT requires the data to fit the model (confirmatory and predictive). Fit is determined by the size and structure of residuals, to assess the extent to which the requirements of probabilistic conjoint measurement have been sufficiently realized to claim that the item scale has approximate interval properties.

The authors of the various multiattribute generic preference scales overlooked the role of RMT in developing an interval scale. This is only a first step, because while an interval scale can support claims for response to therapy, an interval scale cannot support multiplication and thus the mathematically impossible (or I-QALY) QALY. To achieve this we require a second step, a further transformation to a bounded ratio scale. This has been achieved but only in a restricted application in the case of need fulfillment QoL interval measures [11].

THE IRRELEVANCE OF EQ-5D SCORES

Despite the apparent popularity of the PRO EQ-5D instruments, the EQ-5D-3L and EQ-5D-5L, it is impossible to justify their application to support value claims for competing therapies either as part of randomized clinical trial (RCT) protocols (where they are invariably secondary and underpowered end points) or to support observational studies. Indicative of their popularity a PubMed count (24 February, 2022) of combined terms EQ-5D OR EQ-5D-3L OR EQ-5D-5L yielded 10,389 hits for publication since 1981. At the same time, given the role of multiattribute instruments such as the EQ-5D ordinal preference scores for creating the mathematically impossible QALY, a count for QALY OR QALYS since 1981 yielded 21,976 hits. These are incredible figures when it is considered that the family of EQ-5D scores yield only composite ordinal measures and, as a result, the QALY is an impossible mathematical construct. The unfortunate fact is that few if any of these authors seem aware of the limitations of fundamental measurement, modern measurement theory and the fact that the various multiattribute preference scores have only ordinal properties.

If we accept the argument that for latent value claims such as quality of life (QoL) in health technology assessment to be acceptable, the claim should refer to a single attribute with unidimensional properties, which are credible, evaluable and replicable, then the EQ-5D instruments fail these criteria. They are, by description, multiattribute which means they lack dimensional homogeneity and construct validity. They are also ordinal scales a characteristic they share with the overwhelming majority of disease specific PROs. This is an unremarkable observation for a simple reason: when the instruments were being developed no one thought about measurement properties; as a result the default state is always an ordinal score. Importantly, their respective algorithms, irrespective of the weights attached, can all produce negative scores. This means they fail to support a true zero which requires that under no circumstances should the instruments support negative scores or states worse than death. The EQ-5D-3L, for example, can produce scores in the range -0.58 to 1.0 with UK weights for the 243 possible health states while the application of the EQ-5D-5L in a recent US study found that 20% of health states had negative scores out of a total of 3,125 health states [12]. Irrespective of the other limitations in multiattribute generic preference scores, the fact that the various algorithms can produce negative scores is the deciding factor: the possibility of negative scores defining states worse than death invalidates and belief that the EQ-5D instruments can create interval, let alone ratio scales.

THE TUFTS UNIVERSITY ORDINAL REGISTRY

As evidence for the pervasive belief in PROs that have ratio properties, we can consider the Tufts University Medical Center's Cost-Effectiveness Analysis (CEA) established some 46 years ago in 1976 [13]. Since then the database has grown to include some 10,000 studies involving some variation of cost utility analysis and QALY belief. The common element has been the presence of preference or utility scores, typically from multiattribute instruments such as the EQ-5D-3L/5L with the database reporting

on these together with various preference ratios and health state utilities. In addition, there is the more recently developed disability adjusted life year (DALY) GH-CEA database which exhibits the same fatal flaws as the Tufts CEA database. The health state utilities from the Tufts CEA have been used widely, for a fee, with the database receiving commendations.

The Tufts webpage allows a review of the preference scores presented for 100 health states; although whether these are representation of the entire registry is uncertain. Putting to one side the possibility that the weights come from the various direct and indirect preference scales which render comparisons impossible (but may be overlooked by users). the distribution of scores is of some interest. Of the 100 entries accessible through the webpage, 47% of health states have negative values or 'states worse than death'. The range of negative health state scores is from -0.01 to -0.55; the range for positive scores is from zero to 0.93. As these are averages, presumably, of the distribution of individual responses they are technically false as only a ratio measure can support addition and division. At the same time, averaging incorrectly over individual responses implies that the underlying distribution of scores can encompass both all negative values or a combination of negative and positive values; demonstrating once again the absence of a true zero which would require only positive scores under all circumstances. With such a high prevalence of negative scores it is surprising that after 46 years no one in the Tufts Medical Center thought, apparently, about the implications of this; the insights negative scores give to the underlying measurement properties and the requirement for a bounded ratio scale to create QALYs. Median values for the various multiattribute distributions of scores are not reported, although appropriate for ordinal distributions. This is due, no doubt, to the failure of most authors reporting ordinal preference or utility scores to present only averages and standard deviations (both disallowed) and not the actual ranked score values to assess the distributional characteristics.

