MAIMON WORKING PAPER No.25 NOVEMBER 2020

TO DREAM THE IMPOSSIBLE DREAM: THE COMMITMENT BY THE INSTITUTE FOR CLINICAL AND ECONOMIC REVIEW TO REWRITE THE AXIOMS OF FUNDAMENTAL MEASUREMENT FOR HEMOPHILIA A AND BLADDER CANCER VALUE CLAIMS

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

Abstract

Understandably, after 30 years of ignoring the axioms of fundamental measurement, advocates of creating approximate information through the construction of lifetime cost-per-QALY worlds are somewhat unnerved by the realization that their methodology is incompatible with those axioms. This is made all the more unnerving when it is pointed out that this incompatibility was pointed out over 30 years ago, following the formalization of those axioms almost 80 years ago. Why this was overlooked is a mystery. The result was a commitment to the application of ordinal utility and other patient reported outcome measures to support claims for response to competing therapies; most egregiously, the advocacy of cost-per-QALY lifetime models and willingness to pay thresholds to support recommendations for pricing and access to pharmaceutical products and devices. Although this incompatibility has been pointed out in respect of simulation modeling, groups such as the Institute for Clinical and Economic Review (ICER) they press on, producing evidence reports and recommendations for emerging products that fail the standards of normal science. While these are an analytical dead end, ICER has nowhere else to go. This is their business model; to admit otherwise would mean withdrawing their many evidence reports and admit they were wrong. ICER has rejected this; rather it has decided, together with its academic consultants, to challenge the axioms of fundamental measurement, to produce a parallel measurement universe that can sustain QALYs and the imaginary simulation lifetime models. The purpose here is to make clear that ICER is manifestly wrong and that there is no way it can maintain its credibility in pursuing this path. This is achieved by a deconstruction of the arguments put forward by ICER to defend its new vision of the axioms of fundamental measurement, a vision which provides a case study in the distinction between justified belief and opinion. Fortunately, we have the framework for a new paradigm in value assessment; a paradigm that recognizes the standards of normal science and rejects belief in an alternative reality consistent with fundamental measurement axioms.

Key Words: measurement axioms, alternative reality, impossible belief, I-QALY

Introduction: The Road Not Taken

Abandoning the standards of normal science in health technology assessment was a deliberate decision. Rather than adopting a research program for new products that focused on meeting and reporting on evidence gaps, providing credible, empirically evaluable and replicable claims, leaders in the field chose to create approximate information [1]. At product launch, it was

recognized that information over and above that created by Phase 2 and Phase 3 clinical trials was limited. The solution was not to attempt to meet evidence gaps with real world evidence but to create assumption driven lifetime cost-per-QALY 'claims' that were impossible to validate; claims credibility, empirical evaluation and replication were thrown out., Instead we have a range of technical standards to apply to these imaginary lifetime constructs such as probabilistic sensitivity analysis to try and convince decision makers that the imaginary model with its approximate information is in the right ball park; and that we can use this to pivot through scenarios, claims for incremental cost-per-QALYs and the application of willingness to pay thresholds to make recommendations for pricing and access.

Unfortunately, even those advocates of assumption driven approximate non-evaluable information failed to consider a fact central to the standards of normal science: the limitations imposed by the axioms of fundamental evidence. Even if one might concede that approximate information filled some imaginary information void to some unknown extent, the analyst still runs into a brick wall: the pervasiveness of ordinal scales. This is seen most obviously in the creation of multiattribute utility scales, the foundation for creating QALYs. In order to multiply time spent in a disease stage by an index of utility on a range 1 = perfect health and 0 = death you need a utility scale that has ratio properties; to support multiplication it must have a true zero. Utility scores clearly do not have this property (as negative utilities can be created); nor do utility scales have interval properties; even if they did they could not create QALYs.

