MAIMON WORKING PAPER No. 10 MARCH 2022

FUNDAMENTAL MEASUREMENT AND THERAPY RESPONSE: A PRIMER

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

ABSTRACT

The principal purpose of this Commentary is to alert those undertaking assessments of competing therapies as part of health technology assessment modeling to the critical role played by the axioms of fundamental evidence. This is of particular application to claims based on patient reported outcomes (PRO) instruments. Judged against claims based on levels of evidence, the overwhelming majority of both generic and disease specific PRO claims fail. They are unacceptable as measures of response to therapy. This overturns some 30 years of instrument development. The reason for this is quite clear: over the past decades instrument development and application reflects a profound ignorance of measurement theory by health economists, pharmacists and medical practitioners. The required standards have been known and applied, notably in the physical sciences, since the scientific revolution of the 17th century and, with specific application to the social science, since the 1960s, They have been ignored in health technology assessment or, more accurately, have never been considered.

The purpose of this commentary is to detail the required standards for measuring therapy response. None of this is new. The standards are quite straightforward, recognized for over a century. There is a single message: if you want an instrument to measure response to therapy than it has to be designed to have the required measurement properties. There are only two acceptable measurement standards: interval measures and ratio measurers, where the latter subsumes the properties of the former. A failure to recognize this renders claims for therapy response meaningless.

As simple as this requirement is, it is difficult to achieve in practice. To adopt this standard reflects a seismic memeatic shift and one that will foster significant opposition. For too long we have been prepared to put up with approximate invented evidence to support formulary decisions; to include measures of therapy response which also fail the standards of normal science.

INTRODUCTION

Accurate measurement is the key to value claims that are credible, evaluable and replicable; if measurement fails to meet these required standards then the value claim fails. The often quoted statement by Lord Kelvin (William Thomson 1824-1907), inscribed above the entrance to the University of Chicago, Social Science Research Building, is the touchstone: *If you cannot measure, your knowledge is meager and unsatisfactory.* Value claims for comparative response to therapy required in health technology assessment can only survive if they respect the axioms of fundamental measurement; so far they have failed.

The purpose of this commentary is to demonstrate that the widespread failure of leaders and their followers in health technology assessment to recognize the limitations imposed by the axioms of fundamental evidence on value claims is a disaster of e first magnitude; the insistence on the relevance in decision making of assumptions driven lifetime simulation modeling has resulted in 30 years of wasted effort. An effort that has been made all the more wasteful by the explicit rejection of the standards of normal science in favor of non-science, of metaphysics and pseudoscience (otherwise described as bunk), where approximate information takes precedence over hypothesis testing and the discovery of new, yet provisional, facts [1]. Unfortunately this commitment to the creation of imaginary claims continues with the recent CHEERS 22 guidance for submitting such claims to leading technology assessment journals [2], complementing the leading textbook in the field which for 20 years in various editions has promoted imaginary claims with the focus on the impossible or I-QALY quality adjusted life year and incremental cost per I-QALY imaginary claims for cost-effectiveness [3] [4]. In this technology assessment has the unequivocal support of professional groups such as the Institute for Clinical and Economic Review (ISPOR) and self-proclaimed US national arbiter for imaginary cost effectiveness recommendations, the Institute for Clinical and Economic Review (ICER).

FUNDAMENTAL MEASUREMENT

Following the formalization by Stevens and others in the 1930s and 1940s, scales or levels of evidence used in statistical analyses are classified as nominal, ordinal, interval and ratio [5]. Each scale has one or more of the following properties: (i) identity where each value has a unique meaning (nominal scale); (ii) magnitude where values on the scale have an ordered relationship with each other but the distance between each is unknown (ordinal scale); (iii) invariance of comparison where scale units are equal in an ordered relationship with an arbitrary zero (interval scale) and (iv) a true zero (or a universal constant) where no value on the scale can take negative scores (ratio scale). The

implications for the ability to utilize a scale to support use of arithmetic operations (and parametric statistical analysis) are clear. Nominal and ordinal scales do not support any mathematical operations; only nonparametric statistics. Interval scales can support addition and subtraction while ratio scales support the additional operations of multiplication and division, as they have a true zero. This zero point characteristic means it is meaningful to say the one object is twice as long as another. To measure any object on a ratio scale it has to be demonstrated that all criteria for an interval scale have been met with a true zero. It is impossible to take an ordinal score and translate that to a ratio score via a simple linear transformation or to put ordinal scores on an interval scale and then assume the ordinal scale has interval properties and, hopefully a true zero. This is the mistake made in applying multiattribute preference scores, with the added problem that none have true zeros. Preference scores have to meet ratio measure standards; if they produce negative values by applying their various system algorithms, described as 'states worse than death', then they fail. All generic preference scores support negative values.