THE IMPOSSIBLE QALY (I-QALY)

Even if, by some sleight of hand or dint of the imagination we were to agree that preference scores had interval properties, including both negative and positive vales (the latter including zero), it would not suffice to create QALYs; there is no true zero. The possibility of the various preference algorithms to create negative values makes this quite clear. If, as noted in previous commentaries, the QALY is mathematically impossible then it overturns any belief in incremental cost-per-QALY claims for HRQoL and the overarching claim that a product is cost-effective. The CHEERS 22 guidance is then just an empty exercise, joining the leading text book in technology assessment as a primer for imaginary constructs that ignores modern measurement theory [41].

If a measure is to be accepted then, to create QALYs or their equivalent it must be a bounded ratio scale capped at unity and with a true zero. These last properties are critical; to meet the standards for modern measurement theory; there must be no circumstances under which in the case of preference scores, the algorithms generating those scores can include a health state with a negative score for all possible target populations. The possibility that a given target population yields only positive scores does not mean that the next application of the instrument will not yield a single negative score (Hume's problem of induction) [14].

The authors of the various multiattribute generic preference scales overlooked the role of RMT in developing an interval scale. This is only a first step, because while an interval scale can support claims for response to therapy, the interval scale cannot support multiplication and thus the mathematically impossible QALY (or I-QALY). To achieve this we require a second step, a further transformation to a bounded ratio scale. This has been achieved but only is a restricted application in the case of need fulfillment QoL interval measures where assumptions can be justified for a true zero [15].

MISAPPLICATION OF ORDINAL MULTIATTRIBUTE PATIENT REPORTED OUTCOMES

The term multiattribute can apply equally to generic preference instruments as well as to disease specific PROs. In the former case we are dealing with composite bundled symptoms (or attributes) that are taken to describe health states. The standard gamble (SG) and time trade off (TTO), described as direct preference scales, are defined by a bundled description which yields both ordinal and negative values while the indirect multiattribute such as the EQ-5D, HUI and SF-6D ask respondents to assess their level of distress (if any) with predetermined health dimensions. Combining these responses yields health states which are then valued by scoring algorithms. In either case the result is the same: an ordinal score which cannot capture response to therapy.

The most significant misapplication is to assume that both direct and indirect preference scores yield bounded ratio scales. This is a misinterpretation. As they can create negative scores either directly with the SG or TTO, or indirectly via system algorithms they must fail the ratio property requirement for a true zero. As detailed below further misapplications occur when the false assumption of a ratio property supports mapping or crosswalking both between generic scores or from disease specific PROs to generic scores.

Disease specific PROs are typically constructed from sub-domains each defined by a series of Likert scales. This falls at the first hurdle because Likert scales are ordinal. Respondents are asked to assign integer values either to a agree/disagree symmetric scale or, more usually in PROs a ranked range of integer values (typically 0[1] to 5 or 0[1] to 7) to items or statements attempting to gauge either level of agreement/disagreement or with a symptom level. While these may be a short cut to assessing collective response in the less sophisticated marketing exercises, they are unacceptable as a quantitative measure of response to therapy (a single attribute ranked Likert pain scale is qualitative). There are a number of issues which the application of Likert scales to developing PRO scores that are not resolved. These include the validity or relevance, including wording, of the item; the intrinsic difficulty of the item; the weighting of the item, the ability of the patient to understand and respond to the item and the absence of any sense as to the psychometric distance between the items. In respect of the distance the Likert integer responses could be equally well be represented by letters from the a alphabet (e.g., A > B > C etc.) or even emojis.