The advocates of what has now come to be known as the imaginary or approximate information I-QALY paradigm, although warned on a number of occasions [2] [3], failed not only to see that the QALY was a mathematically impossible construct (hence the term I-QALY) but that exeunt the I-QALY the entire lifetime cost-per-QALY simulation for approximate information paradigm (or more properly a meme) collapses. We have, in effect, wasted 30 years pursuing a will o'the wisp mathematically impossible analytical framework to support formulary decisions. As Greene notes: 'Of course, truth in science is not determined by polls or popularity. It is determined by experiments, observations, and evidence [4]. Unfortunately, in health technology assessment 'truth is consensus' [5], irrespective of its failure to acknowledge the standards of normal science. It is pseudoscience or pure bunk [6]

Believing the Impossible

The term meme is used deliberately; the term paradigm for the I-QALY approximate information belief system is too strong as it implies a process where, within the framework of normal science, a new framework of analysis emerges to resolve issues but which still accommodates the previous paradigm. Questions that the previous paradigm could resolve, at least provisionally, are still resolved within the new paradigm but questions that could not be resolved now face provisional resolution. We face a quite different potential for transformation: a new paradigm that meets the standards of normal science cannot accommodate the I-QALY approximate information meme as there is no common ground, or common acceptance of evidentiary standards between them.

The extent to which the I-QALY approximate information meme is held should not be understated. For 30 years generations of students and instructors have been told to put the

standards of normal science to one side. Many single payer health systems, most notably the National Institute for Health and Care Excellence (NICE) and the Pharmaceutical Benefits Advisory Committee (PBAC) in the UK and Australia respectively have assiduously promoted approximate information and gatekeeping I-QALY thresholds as sacrosanct; even to the extent of contracting with academic research centers to act as inquisitors in policing the purity of the imaginary simulations developed by manufacturers [7] [8].. Add to this the missionary endeavors and support for the meme's transmission fidelity through advocacy of good practice guidelines in outcomes research reports in the construction of imaginary worlds by the International Society for Pharmacoeconomics and Outcomes Research (ISPOR); most recently in its 2017 release of a series of ISPOR Special Task Force Reports [9] and in 2019 a review of the use of health state utilities to calculate QALYs [10]. In none of the ISPOR publications is there any reference to fundamental measurement; this will be a difficult position to draw back from after 20 years of enthusiastically and uncritically endorsing utilities as ratio measures. Transmission fidelity is also reinforced by journal editors who have been assiduous, with few exceptions, in rejecting papers that challenge to meme belief system. Given these firewalls, there will be, no doubt, substantial resistance to abandoning the I-QALY approximate information meme. After all, it has been argued that the most impossible the belief the stronger it is held. Missionaries are adept at reconciling apparently absurd claims in marketing their product [11]. Last but not least is the Institute for Clinical and Economic Review (ICER) in the US which has assiduously embraced the meme in contracting to academic centers for approximate information I-QALY simulations to support pricing and access recommendations.

Understanding Fundamental Measurement

In the physical sciences, and the more rigorous, and aware, social sciences such as economics, an understanding of the axioms of fundamental measures are recognized and are considered essential in instrument development [12]. Following the formalization by Stevens and others in the 1930s and 1940s, the axioms of fundamental measurement are well understood [13]. The measurement scales used in statistical analysis are nominal, ordinal, interval and ratio. Each scale of measurement has one or more of the following properties: (i) identity where each value has a unique meaning; (ii) magnitude where ordered values on the scale have an ordered relationship with each other but the distance between is unknown; (iii) invariance of comparison where scale units are equal to each other in an ordered relationship and known; and (iv) a true zero where no value on the scale can take negative scores. The implications for the ability to utilize a scale to support arithmetic operations (and parametric statistical analysis) are clear cut. A nominal scale is just a set of unique meanings but nothing else (e.g., gender). An ordinal scale has identity and magnitude in an ordered relationship but we do not know the distance between the values (i.e., it cannot support arithmetic operations, only non-parametric statistical evaluations, modes and medians). An interval scale has known differences but no true zero and can support only addition and subtraction (i.e., it can change the point on an integer line but only relative to other points). A ratio scale can support the additional operations of multiplication and division because it has a true zero (i.e., change the point on an interval line relative to zero).

We cannot assume a given scale (e.g., utilities) is an ordinal, interval or ratio scale. The default scale is an ordinal scale. Unless a scale is designed to have interval or ratio properties it is an ordinal scale; a ranking of raw scores. Understanding this points to the importance of Rasch

measurement theory (RMT) which is quite clear in that it is only possible to create an interval scale if there are techniques of translating raw scores or ordinal values to an interval scale [14].