Unfortunately this failure to recognize psychometric distance is pervasive in patient reported outcome measures (PROs); discussed below as the Likert fallacy. It does not apply, of course to clinical measures of response that are designed and calibrated as either interval or ratio measures. Nor does it apply to statistical series or data repositories which have known integer values and with invariance of comparison. The focus here is on PRO measures that clearly fail the standards required, which leads us to Rasch Measurement Theory (RMT), first developed in the 1960s, but seldom referred in health technology assessment as the ideal framework for creating measures to evaluate response to therapy [6].

THE UNWELCOME CHALLENGE

If we are to accept the importance of measurement in evaluating comparative therapy response then we have to overcome a meme, a meme or belief system that has been ingrained in health technology assessment for over 30 years [7] [8]. A belief system that has thousands of adherents who have never challenged the role of imaginary assumption driven simulated or modeled claims, and who believe that imaginary claims have to be accepted as gospel, as the basis for formulary decisions. While this may seem, from the perspective of normal science as ridiculous, it is the standard in health technology assessment, as attested to by many academic groups [9]. Indeed, as noted, the leading textbook in technology assessment is nothing more (or less) than a primer for constructing multitudes of assumption driven simulation models to invent deliberately non-empirically evaluable claims [10].

Fundamental measurement has, as far as can be ascertained, never been a concern of those promoting the current technology assessment meme. Neither, might it be added have the standards of normal science; proposing credible, evaluable and replicable value claims, been of interest; assuming these standards are understood. This means, in the commitment of the existing meme to approximate information, the continued embrace of metaphysics and pseudoscience[11]. This applies equally to multiattribute generic preference scores as well as to the many disease specific PRO instruments with their claims for 'measuring' quality of life as well as response to therapy.

REALISTIC ASSUMPTIONS

It is common to find belief systems with thousands of followers built entirely from assumptions that have no possibility of being challenged empirically; if you are raised or trained in an assumption driven belief system then you are likely to continue to believe in its mysteries despite its denial of common sense. Denial is common in health technology assessment with ICER, for example, believing in the ordinal preference measure actually having mystical ratio properties. These belief systems, or faith driven memes, while common are not encountered in the physical or mature social sciences, with the exception of the existing health technology assessment meme. This is unique in its belief in realistic assumptions to drive competing claims for therapies that extend decades into the future. Indeed, in a number of countries (e.g., United States, Canada, United Kingdom, Australia) academic centers in leading universities are charged by agencies such as the UK National Institute for Health and Care Excellence (NICE) and the Australian Pharmaceutical Benefits Advisory Committee (PBAC) with evaluating the realism and internal construction of assumption driven imaginary simulations submitted by manufacturers to ensure their 'realism' and consistency with the agency reference case. Parallels with religious groups and extreme political parties need not be explored.

While the more cynical observer might suggest that health technology assessment simulation modeling should be taught in a College of Theology, the fact remains that the assumption driven simulation modelling that is the mainstay of the meme defies the rules of elementary logic. We have known since the mid 18th century that we cannot assume that claims or observations from the past can support claims on the future: Hume's problem of induction. In other words we cannot secure assumptions regarding the future because we cannot observe future events. *The security of future assumptions cannot be achieved through logical argument; it does not follow from the fact that all past futures have resembled past pasts that all future futures will resemble future pasts* [12].

This does not mean that assumption do not pay a role in both the physical and mature social sciences. The difference is the empirical status of the claims made. Following Popper, science looks to falsification of claims as the key element in the discovery of new facts. If a claim is falsified following an appeal to real world data, then the role of selected assumptions can be evaluated. If the value claims, as with the health technology assessment meme are designed deliberately to be immune to any appeal to real world evidence, then we are in the belief system of non-science such as intelligent design..