A representative example of a 'Likert score' representing disease specific quality of life is the Asthma Quality of Life Questionnaire (AQLQ). Consider how the AQLQ is assembled and the measurement implications of Likert scales [16]. This is a 32 item-questionnaire used to assess the physical, occupational, emotional and social qualities of adults 17 to 70 years exhibiting mild to moderate asthma. It is a multiattribute instrument with four domains: symptoms (12 items), activity limitation (6 generic and 5 patient-specific items), emotional function (5 items), and environmental stimuli (4 items). Each item response is on a 7 point Likert scale with responses ranging from 1 = maximal impairment to 7 = minimal impairment. The items are in the form of questions with each of the scale points anchored on a word or phrase and not just the extreme values; descriptors include "totally", "extremely", "very", "moderate", "some" "a little". As Wilson et al note, some of these scales may be confusing to respondents as they mix adjectives with other grammatical elements and that there is no published evidence that the anchor words and phrases can be consistently ordered independently of their numerical positioning on the response scale or that the relative positions of different phrases represent approximately equal psychometric intervals [17]. The fact that the AQLQ has shown strong classical measurement properties based on integer ratio assumptions, is irrelevant; this only occurs if you ignore modern measurement theory and assume the AQLQ has ratio properties for the Likert scores or even an emoji for each Likert space.

Nevertheless, the scoring of the AQLQ ignores the requirements of fundamental measurement and treats the scales as if they had ratio properties. This allows an average score to be created for each ordinal Likert scale with domain and aggregate scores created by merging the average Likert values for each item (mathematically disallowed). To describe the average AQLQ score as a 'score' is a misnomer; it is a value that results from illegitimate manipulations of Likert scales to produce a 'number' that is meaningless in response to therapy terms. Although seldom articulated, the issue of combining negative and positive scores is of interest. The possibility exists that the (disallowed) average of the preference score may be negative (leading to negative I-QALYs) or zero (an average death state) which makes no sense other than that on average they certainly would have no discernible, in this life at least, quality of life.

The AQLQ ordinal scale stands in contrast to the RMT based Asthma Life Impact Scale (ALIS) which meets the requirements for an interval score for need-fulfillment QoL to capture response to therapy [18]. The ALIS interval score can also be transformed to a bounded ratio scale for further response assessment. ALIS is not the only example, some 30 other RMT interval measures have been develop and applied mainly in clinical trials; including psoriasis [19], Crohn's disease [20] and migraine [21]. All can be transformed to bounded ratio scales to report response to therapy in needs fulfillment terms.

IMPOSSIBLE ORDINAL MAPPING

The axioms of fundamental measurement are quite clear: unless designed to have interval or ratio properties, PRO instruments are, by default, only capable of creating ordinal scores. This means that mapping between PROs that each has an ordinal score is mathematically impossible. Unfortunately, this limitation is not recognized by professional groups such as ISPOR and by its house journal *Value in*

Health which, in addition to publishing research practice guides to ordinal mapping has published a significant number of mapping studies. It might be noted that this impossible mapping was designated by ISPOR as one of the two top priorities for development of good practice research guidelines; in retrospect, an unfortunate choice.

The ISPOR recommendations for mapping methods are to estimate health state utilities to estimate health state utilities from non-preference based outcome measures [22]. These recommendations follow from an earlier good research practice guidelines for estimating health-state utilities for economic models in clinical studies [23] . In neither of these practice recommendations is there any consideration of modern measurement theory or the question of the impossibility, in particular, of either selecting instruments with ordinal scores to support imaginary modeled claims or the limitations imposed when the disease specific PRO has only ordinal properties; which is invariably the case. Given the limitations of multiattribute generic instruments in creating ordinal preference scores and the impossibility of the I-QALY it seems odd that there should be a focus on attempting to emulate this limitation; translating one ordinal disease specific instrument to a generic ordinal disease specific scale. Certainly it is important, as the earlier recommendations emphasize, to plan early in product development for the most appropriate instrument for quality of life value claims. This choice must, of course, be disease specific and informed by modern measurement theory to provide evaluable value claims for single attributes that reflect the patient voice..