Understanding Construct Theory in Instrument Development

Objectivity is a fundamental requirement of valid measurement [15]. If this is achieved then a unit amount of the variable being measured maintains its size irrespective of the instrument being used or what is being measured (e.g., thermometers). Local objectivity is defined by relative differences between locations or invariance of comparisons. In the case of PROs this is achieved by application of Rasch Measurement Theory (RMT). Depending on the latent construct or attribute being measured an individual scoring higher is assumed to have more of what the construct is measuring. Failure to meet RMT standards means that claims for relative difference are impossible. The instrument is in the default ordinal state.

General objectivity takes one further step and considers the absolute, not the relative position of objects. In the sciences, this is approximated by measures that are independent of the respective instruments and conditions of measurement. The absolute location implies a defined absolute zero and measurement units. While this is achieved with a variety of instruments in the physical sciences, its application in the social sciences with the focus on latent constructs is more problematic. Certainly, we could utilize a construct theory to build calibration equations to specify and maintain a zero point and unit of measurement independent of any instrument and indication, but this is considered practically impossible for latent or non-physical constructs. If meeting the standards for general objectivity through calibrating instruments or linking the construct theory to scores created by the instrument is difficult if not impossible, then we have for the moment to put to one side any thought of instruments attempting to capture a construct anchored on an absolute or true zero with invariance of comparisons. The claim that an instrument has these ratio properties is nonsensical; at best we are dealing with instruments attempting to measure latent constructs that have interval properties. This is the contribution of RMT.

Three qualities are necessary for a PRO measure to be truly valid in assessing health outcomes [16]. These are:

- The instrument should be based on a coherent construct theory or conceptual model of the outcome to be measured
- There should be a specification equation to link the construct theory to scores produced by the questionnaire
- Data collected with the instrument should fit Rasch Measurement Theory (RMT) to translate ordinal to interval scores and achieve local objectivity (i.e., an interval scale)

If an outcome measure is to be truly valid, the scores it generates should be predictable from its construct theory by means of a specification equation. Achieving this would match the quality of measurement in the physical sciences. To date, the absence of specification equations equation means that we have a qualified yet valid instrument. Given a coherent construct and the guide of

RMT, we have the framework and the tools to develop a PRO instrument that can be justified as providing a meaningful estimate of response to therapy.

The failure of the EQ-5D-3L to provide a valid interval PRO measures is because it fails on all three criteria. It fails from first principles. A PRO measure must have a clearly defined construct; a conceptual model that is clinically meaningful and interpretable; defining a variable or an outcome in terms of a model with a limited predictor variable set. The EQ-5D-3L is based on a set of generic symptoms deemed appropriate by clinicians to capture what they define as a measure of health related quality of life (HRQoL) with some 'agreement' by patient focus groups. This is diametrically opposed to item generation from a latent construct theory of quality of life (QoL). If we want to understand the impact of therapy interventions then we must start with the patient to generate questionnaire items that are consistent with the conceptual model driving instrument development. Rather than HRQoL the focus should be on the needs of patients and the impact of therapy options on meeting those needs for target patient groups within disease areas. This is the only basis for linking the instrument to the latent construct. The recommended model structure is RMT. The most important requirements of a latent PRO measure are a credible construct theory and a fit to RMT.

Consider needs based QoL in contrast to HRQoL multiattribute measures of utility where the EQ-5D-3L is the classic example. The EQ-5D-3L lacks a conceptual framework, a construct, which might guide the items selected to measure HRQoL. It is just a selection of symptoms and functions defined by 3 response levels which may be of no interest or relevance to patients in calibrating response to therapy. The needs based QoL on the other hand has a well-defined construct designed to provide a framework for item selection and hypothesis testing.

Constructs and Dimensional Homogeneity

Advocates of multiattribute utility or preference scores fail to recognize that constructs refer to single attributes. The response scale must be unidimensional, reflecting the dimensional homogeneity of the construct. The axioms of fundamental measurement are quite clear in rejecting composite measures; that is, a measure that is made up of one or more variables that are related either conceptually or statistically. Examples in PROs abound; in fact the majority of PROs, whether generic such as the EQ-5D-3L or disease specific such as the PHQ-9 for anxiety and depression, are composite measures generating a single score by adding or combining different variables to create an ordinal score.

Accurate measurement of latent constructs such as QoL requires unidimensionality. All items in a scale (e.g., a QoL needs scale) should reflect a single construct if a single score is to have any meaning. This is achieved by the specification of a response model where the probability of a given item response is linked to characteristics captured in the construct. The most widely used response model is the Rasch model.