To try to claim that one assumption is 'more realistic' than another is absurd; an assumption is an assumption. Belief in its 'realism' is psychological; otherwise it is a logical absurdity. This leads of course to the bizarre situation where any number of different assumptions regarding the future lead only to any number of 'competing' assumption driven simulation models. In the absence of evaluable value claims there is no basis for claiming one model is 'superior' to another. In practice, the progress of discovery of new, yet provisional, facts is abandoned or never considered; we rely on convincing formulary committees not that we have new data but that they can usefully base decisions on invented value claims that can never be challenged, unless by an other equally invalid value claims for competing therapies.

MULTIATTRIBUTE GENERIC ORDINAL SCORES

Looking back over the past 30 or more years, what stands out in health technology assessment is the attention given to instruments such as the Standard Gamble (SG) and the Time Trade Off (TTO) preference measures as well as the preference scores generated by multiattribute generic instruments such as the EuroQol (EQ-5D-3l/5) the HUI Mk2/3, the SF-36/6D and country specific variants such as the Australian AQOL. It has been universally assumed that these instruments have ratio properties. Indeed, ICER defies logic (and common sense) in going, as noted. so far as to say, without any proof whatsoever, that health economists 'have the confidence' that these instruments have ratio properties despite all evidence to the contrary [13] . If you admit that the score is actually ordinal, then ICER's business case in modelling imaginary worlds falls over. This, of course, cannot be allowed to happen.

The fact that these multivariate generic instruments have only ordinal score properties essentially destroys some 30 years of health technology assessment activities for the simple reason that as an ordinal score it can only support nonparametric statistics; it cannot support standard parametric statistical operations involving division, multiplication, additional subtraction. The preference scores cannot support value claims for response to therapy as the 'psychometric' distance between ranked ordinal preference scores is unknown. This

means, as detailed below, that these preference scores cannot be applied to create quality adjusted life years (QALYs). As well, possibly as a surprise to those critics of the QALY, the fact is that it is a mathematically impossible construct means that their criticisms of the QALY (e.g., on age and equity terms) are irrelevant. Indeed, attempting to make such criticisms could imply support for the impossible QALY, instead of dismissing it out of hand.

In fact, there is no evidence whatsoever that these various direct preference measures such as the SG and TTO and multiattribute instruments have anything other than ordinal properties. The algorithms that are applied to create the various scores involve attaching TTO or equivalent weights were never designed to have to create ratio measures. In the case of the EQ-5D-5L, patients describe their current health status in terms of 5 symptoms (or attributes): mobility, self-care, usual activities, pain/discomfort and anxiety/depression and 5 response levels: no problems, slight problems, moderate problems, severe problems or extreme problems (with some wording variations). The response levels are, of course, ordinal but are combined for a single valuation for the 3,125 possible health states. Weights are attached to the responses as part of an algorithm designed to create a preference score in the range 1 = perfect health and 0 = death. The developers hoped in developing the preference score algorithm, fitting (or tweaking) their model to the data, that the individual patient preferences would fall in this range (the ratio properties of the preference scores were assumed; at least no attention was in developing the instrument to ensure the standards of fundamental measurement were met). The algorithm works by subtracting preference decrements from unity, hoping that under no circumstances would the algorithm 'overshoot'; and create negative score. Unfortunately, much as they attempted to 'squeeze' the preference scores, the preference algorithms generate negative scores or 'states worse than death'. Much as efforts were made to hide or dismiss the implications of the presence of negative preferences, the outcome remains: there is no absolute zero for the preference scores. A recent attempt to produce US valuations for the EQ-5D-5L with the 3,125 health states found that 20% of the health stats had negative values (N=620) or the euphemistic states worse than death [14]. Once again, a classic example of attempting to fit a model to the data, rather than attempting to develop an instrument with the required measurement properties. Fitting a model to the data (or attempting several model fits) guarantees, by defauly, that the resulting instrument will have only ordinal measurement properties and, unless tweaked will not automatically create preference in the required zero to unity range with interval scores. This applies to the SG and TTO as well as the various multiattribute instruments.

As well as the failure to create a ratio score for the various preference instruments, they suffer from two additional characteristics: a lack of dimensional homogeneity and hence

construct validity [15]. The preference instruments are characterized by either asking respondents to 'value' a basket of attributes (including the SG and TTO) or to detail current experience with symptoms and response which may be of no interest to patients in a specific disease state. If a campaign to discredit the QALY is proposed, then a first step should be to criticize the preference scores and the algorithms on which they are based. This has never been done. For purposes of response evaluation, all multiattribute preference instruments should be abandoned.