Yet ISPOR perseveres with *Value in Health* continuing to publish ordinal mapping studies. The latest contribution mapping from the EQ-5D-3L to the EQ-5D-5L takes the position that as the latter is the multiattribute scale of the future than the existence of numerous EQ-5D-3L scores will require 'updating' [24]. Given the ordinal nature of the EQ-5D-3L this seems a patent waste of time; if after the proposed transformations you end up with just another ordinal scale then you are no further ahead given the requirements for single attribute unidimensional value claims with interval or ratio properties. The authors seem unaware of the ordinal nature of the EQ-5D-3L scale, the potential for negative scores and the lack of dimensional homogeneity and construct validity. The fact that the QALY is an impossible mathematical construct appears not to be on the radar. An understanding of modern measurement theory is entirely absent; if analysts see the EQ-5D-5L as the multiattribute of the future then they should recognize it as an analytical dead end; one that fails to meet health system decision requirements unless you believe in CHEERS 22. While one might appreciate what ISPOR sees as a pressing need to create preference scores from ordinal PROs to populate assumption driven imaginary cost-utility simulations, the entire mapping exercise is a waste of time and effort.

MISINTERPRETING VISUAL ANALOG SCALES

Visual analog scales in the case of the EQ-5D instruments are multiattribute ordinal scores despite responses being forced to an interval scale. Respondents are presented with a bundle of symptoms and response levels to be subjectively placed on the VAS hierarchy. This process fails at this stage as the bundle lacks dimensional homogeneity and construct validity; as the assignment to the VAS is by a community sample the link to patients is also lost. Typically the VAS is 'scored' on a forced tableau

with a range 0 to 100 (or unity). It is, in effect, a more crowded Likert scale. As a Likert scale it lacks, or was never designed, to have interval properties; the distance between the assigned marks on the scale is unknown. Although negative values are absent by construct it is not a ratio scale for no other reason that it lacks invariance of comparison. If analysts believe that a VAS score has ratio or interval properties then they are mistaken; ignoring questions of relative difference and the potential for clustering at the midpoint. To address the question of relative distance, the transformation of the VAS to the mathematically more useful log-odds scale, the odds or probability of success, an approximation to an interval scale would be required; a transformation that was made clear by Thurstone a century ago [25]. This transformation is never attempted with authors assuming incorrectly that it has not just interval but bounded ratio properties.

ALL MULTIATTRIBUTE PREFERENCE SCORES ARE REALLY THE SAME

It is not unusual in assembly of the assumptions to drive a modeled simulation to come up short on the search for preference scores from a selected generic instrument, whether the measure is direct, indirect or is the product of mapping. One solution, but time consuming and irrelevant, is to attempt a mapping from a disease specific PRO to the 'required' preference scale. A less demanding tactic is to assemble the required preference scores from a number of direct and indirect preference instruments, putting on one side the different health dimensions covered, the different response levels, and the basis on which the preference scoring is created. A recent example of this is the ICER model, an off-the shelf model from academic consultants, for trilaciclib and plinabulin to prevent chemotherapy induced neutropenia [26].This model presents EQ-5D scores (it is not clear if these are EQ-5D-3L or EQ-5D-5L) [27] [28] lumped together with utility scores from a variant of the standard gamble technique [29] and time trade off technique (TTO) [30]. The willingness to combine preference score from different generic multiattribute sources makes a mockery of the notion of an assumption driven simulation when the assumptions are 'violated' in this way. Of course, the criticism rings hollow when it is recognized that any number of models driven by any arbitrarily selected set of assumption can drive imaginary cost-effectiveness claims for the same product combination. Given this, the inclusion of preferences from different systems seems inconsequential as there is no basis for empirical evaluation of value claims in any event.

MODERN MEASUREMENT AND THE COSMIN DEBACLE

One of the more disheartening aspects of health technology assessment is the presence of checklists for the construction and acceptance of imaginary, non-evaluable value claims. This is apparent in the CHEERS 22 guidance for constructing imaginary assumption driven simulation models to meet acceptance standards with leading journals willing to publish such imaginary claims; opening the door to a multitude of competing model claims for disease areas and targeted therapy comparisons. In this tradition is the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) [31] [32] [33] . The apparent purpose of this checklist, which has been available or over 10 years, is to apply standards to assess the 'quality' of PRO instruments and avoid bias, including systematic reviews. A recent review and exchange with the COSMIN authors make the point that the COSMIN

checklist is not fit for purpose [34], noting that it is not clear what the COSMIN purpose actually is in its defense of CTT in PRO development and assessment, with studies declaring they have applied the COSMIN checklist in systematic reviews showing little understanding of CTT, let alone RMT. The defense by the COSMIN authors is unconvincing as is their lack of appreciation of the standards of fundamental evidence, notably the fact that all CTT based PRO scores are ordinal [35].