Dimensionality refers to the characteristics of quantities or items; these can only be compared if they have the some dimension [17]. If not, as in the typical case of composite measures, dimensional homogeneity breaks down; the measure lacks construct validity. The EQ-5D-3L is a dimensionally heterogeneous composite measure as it adds together symptoms that

are dimensionally distinct. Interval and ratio scales are only viable if they measure a single attribute. That is, they are unidimensional or dimensionally homogeneous. Response to therapy is defined in terms of single attributes. Attempting to create multiattribute scales yields impossible values or raw scores . The EQ-5D-3L utility, for example is not only a raw ordinal score but the score itself is an impossible multiattribute composite which fails the standard of dimensional homogeneity. Certainly, we can combine dimensionally distinct unidimensional scores to create a composite score (e.g., body mass index) but this requires the components (mass, height) to be ratio scales (i.e., with a true zero).

Combining attributes with different dimensions into a single composite score creates its own problems. Depending on the symptoms and descriptors, a composite score such as the EQ-5D-3L may create a quite different response from the HUI Mk 3. Response may be attributable to only a subset of the symptoms covered and the response levels for those symptoms. Gaming of composite scores is possible where a score is chosen because it favors a particular product response. If a product is indicated for depression but has significant side effects, then choose or create a composite score that captures depression but neglects particular side effects. It is more appropriate to focus on reporting specific attributes.

Paradigm Failure

We are, in fact dealing with two different measurement paradigms; one which conforms to the standards of fundamental measurement and one that does not. Unfortunately, probably by accident rather than design, the approximate information I-QALY meme (not paradigm) has locked itself into a parallel measurement universe which has no link to reality. RMT is not compatible with either classical test theory (CTT) or item response theory (IRT). They are, as Bond and Cox point out, competing paradigms. RMT takes the perspective that if the instrument is to meet fundamental measurement standards then we should adopt the Rasch *data-to-model* paradigm. If we are not concerned with, or are happy to ignore, questions of fundamental measurement, then we can follow the CTT or IRT *model-to-data* paradigm. The key distinction is that *RMT uses the measurement procedures of the physical sciences as the reference point.* We can aim for the standards in the physical sciences by, as Stevens pointed out in the 1940s, allocating numbers to events *according to certain rules.* It is these rules that comprise RMT. To reiterate: RMT is designed to construct fundamental interval measures. CTT and IRT focus on the observed data; these data have primacy and the results describe those data. RMT provides a framework for translating single attribute ordinal scores to interval scores. As Bond and Cox emphasize: in general, CTT and IRT are *exploratory* and *descriptive* models; the Rasch model is *confirmatory* and *predictive.* If RMT is ignored then, by default, instruments utilizing Likert scales or similar frameworks will fail to meet the required axioms of fundamental measurement and remain ordinal scales. This means that claims for response to therapy utilizing CTT or IRT do not meet the required measurement standards; they are not interval scales.

Defending the I-QALY Belief Meme

As illustrative of the confusion that arises when the inappropriate use of ordinal or raw utility scores is pointed out, we can consider the response of ICER to questions raised in the public review period responding to comments on two recent ICER draft evidence reports for hemophilia

A and bladder cancer [18] [19] [20] [21]. Both the I-QALY hemophilia A and bladder cancer simulated lifetime I-QALY models; these were created by the Center for Pharmacoepidemiology and Pharmacoeconomics Research, College of Pharmacy, University of Illinois at Chicago. The evaluation reported here follows from a previous assessment of ICER responses for their ulcerative colitis draft evidence report [22] [23]. As will be demonstrated, ICER and its academic consultants have no apparent appreciation of the role of fundamental measurement. Until the questions were raised, notions of the different properties of ordinal, interval and ratio scales were apparently quite foreign.

ICERs first line of defense in utilizing the I-QALY (typically based on the EQ-5D-3L instrument) is that 'everyone else does it' (bladder cancer) or, as stated in the case of ulcerative colitis, ICER, in common with other health economists 'has an understanding' that instruments such as the EQ-05D-3L have ratio measurement properties.