THE IMPOSSIBLE QALY (I-QALY)

The pre-eminent role the QALY has played over the last 30 years in health technology assessment and ersatz value claims is undeniable. Unfortunately, this role is totally misplaced because, once fundamental measurement is considered, the QALY (or I-QALY) is mathematically impossible [16]. Given this fairly obvious criticism, the question is why has the I-QALY continued to play a central role in health technology assessment. One answer, which goes to a lack of knowledge of measurement theory, is simply a failure to understand, or even be aware of, the axioms of fundamental measurement. Another response is to claim that while the case for an ordinal score is clear cut, technology assessment needs to believe the preference scores are ratio measures (in disguise or one more assumption) in order to apply the cost-effectiveness claims of the technology assessment model with an ersatz blanket claim for cost-effectiveness. This requires a belief among participants in health technology assessment claims that we are actually dealing with a ratio measure that lacks ratio properties; overlooking the facts that our mystical imaginary ratio scale lacks dimensional homogeneity, construct validity, a true zero and invariance of comparisons.

The issue of relativism is relevant here. There is a tradition, the so-called strong program in the philosophy of science, where the 'content of science' has a sociological explanation. As Wootton describes it, the essence lies in the principle of symmetry: *the same sorts of explanations must …be given for all types of knowledge claims, whether they are successful of not* [17]. If an explanation is sought for a belief, such as the approximate information meme, we should look to a sociological/psychological explanation. This puts forward the assertion that it is not the role of science to provide a framework for coming to grips with reality. It is illegitimate to say that we can exclude the approximate information meme because it is wrong or that we have good evidence for it being wrong. It stands on an equal footing with a competing meme that recognizes fundamental measurement and the standards of normal science. As Wootton notes, if we accept that the content of science can be explained sociologically and not in the way science is organized and the aspirations of scientists, *then it*

systematically excludes from consideration the feature that makes scientific arguments distinctive: their appeal to superior evidence [14]**.**

A further attraction of the QALY is that it is the ideal bundled metric. By putting the I-QALY as central to the modelled imaginary claims which include assumptions as to unknown future costs and unknown future responses to therapy (both clinical and PRO) the analyst combines (or cobbles together), by assumption, all the various elements that are required to support a universal claim for cost effectiveness. This of course, appeals to the media and others who can proclaim a therapy is cost-effective at a cost-per-QALY threshold value that ICER (or an agency, and even a manufacturer who has paid for a model) has determined. No one looks critically at the model to determine how well dressed the emperor actually is.

A more worrying concern is that the belief in the I-QALY is held so strongly because the I-QALY is an impossible construct. This is a common feature, that mystery is a virtue, even for the more acceptable religions; as Tertullian is quoted *It is by all means to be believed, because it is absurd* or Sir Thomas Browne's claim: *Methinks there be not impossibilities enough in religion for an active faith* [18]*.* Presumably those inhabiting the technology assessment meme would respond: 'Amen'.

To be frank, it is time we stopped criticizing the mystery of the QALY as though it had any meaning. It is a mathematically impossible construct; that is the end of the story. It is time that we turned our attention to measures of quality of life that met the standards of fundamental measurement, representing a coherent, single attribute or latent construct. This has been achieved (detailed below) in the need-fulfillment (N-QOL) measure as a bounded ratio score.

Unless there is an understanding of the limitations imposed by the axioms of fundamental measurement in health technology assessment techniques constructs such as the QALY will continue to be promoted and accepted by less-informed decision makers, driven by the beliefs of those accepting the existing technology assessment meme. If challenges are to be mounted, the challenger and the audience must be in a position to understand the challenge. This is seldom the case. It is not as though the standards are overly technical; indeed it is the opposite. Once standards are recognized the failure of instrument developers is readily apparent. Unfortunately, even when made clear to the authors of modeled imaginary cost-per-I-QALY claims that these lack any pretense to scientific credibility, they persevere.

IMAGINARY WORLDS

The fundamental difference between health technology assessment and the physical sciences, and the more sophisticated social sciences, is the reliance upon invented evidence to support formulary submission. This was decided in the early 1990s when 'leaders' in the field abandoned hypothesis testing as too time consuming to fill evidence gaps in favor of approximate invented evidence to fill those gaps. This enabled them, when products were first approved by the FDA, to overcome any constraints imposed by limited evidence. ICER occupies pole position in this race. Rather than proposing a research strategy to fill gaps, the evidence was invented. It was assumed that formulary committees were perfectly happy to list products, determine access to therapies and negotiate prices on invented non-evaluable claims. The smoke screen was a claim for the relevance of 'approximate information' defended by sensitivity analyses, in particular the imaginary probability sensitivity analysis, as a defense for ersatz value claims delivered by lifetime assumption driven imaginary and non-evaluable simulations; no one asked the question 'approximate to what'. In fact, once the limitations of fundamental measurement and simple logic are recognized there is nothing to be approximated, unless to future events that are, by definition, unknown.