In these respects, it is worth noting that there is a failure by ISPOR in its series on recommending practice standards for health technology assessment to make reference to the need to meet the standards modern measurement theory. The issue of empirical evaluation as part of a process of modeled value claim validation is apparently only an option [36]. In particular, there is no attempt to introduce the ISPOR membership to the distinction between classical test theory (CTT) and item response theory (IRT), and RMT. Certainly, CTT can play a role in RMT instrument development, but the belief holds that CTT is all that is required to develop response measures, which is incorrect.

CONCLUSIONS

If we accept the standards of normal science, including those of fundamental measurement, then we are a long way from thwe application in PROs of modern measurement theory which meet the required standards for value claims: single attribute, unidimensional claims that are credible evaluable and replicable. If we also accept that latent constructs are most appropriately measured by application of RMT, then we are faced with, by default, ordinal claims for response to therapy rather that the required interval or ratio measures. The most singular criticism is that no one, in developing generic multiattribute PROs or those for target populations in disease areas, gave a thought to the need for a PRO instrument that was designed, for measuring latent constructs, to have interval or, more appropriately bounded ratio properties to capture, as accurately as possible, response to therapy.

Certainly, there has been a degree of belated recognition that Rasch measurement may support item selection from ordinal and multiattribute PRO instruments. Unfortunately, this is a false belief as it still ignores the question of fundamental measurement and the need, not to try and capture observations, but to fit data elements to a prior model for the latent construct of interest through capturing the needs of the target patient group. As the various generic PROs and disease specific PROs are typically multiattribute, lacking dimensional homogeneity and construct validity, not to mention a true zero in the obvious case of generic instruments, they are, as ordinal scales, incapable of measuring response to therapy. Attempting to map from either one multiattribute generic score to another or mapping from a disease specific PRO to a generic PRO is both mathematically impossible and a waste of time. Perhaps those endeavoring to transform one ordinal score to another should remember that *you can't make a silk purse out of a sow's ear* [37].

There will undoubtedly be pushback in an attempt to delay the abandoning of generic and disease specific PROs; possibly exemplified in the latest COSMIN defense of traditional PROs that an instrument should be selected if, given the COSMIN criteria, is the least worst for the construct and

population of interest [38]. Again, this fails completely to recognize the standards of modern measurement theory. Of course, if the analyst is satisfied with approximate multiattribute response claims, then they will be continually on the defensive with criticism that they are creating only ordinal claims that have no basis in science, only non-science or metaphysics and pseudoscience.

The fact remains: the overwhelming majority of PRO instruments produce only ordinal scores. They represent an analytical dead end from which no amount of attempted transformations to ersatz interval and the equally impossible ratio score will extricate them. We have experienced, as note in a letter to *Value in Health,* the ISPOR house journal, 30 (or more) wasted years in PRO development, failing to meet measurement standards that have been recognized for some 60 years [39]. It is time to accept modern measurement theory consistent with RMT standards. Will this happen? Or is the willingness to reject the standards of normal science and the axioms of fundamental evidence too well ingrained?

REFERENCES

[1] 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

[2] 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

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

[4] Langley P. Nothing to Cheer About: Endorsing Imaginary Economic Evaluations and Value Claims with CHEERS 22. *F1000Research. 2022*

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

[6] Langley P. Peter Rabbit is a Badger in Disguise: Deconstructing the Belief System of the Institute for Clinical and Economic Review. *InovPharm*. 2021; 12(2): No 22

[7] Langley P. Supping with the Devil: Belief and the Imaginary World of Multiple Myeloma Therapies Invented by the Institute for Clinical and Economic Review. *Inov Pharm*. 2021; 12(3): No. 6.