ICER's response to questions regarding the ability to demonstrate that the EQ-5D--3l utility scale has ratio properties was, in the hemophilia A response, to cite a recent paper on setting dead at zero [24]. This paper did not address the issue of how negative utilities were consistent with an assumed ratio scale but merely asserted that they were consistent. The 'proof' presented of the value of zero (as a true zero) lacked credibility. If the authors wanted to argue that if the utility scale was a ratio scale in disguise, then this had to be by assumption. Setting dead at zero, with states worse than death, does not invalidate the axioms of fundamental measurement. We can't overturn these axioms except by assumption.

ICER's response to the question of the measurement properties of the EQ-5D-3L in the bladder cancer report is even more confusing. ICER makes its case for the I-QALY and the setting aside of any concerns with the axioms of fundamental measurement in the following responses:

- Cost-effectiveness analyses including cost per-QALY estimates have been used for decades by academic researchers, international health technology assessment agencies and pharmaceutical manufacturers

  Response: this is beside the point if they have failed to recognize the limitations of fundamental measurement and the mathematical impossibility of creating a QALY. After all, the use of leeches was abandoned as a key medical technique after centuries of use. The widespread 'belief'; in these cost-effectiveness analyses is indicative of a fundamental misunderstanding of the standards of normal science and, in particular measurement theory.

- There is a 'widely held belief' that the EQ-5D-3L and -5L can estimate scores for 243 and 3125 health states and is 'widely accepted' to have interval properties (referencing a paper by Weinstein et al [25]) for criteria for multiattribute utility instruments to be considered for estimating QALYs.

  Response: Certainly these instruments can provide, ignoring the absence of construct validity, raw scores for these number of health states (including negative utility scores) by applying the EQ-5D-3L scoring algorithm. But this does not mean the scores have any

intrinsic meaning let alone interval or even ratio properties (we may 'believe' and 'accept' but that is not evidence for measurement). Part of the problem is that these scores are typically presented on a number line with equal intervals which gives the false impression that they have interval properties. You might more usefully put these scores as a ranked column vector without knowing the distance between the scores. In any event, the multiattribute framework for construction these dimensionally heterogeneous scores is sufficient for their rejection.

As to the Weinstein et al paper, their only mention of measurement is in the following paragraph (pg. 1):

> *Health states must be valued on a scale where the value of being dead must be 0, because the absence of life is considered to be worth 0 QALYs. By convention, the upper end of the scale is defined as perfect health, with a value of 1. To permit aggregation of QALY changes, the value scale should have interval scale properties such that, for example, a gain from 0.2 to 0.4 is equally valuable as a gain from 0.6 to 0.8. States worse than dead can exist and they would have a negative value and subtract from the number of QALYs. These conditions, along with an assumption of risk neutrality over life-years, are sufficient to ensure that the QALY is a useful representation of health state preferences.*

Response: The authors are confused. While the valuation scale 'must' have a 0 – 1 range they do not consider how QALYS are created or demonstrate that the scale 'must have' interval scale properties (in fact they should be referencing ratio scales). The scale needs a true zero (which utility scales don't have) in order to create and evaluate QALY change. As the QALY is an impossible mathematical construct the argument for QALY required characteristics collapses. If the authors understood measurement theory it is not the invariance of comparison of an interval scale that is the only critical issue but the presence of a ratio scale that lacks construct validity and lacks both invariance of comparisons, and a true zero. There is no suggestion in the paper that in creating utilities for QALYs you need to recognize the importance of actually creating a ratio scale when the instrument is being developed. Creating such a scale for latent constructs such as need based QoL is impossible; we rely on interval scales to assess response. They also fail to appreciate that multiattribute scales such as the EQ-5D-3L are not dimensionally homogeneous in the symptoms captured and hence cannot be used to create a single score (they lack construct validity).

- ICER agrees that the EQ-5D itself is not a ratio scale. However they disagree that a ratio scale is necessary for estimation of utility for use in producing QALY estimates. They argue that ratio scales are only necessary when needing to multiply or divide values along the continuum of the scale (*not for multiplying some other quantity such as time?*). The requirements for calculating a QALY in their model apparently required only that the scale produced an equal magnitude for each point on the scale. Therefore only an interval scale was needed.