There is, of course, not just one ideal imaginary future world. Certainly, NICE and its inquisitors for the structure and content of their choice of an imaginary world, may prefer one imaginary construct over another, but there is no logical basis for choosing one set of assumptions over another to support claims for pricing and access. The fact remains that there is a potential multiverse of models each supporting and equally justifying thresholds and claims for ersatz cost-effectiveness, pricing and access; this effectively disables any specific model claim.

The 'success' of the approximate information or imaginary simulation was due in large part to the enthusiastic advocacy of academic groups (and the uptake of imaginary modelling by NICE in the UK). The ability of these thought leaders to maintain the transmission fidelity of the meme through their acolytes globally with agencies such as ISPOR was crucial; and will continue to be so. This is not to say that red flags were not posted; they were either overlooked or just ignored[19] [20] [21] .

PATIENT REPORTED OUTCOMES: THE LIKERT FALLACY

The Likert scale, first proposed in 1932, is a five or seven point scale which allows the respondent to express how much they agree or disagree with a statement. Likert scales are used widely in disease specific PROs, although the various authors fail to appreciate the fact

that they are ordinal scales and however you manipulate them they remain ordinal or, more simply, impossible mathematical constructs. As such they can tell us nothing about therapy response, apart from claims based on non-parametric statistics. They are not, it should be noted, 'measures'; a term that applies exclusively to interval and ratio constructs and not to ersatz creations that are claimed to be equivalent..

Consider, as an example of the Likert fallacy, the Asthma Quality of Life Questionnaire (AQLQ) [22]. Although widely used over the past 25 years, the AQLQ fails to meet the standards of fundamental measurement; it is a multiattribute ordinal scale. Consider how the AQLQ is assembled and the measurement implications of Likert scales. 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 [23]. The fact that the AQLQ has shown strong classical measurement properties based on integer ratio assumptions, is irrelevant; this only occurs if you ignore the axioms of fundamental measurement and assume the AQLQ has interval properties for the Likert scores (which could equally well be designed on a 7 point scale as A >B >C >D >E >F > F >G rather than with a numeric assignment 1 >2 > 3 >4 >5 >6 >7 > 8 (or even an emoji for each Likert space).

The assumptions made, without any justification, is that (i) all items in the AQLQ are of equal difficulty, irrespective of sub-domain; (ii) the distances between the scale integer values are psychometrically equal; (iii) that where a number of Likert scales are scored and aggregated the this claim for invariance of comparisons is identical for all distances for each scale and identical item difficulty; (iv) that each Likert scale has the same weight in the overall scores for sub-domains and the total score (unless arbitrary weights are applied) and (v) that the aggregation over Likert scales actually produces, not an interval measure, but a ratio measure. In effect, the Likert scale which is ordinal can be believed to be translated, without proof, into a required ratio scale. This is a reach too far. Assumptions have to be justified;

what we don't need at this juncture is to defend aggregation of Likert scales on the basis this this just one more acceptable assumption in the commitment to metaphysics and pseudoscience in therapy evaluations.

Certainly, there is an extensive literature and debate between those who believe ersatz interval scores can be 'created' from Likert scales, introducing concepts of a balance and symmetry in designing Likert scales, and those who maintain they are just ordinal scores. This overlooks RMT and the creation of interval measures from ordinal scores in a mathematically acceptable process.

RASCH MEASUREMENT THEORY

If we are to escape the morass of value claims that fail to meet the requirement of fundamental measurement, one avenue is to consider the application of RMT to developing value claims that meet the required evidence standards for response to therapy. It cannot be assumed, *ex post facto*, that a given scale has interval, invariance of comparison, or ratio properties. This point is made by Bond and Cox [12] in their discussion of Rasch Measurement Theory (RMT) theory and its contribution to fundamental measurement: in traditional test theory (TST) and item response theory (IRT) the observed data have primacy; results are exploratory and descriptive of those data. Rasch models are, on the other hand, confirmatory and predictive; a confirmatory model requires the data to fit the model where, following the principles of conjoint measurement, they are sufficiently realized to claim the results are a measurement scale with interval measurement properties detecting measurement structures in non-physical attributes [12].