[8] Luce R, Tukey J. Simultaneous conjoint measurement. A new type of fundamental measurement. *J Math Psychol*. 1964; 1(1);1-27

[9] Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedagogiske Institut, 1960

[10] Bond T, Fox C. Applying the Rasch Model: Fundamental Measurement in the Human Sciences (3rd. Ed.). New York: Routledge, 2015

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

[12] 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

[13] Tufts University Medical Center. Cost-Effectiveness Analysis (CEA) Registry. https://cevr.tuftsmedicalcenter.org/databases/cea-registry

[14] Magee B. Popper. London: Fontana; 1974

[15]

[16]Juniper E, Guyatt G, Epstein R et al. Evaluation of impairment of health-related quality of life in asthma: development of a questionnaire for use in clinical trials. *Thorax*. 1992;47:76-83

[17] Wilson S, Rand C, Cabana M et al. Asthma Outcomes: Quality of Life. *J Allergy Clin Immunol*. 2012;129(3 0): S88-123

[18] Meads D, McKenna S, Doward L et al.Development and validation of the Asthma Life Impact Scale (ALIS). Respir Med. 2010;104(5):633-43

[19] McKenna S, Cook S, Whalley D et al. Development of the PSORIQoL, a psoriasis specific measure of quality of life designed for use in clinical practice and trials. *Br J Dermatol*. 2003;149(2):323-31

T

[20] Wilburn J, McKenna S, Twiss J et al. Assessing quality of life in Crohn's disease: development and validation of the Crohn's Life Impact Questionnaire (CLIQ). *Qual Life Res*. 2015;24(9):2279-88

[21] McKenna S, Doward L, Davey K. The development and psychometric properties of the MSQoL: A migraine – specific quality-of-life instrument. *Clin Drug Investig*. 1998;15(5):413-23

[22] Wolowacz S. New ISPOR Recommendations – Mapping Methods for Estimation of Health State Utility (Editorial). *ValueHealth*. 2017;20;28-29

[23] Wolowacz S, Briggs A, Belozeroff V et al. Estimating Health-State Utility for Economic Models in Clinical Studies: An ISPOR Good Research Practices Task Force Report. *Value in Health*. 2016;19:704-19

[24] Van Hout B, Shaw J. Mapping EQ-5D-3L to EQ-5D-3L. *ValueHealth*. 2021;24(9):1285-93

[25] Thurstone, L. Theory of attitude measurement. *Psych Rev*. 1929*, 36*(3), 222–241.

[26] Tice JA, Bloudek L, McKenna A, et al. Novel Agents to Prevent Chemotherapy-Induced Neutropenia and Other Myelosuppressive Effects; Draft Evidence Report. Institute for Clinical and Economic Review, January 25, 2022

[27] Kuehne N, Hueniken K, Xu M et al. Longitudinal assessment of health utility scores, symptoms and toxicities in patients with small cell lung cancer using real world data. *Clin Lung Cancer*. 2021; September (pre print)

[28] Chouaid C, Luciani L, LeLay K et al. Cost-effectiveness analysis of afatnib versus gefitinib for firest-line treatment of advanced EGFR-mutated advanced non-small cell lung cancers. *J Thorac Oncol*. 2012; 12(10):1496-1502

[29] Nafees B, Stafford M, Gavriel S et al. Health state utilities for non-small cell lung cancer. *Health Qual Life* Outcomes. 200;6:84

i

[30] Tolley K, Goad, Yi Y et al. Utility elicitation study in the UK general public for late-stage chronic lymphocytic leukaemia. *Eur J Health Econ*. 2013;14(5):749-59

[31] Mokkink L, Terwee C, Patrick D et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study. *Qual Life Research*. 2010;19(4):539-49

[32] Mokkink J, Prinsen C, Bouter L et al. The Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) and how to select an outcome measurement instrument. *Braz J Phys Ther*. 2016;20(2):105-13

[33] Mokkink J, L, De Vet H, Prinsen J et al. COSMIN risk of bias checklist for systematic reviews of patient reported outcome measures. *Qual Life Res*. 2018;27(5):1171-79

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

[35] Mokkink L, Terwee C, de Vet H et al. Reply to the concerns raised by McKenna and Heaney about COSMIN. *J Med Econ*, 2021;24(1):857-59

[36] Eddy D, Hollingworth W, Caro J et al. Model transparency and validation: a report of the ISPOR-SMDM Modeling Good Research Practices Rask Force – 7. *Med Decis Making*. 2012;32(5):733-43

[37] Glosson S.. *The Ephemerides of Phialo: Deuided Into Three Bookes* in The Ephemerides of Phialo (1579)

[38]

[39] Langley P, McKenna S. Fundamental Measurement and Quality Adjusted Life Years. *ValueHealth*. 2021;24(3):461[letter]