Response: Apart from the fact that the EQ-5D-3L does not have demonstrated interval properties (because it was never designed to have them) I don't see how you create QALYs without multiplication? You have a ranking of raw scores, distance unknown, and you multiply time spent by an ordinal raw score (which is mathematically impossible). I fail to see how your model overcomes this. How do you apply an equal magnitude of difference (which the EQ-5D-3L does not have) to create a QALY? Does this mean that a QALY is not created by multiplying an EQ-5D-3L raw score by time but by some other magnitude? Yet you present estimates of simulated QALYs? How can you multiply time spent in a disease state by an equal magnitude of difference in ordinal utilities (?) when the utilities in question do not have interval properties? Again, you point out that the EQ-5D-3L is 'considered' as meeting criteria for creating QALYs yet you are unable even to demonstrate it has interval properties. Why is everything 'considered' but never proved? Is this just belief by unprovable assumption? Is belief strongest when the object of that belief is patently impossible? You raise some, possibly interesting, epistemological questions on the distinction between justified belief and opinion.

- Apparently, there is a belief that the QALY may satisfy ratio properties. The reference is to Roudjik et al [20].

  *Response: This was raised in the hemophilia A evidence response. "Ratio properties are not necessary for estimation of utility for use in producing QALY estimates". You confuse the question: a utility score either has or has not ratio properties (*we know it does not). We can, as you clearly do, suspend belief in fundamental measurement and enter a fantasy parallel measurement universe where ordinal scales are ratio/interval scales in disguise and interval scales have ratio properties. I have already commented on the reference provided [26] [27]. The claims made are nonsensical where a true zero for the EQ-5D-3L is just an assumption. To assert that the EQ-5D-3L scale is a ratio scale in disguise is a far cry from actually proving it. Belief is not proof.*

Abandoning an Alternative Reality

If we are prepared to abandon the I-QALY approximate information meme, an unlikely event (at least in the near future), as well as putting aside the belief in composite, multiattribute dimensionally heterogeneous raw scores, then the issue is one of the requirements for a new value assessment paradigm. Fortunately, we have a paradigm for health technology assessment that meets the standards of normal science without being encumbered by the I-QALY. The elements have been detailed in the recently released version 3.0 of the Minnesota formulary guidelines [28]. The answer lies in rejecting impossible ordinal scales from multiattribute instruments, focusing instead on the construction of disease specific single attribute claims. These claims must be credible, empirically evaluable and replicable. Claims should specify a particular value attribute, whether clinical, quality of life or as elements of resource allocation. Claims will have either demonstrated interval or ratio properties. Claims that refer to latent constructs such as needs fulfillment QoL will have interval properties. All claims should be accompanied by an assessment protocol detailing the real world evidence base for claims assessment and the timelines for reporting to a formulary committee. We don't need assumption driven imaginary lifetime claims that are mathematically impossible.

Contributions to this new formulary submission framework have been pursued for the last 20 years. In terms of a latent construct or attribute such as needs fulfillment QoL a recent paper details how, with RMT as a guide, a valid interval scaled instrument single attribute instrument, with excellent psychometric properties, can be developed following application of Rasch measurement theory; the development of the Alzheimer's Patient Partners Life Impact Questionnaire (APPLIQue)[29] [30]. The theoretical basis for this measure is the needs-based QoL model. The basic premise is that if the impact of a disease or condition on a patient (or caregiver) is to be assessed then this must include both clinical and non-clinical influences; life gains its quality from the ability of individuals to meet their basic needs [31]. The APPLIQue captures a unidimensional construct: needs-based QoL. It is assumed that QoL will be higher when most needs are fulfilled; lower when less are fulfilled. Designed to have an interval measurement property, the APPLIQue can measure response to therapy and employ a range of statistical techniques to evaluate change over time and compare competing therapies. This has been true of the many RMT disease specific instruments developed over the past 20 years. In all cases care was taken to ensure that they generated interval scores and that the steps in instrument development were documented in peer reviewed publications. With such transparency it is no longer necessary to say that there is an understanding, deeply held yet false, belief or faith that an instrument such as EQ-5D-3L has ratio properties without any evidence for these assertions in instrument development.