The Rasch model is designed to analyze categorical data where the likelihood of a positive response is a function of the trade-off between item difficulty and the respondent's abilities or proficiency, as locations on a continuous latent variable (e.g., need fulfillment quality of life). The object in RMT is to develop an index of response, from ordinal scales, that has interval level properties. In certain circumstances this index can be translated to a bounded ratio scale (a true zero). This achieves, for the first time, a PRO measure that has the required properties for meaningful response to therapy supporting multiplication and division.

In the case of the recently proposed need fulfillment quality of life measure (N-Qol), this meets all required RMT standards as well as allowing the analyst, for the N-QoL class of attributes for specific disease areas, to develop a bounded ratio scale [24]. This allows direct comparison of this dimension of benefit as well as allowing, if required, the construction of quality adjusted time spent in disease states; it effectively, eliminates the I-QALY.

EXEUNT COST-EFFECTIVENESS

A central belief of the current technology assessment meme is that we can make blanket claims from the assumption driven simulation that one product is superior, or more effective, than another. While this is an impossible claim given the imaginary nature of the modelling and its reliance on the mathematically impossible QALY, it points to a key fact: competing value claims for therapy options cannot be reduced to a single metric. The axioms of fundamental measurement are quite clear: measurement must relate to single attributes, whether physical or latent. We can only combine single attributes if they each have ratio measurement properties (a fact that seems to have escaped the attention of supporters of multiattribute decision theory). The easiest example is the ratio scale body mass index which combines two ratio scales: height and weight.

While it might appear less elegant than the ersatz blanket claim for cost-effectiveness, value claims for competing therapies must rely on an agreed set of attributes spanning clinical measures, PROs (including quality of life as a single attribute) and resource utilization. All can be separately defined by an assessment protocol, submitted to formulary committees. It is up to the formulary committee to judge the merits of the value claims as part of a negotiation to set provisional pricing and access requirements. These would be subject to the evaluation of the attribute reported back to formulary committee in real time and, potentially, part of longer term commitment to discover new facts.

In short, if we are to reject a belief system that is clearly a failure, then we need to consider a successor. The focus must be on single attributes as the basis for value claims. This applies clearly to patient reported outcomes (PROs) but also to clinical claims and those for resource allocation. Each PRO must be based on a coherent latent construct, with the measure having unidimensional and construct validity properties. It must be empirically evaluable, supported by an assessment protocol, which proposes a timeline within which the results of an empirical assessment can be reported to a formulary committee. This means, of course, dispensing with the overwhelming majority of existing generic and disease specific PRO measures. Measurement and response are the keys.

CONCLUSIONS

The ability to make precise measurements …. remains an essential step in building explanatory bridges, in explaining, predicting and controlling the nature of subjective experience ….measurement turns the qualitative into the quantitative, the vague into the precise [25] **. The existing health technology assessment meme is not about measurement;**

unless you are prepared to jettison the recognized axioms of fundamental measurement. The fact that the current meme for health technology assessment has survived for 30 years is testament, not to its intellectual dominance, but to the failure of critics to appreciate its manifest failings, notably its inability to respect the standards of normal science. It is relatively easy to document these manifest failings; what is missing is the ability of critics to address these failings and present an alternative analytical framework to support value claims. As long as ICER, ISPOR and other organizations, including academic research centers, feel they can deflect criticisms (or just ignore them), ICER and the assumption driven I-QALY simulations will continue to be the basis for pricing recommendations and the denial of care. It is undeniable, however, that escaping the belief system that has been in place for 30 years will not be easy; the many followers have to be convinced that truth is not consensus. Mass Damascene conversions are unlikely. in the approximate information or assumptions simulation meme, we have no possibility of ever being able challenge imaginary claims; we have no idea if they are right or if they are wrong; we will never know and we were never intended to know.

The purpose of this commentary has been to set out the basis for ignoring ICER recommendations. The task for patient groups and others is to sustain these criticisms for a wider and wider audience. The principle obstacle is the ability to understand and sustain the arguments against the ICER reference case. The intellectual challenge is not difficult; all we need to recognize is the failure to meet the demarcation test between science and non-science. Once it is accepted that 'science' requires all value claims to be credible, evaluable and replicable, the case is made. Each value claim should be for a single attribute that meets ratio (or at least interval) measurement properties. Value claims can be for clinical attributes, patient reported outcome (PRO) attributes and for resource utilization and compliance.