Conclusions

Whether or not the responses from ICER are necessarily indicative of those that might be received from other analysis groups, it is disquieting that there seems to be no awareness of the difference between the data-to-model paradigm and the model-to-data paradigm as described by Bond and Cox. The key point is that there is no perception, common in the physical sciences since the 17th century, that if an instrument is to have required measurement properties then it has to be designed to have those properties [32] . We cannot assume that *ex post facto* these properties are mysteriously present, demonstrated by placing raw or ordinal scores on a number line with equal scale units and assuming they had ratio properties. Even the presence of negative utilities failed as a red flag to question the implications of the absence of a true zero. Indeed the hallmark of this approximate information I-QALY parallel measurement universe is that everything is considered 'to be', never proved. They are just one more in a series of assumptions that support non-evaluable claims. Perhaps they should be honest and simply state that the required measurement properties are present only by assumption..

Possibly more disconcerting is the willingness to 'understand' that a measure has certain properties without bothering to challenge this assumption. Even more disconcerting is the confusion that defenders of the I-QALY exhibit in their attempt(s) to defend the impossibility of an I-QALY construct. In the response to issues raised in the bladder cancer ICER evidence report, ICER appears to believe: (i) in an ordinal scale that is actually an interval scale; (ii) in an assumed interval scale (the EQ-5D-3L) that has the ratio properties of multiplication and division; (iii) that the EQ-5D-3L ordinal scale is actually a ratio scale in disguise; (iv) that ratio properties are not necessary for estimation of utility for use in producing QALY estimates; (v) that to create QALYs all you need is an interval scale without a true zero; (vi) that the EQ-5D-3L

needs only to have interval properties to produce QALYs without any consideration of ratio scales; and (vii) it is acceptable to construct ordinal dimensionally heterogeneous multiattribute utility scores that lack construct validity and interval properties yet consider them ratio scales.. This is an amazingly complex and contradictory belief system; but perhaps the I-QALY approximate information missionaries from ISPOR will resolve these issues. Or, on a more positive note, perhaps ICER could propose a new standard for the axioms of fundamental measurement. If so, it would be the most significant contribution to measurement theory for the past 80 years.

Fortunately, we have a paradigm for health technology assessment that meets the standards of normal science without being encumbered by the I-QALY. The elements have been detailed, as noted, in the recently released Minnesota formulary guidelines. The answer lies in rejecting ordinal scales from multiattribute instruments and all measures that fail to meet the standards of fundamental measurement. Indeed, a rejection of the notion that it is possible to subsume blanket claims for cost-effectiveness within a single score. This cull would be extensive as the construction and belief in ordinal measures or raw scores is widespread. We need to focus on the construction of disease specific single attribute RMT standard scales, notably where we are dealing with latent constructs such as QoL. Claims must be credible, empirically evaluable and replicable. Claims should specify single attributes relevant to value assessment in that population, whether clinical, quality of life or as elements of resource allocation. Individual attribute claims in the clinical area would include a range of ratio scales and interval scales. Where PRO latent construct claims are proposed then we must assess their fit to RMT and their interval basis for assessing response to therapy. For resource utilization impacts we return to ratio scales. All claims should be accompanied by an assessment protocol detailing the real world evidence base for claims assessment and the timelines for reporting to a formulary committee. We don't need assumption driven imaginary claims that are mathematically impossible; nor do we need lifetime imaginary simulations built on discredited multiattribute measures.

:

REFERENCES

[1] 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

[2] Merbitz C, Morris J, Grip JC. Ordinal scales and foundations of misinference. *Arch Pjys Med Rehabil*. 1989. 70:308-312

[3] Grimby G, Tennant A, Testo L. The use of raw scores from ordinal scales: Time to end malpractice (Editorial) *J Rehab Med*. 2012;l44:97-8

[4] Greene B. Until the end of time. New York: Alfred A. Knopf, 2020

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

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

[7] Langley PC. Sunlit uplands: the genius of the NICE reference case. *Inov Pharm*. 2016;7(2): No.12.

[8] Langley PC. Dreamtime: Version 5.0 of the Australian Guidelines for Preparing Submissions to the Pharmaceutical Benefits Advisory Committee (PBAC). *Inov Pharm.* 2017;8(1): No. 5

[9] 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

[10] Brazier J, Ara R, Azzabi I, et al. Identification, review, and use of health state utilities in cost-effectiveness models: an ISPOR Good Practices for Outcomes Research Task Force Report. *Value Health*. 2019;22(3):267–275.

[11] Dawkins R. The Devil's Chaplain. New York: Houghton Miflin, 2003

[12] Langley PC, McKenna SP. Measurement, modeling and QALYs [version 1; peer reviewed] *F1000Research* 2020, 9:1048 https://doi.org/10.12688/f1000research.25039.1

[13] Stevens S. On the theory of scales of measurement. *Science*. 1946;103:677-680

[14] Bond T, Fox C. Applying the Rasch Model . New York: Routledge, 2015

[15] McKenna S, Heaney A, Wilburn J et al. Measurement of pAtient-reported outcomes. !:The search for the holy grail. *J Med Econ*. 2019;22(6): 516-22

[16] McKenna SP, Heaney A, Wilburn J. Measurement of patient reported outcomes. 2: Are current measures failing us? *J Med Econ*. 2019;22(6):523-30

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

[18] Rind DM, Walton SM, Agboola F, Herron-Smith S, Quach D, Chapman R, Pearson SD, Bradt P. Valoctocogene Roxaparvovec and Emicizumab for Hemophilia A: Effectiveness and Value; Draft Evidence Report. Institute for Clinical and Economic Review, August 26, 2020. https://icer-review.org/material/hemophilia-a-update-draft-evidence-report/

[19] Institute for Clinical and Economic Review. Valoctocogene Roxaparvovec and Emicizumab for Hemophilia A without Inhibitors: Effectiveness and Value Response to Public Comments on Draft Evidence Report October 16, 2020 https://icer-review.org/wp-content/uploads/2019/12/ICER_Hemophilia-A_Public-Comment-Responses_101620.pdf

[20] Atlas SJ, Touchette DR, Beinfeld M, McKenna A, Joshi M, Chapman R, Pearson SD, Rind DM. Nadofaragene Firadenovec and Oportuzumab Monatox for BCG-Unresponsive, Non-Muscle Invasive Bladder Cancer: Effectiveness and Value; Draft Evidence Report. Institute for Clinical and Economic Review, September 17, 2020. https://icer-review.org/material/bladder-cancer-draft-evidence-report/

[21] Institute for Clinical and Economic Review. Nadofaragene Firadenovec and Oportuzumab Monatox for BCG-Unresponsive, Non-Muscle Invasive Bladder Cancer: Effectiveness and Value Response to Public Comments on Draft Evidence Report November 6, 2020 https://icer-review.org/wp-content/uploads/2020/02/Bladder-Cancer_Public-Comment-Response_110620.pdf

[22]Ollendorf DA, Bloudek L, Carlson J, Pandey R, Fazioli K, Chapman R, Bradt P, Pearson SD. Targeted Immune Modulators for Ulcerative Colitis: Effectiveness and Value; Draft Evidence Report. Institute for Clinical and Economic Review, May 26, 2020. https://icerreview.org/topic/ulcerative-colitis/.

[23] 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). *InovPharm*.2020;11(2): No 17
https://pubs.lib.umn.edu/index.php/innovations/article/view/3330/2533

[24]Roudijk B, Donders R, Stalmeier P. Setting dead at zero: Applying scale properties to the QALY model. *Med Decis Making*. 2018;38(6): 627-34

[25]Weinstein M, Torrance G, McGuire A. QALYs: The basics. *Value Health*. 2009;S5-S9

[26]Roudijk B, Donders R, Stalmeier P. Setting dead at zero: Applying scale properties to the QALY model. *Med Decis Making*. 2018;38(6): 627-34

[27] Lugnér AK, Krabbe P. An overview of the time trade-off method: concept, foundation, and the evaluation of distorting factors in putting a value on health. *Exp Rev Pharmacoeconomics Outcomes Res*. 2020; 29(4):331-342

[28] Langley P. Value Assessment, Real World Evidence and Fundamental Measurement: Version 3.0 of the Minnesota Formulary Submission Guidelines. *Inov Pharm*. 2020;11(4): No 12
https://pubs.lib.umn.edu/index.php/innovations/article/view/3542/2613

[29] Hagell P, Rouse M, McKenna S. Measuring the impact of caring for a spouse with Alzheimer's disease: Validation of the Alzheimer's Patient Partners Life Impact Questionnaire (APPLIQue). *J Applied Measurement*. 2018;19(3):271-82

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

[31] McKenna, SP, Doward, LC. The needs-based approach to quality of life assessment. *Value Health*, 2003; ,7 Suppl 1, S1-3.

[32] Chang H. Inventing Temperature: Measurement and Scientific Progress. New York: Oxford University Press, 2004