As noted, once these requirements are met we can consign 30 years of assumption driven simulation models to the shredder together with ersatz claims that a product is 'cost-effective'. Modeled I-QALY driven claims for thresholds and a 'social' price are similarly rejected. Instead, we focus on the discovery of new yet provisional facts regarding therapy impact and the commitment to an ongoing research program to support future claims as part of disease area and therapeutic class reviews.

REFERENCES

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

[2] Husereau D, Drummond M, Augustovski F et al. Consolidated Health Economic Evaluation Reporting Standards 2022 (CHEERS 22) Statement: Updated reporting guidance for health economic evaluations. *ValueHealth*. 2022;25(1):3-9

[3] Drummond M, Sculpher M, Claxton K et al. Methods for the Economic Evaluation of Health Care Programmes (3rd Ed.). New York: Oxford University Press, 2015

[4] Langley P. Nothing to Cheer About: Endorsing Imaginary Economic Evaluations and Value Claims with CHEERS 22 [version 1; peer review: awaiting peer review] F1000Research 2022, 11:248 https://doi.org/10.12688/f1000research.109389.1 First published: 28 Feb 2022, 11:248 https://doi.org/10.12688/f10

[5] Stevens S. On the theory of scales of measurement. *Science*. 1946;103: 677-80

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

[7] Langley P. Nonsense on Stilts – Part 1: The ICER 2020-2023 value assessment framework for constructing imaginary worlds. *InovPharm*. 2020;11(1): No. 12
https://pubs.lib.umn.edu/index.php/innovations/article/view/2444/2348

[8] Langley P. Peter Rabbit is not a Badger in disguise: Deconstructing the Belief System of the Institute for Clinical and Economic Review in Health Technology Assessment. *InovPharm*. 2021;12(2):No. 2
https://pubs.lib.umn.edu/index.php/innovations/article/view/3992/2855

[9] Langley PC and McKenna SP. Measurement, modeling and QALYs. *F1000Research 2020*, 9:1048 https://doi.org/10.12688/f1000research.25039.1

[10] Drummond M, Sculpher M, Claxton K et al. Methods for the Economic Evaluation of Health Care Programmes. 4th Ed. New York: Oxford University Press, 2015

[11] Neumann P, Willke R, Garrison L. A health economics approach to US value assessment frameworks – Introduction: An ISPOR Special Task Force Report (1). *Value Health*. 2018;21:119-123

[12] Magee B. Popper. London: Fontana, 1974

[13] Langley P. The Impossible QALY and the Denial of Fundamental Measurement: Rejecting the University of Washington Value Assessment of Targeted Immune Modulators (TIMS) in Ulcerative Colitis for the Institute for Clinical and Economic Review (ICER). *Innov Pharm*. 2020; 11(3): No 17
https://pubs.lib.umn.edu/index.php/innovations/article/view/3330/2533

[14] 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

[15]McKenna S, Heaney A. Composite outcome measurement in clinical research: the triumph of illusion over reality. *J Med Econ*. 2020 DOI: 10.1080/13696998.2020.1797755

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

[17] Wootton D. The Invention of Science: A new history of the scientific revolution. New York: Harper Collins, 2015

[18] Dawkins R. A Devil's Chaplain. New York: Houghton Miflin, 2004

[19]Merbitz C, Morris J, Grip J. Ordinal scales and foundations of misinference. *Arch Phys Med Rehabil*. 1989;70:308-12

[20] Tennant A, McKenna S, Hagel P. Application of Rasch analysis in the development and application of quality of life instruments. *Value Health*. 2004;7(Suppl 1):S22-S26

[21] Grimby G, Tennant A, Tesio L. The use of raw scores from ordinal scales: Time to end malpractice. *J Rehabil Med*. 2012;44:97-98

[22] 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

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

[24] Langley PC. Value Assessment in Cystic Fibrosis: ICER's rejection of the axioms of fundamental measurement. *Inov Pharm*. 2020;11(2): No. 8 https://pubs.lib.umn.edu/index.php/innovations/article/view/3248/2395

[25] Seth A. Being You: A new science of consciousness. New York: Penguin Books, 2021