MAIMON WORKING PAPERS

Working Paper No. 9 April 2020

TO DREAM THE IMPOSSIBLE DREAM: THE EQ-5D QALY

Paul C Langley, PhD
Adjunct Professor, College of Pharmacy, University of Minnesota

Abstract
One of the unusual features of health technology assessment that has been endorsed over the past 20 to 30 years by professional groups such as the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) and health departments in single payer systems such as the National Health Service in England through the National Institute for Health and Care Excellence (NICE), has been the wholehearted commitment, not just to the creation of imaginary worlds to support health technology assessment recommendations for pricing and access, but the uncritical embrace of the EQ-5D-3L index score for the creation of quality adjusted life year (QALY) claims. Attention has not been given by these professional groups and leaders in the field of technology assessment to the question of fundamental measurement; does the EQ-5D-3L meet the standards for fundamental measurement? We have known for 20 years or more that it does not meet those standards, it is an ordinal manifest score. More to the point is the willingness, without any attempt to evaluate the characteristics of the index, to apply the EQ-5D-3L to modeling diverse disease states. Studies producing EQ-5D-3L scores have taken these at face value to support modeling assumptions. Previous commentaries and reports have made the case that the EQ-5D-3L, from the perspectives of the practice and axioms of fundamental measurement theory have failed to recognize (or have ignored) the fact that the EQ-5D-3L fails standards for interval and ratio scales. This applies to ICER evidence reports and the reference case modeled claims; a complete absence, by the respective contracted academic modeling groups at universities, to include the University of Minnesota, University of Calgary, Colorado University, University of Illinois and the University of Washington, to appreciate the role of measurement theory in instrument development. This is unacceptable. The purpose of this commentary is to build on previous assessments of the pseudoscientific status of ICER evidence reports, examining in detail the properties of the EQ-5D-3L and the impossibility of creating QALYs utilizing its so- called 'utility' index.

Keywords: EQ-5D-3L, profiles, index scores, measurement theory, nonsense QALYs

Introduction

Pseudoscience: a collection of beliefs or practices mistakenly regarded as being based on scientific method (Oxford Dictionaries).
Excellence: the quality of being outstanding or extremely good.

Commentaries by the present author over the past few years, as detailed on the author's website (www.maimonresearch.net), have pointed to the lack of scientific merit in the construction of imaginary, modeled-by-assumption, reference case worlds to support health technology assessment. Outside of professional groups such as the International Society for Pharmacoeconomics and Outcomes Research (ISPOR), the Institute for Clinical and Economic Review (ICER) has attempted to insinuate itself as the principal arbiter for imaginary value assessments in the US. The ICER business model is built around the construction of cost-per-QALY lifetime imaginary simulations which claim to provide a framework relevant to health system decision makers for pricing and access for pharmaceutical products and devices. As detailed in a number of commentaries the ICER modeling approach, its value assessment framework, fails to meet the standards of normal science; the discovery of new facts [1]. It is best characterized as pseudoscience (i.e., bunk) [2]. Constructing imaginary worlds to support pricing and access recommendations has certainly characterized health technology assessment of the past 30 plus years. Indeed, ISPOR makes clear that it is not interested in hypothesis testing or the discovery of new facts in treatment impact [3]. ISPOR with its mission to promote scientific excellence in health economics and outcomes research, sees its principal role in generating 'approximate information'; imaginary world evidence, created by its focus on lifetime incremental cost-per-quality adjusted life year (QALY) estimates and willingness to pay thresholds. Unfortunately, the QALY is a logically impossible creation. This absurd focus on imaginary evidence is in contrast to real world evidence where meaningful claims, the discovery of new facts, for therapy impact and quality of life in disease areas can be evaluated from evidence platforms.

At the same time, the erstwhile contracted ICER model builders at academic institutions, including the University of Minnesota, University of Calgary, Colorado University, University of Illinois and the University of Washington, there seems a lack of interest in confirming the assumed measurement properties of a critical element in construction imaginary worlds: EQ-5D utilities. These 'scores' are applied to the construction of QALYs and subsequent claims for incremental cost-per-QALY over a hypothetical lifetime for a hypothetical target population; this is the foundation for ICER's much publicized recommendations for pricing and access. A review of ICER evidence reports gives no sign that the authors of the various models (nor the leaders of ICER) have considered the question of EQ-5D measurement properties (or, more to the point, their absence). Rather, they seem content to extract EQ-5D-3L measures from a handful of published studies in the particular disease state, adding where deemed necessary scores from other generic instruments and, as a last resort, an educated guess.

This approach is unacceptable. The cavalier attitude in the social sciences towards measurement standards has been commented on for at least 30 years, notably in respect of the misuse of ordinal scales [4] [5]. Measurement is fundamental to the physical sciences, whether direct or indirect. Science cannot progress until the measurement issue and the construction of instrumentation with instrumentation standards has been resolved. Thermometry is a classic example. Temperature is measured indirectly with a range of thermometer types with limited scale range developed for specific applications. As an analogy for the social sciences, thermometry points to the need for a variety of single attribute interval measures. Unfortunately, *most outcome measures used in health care are ordinal in nature, precluding such arithmetic operations* (as addition and subtraction). *Many such measures focus on attributes that are not directly measurable, such as pain, self-esteem and quality of life. These measures give a manifest score of the construct being measured* [6].

The purpose of this commentary is to make the case that, from the perspective of normal science with its commitment to ensuring the application of the axioms of fundamental measurement in instrument development, the continued belief in the generic utility QALY is misplaced. It represents a will o'the wisp that is being slavishly followed by those developing incremental cost-per-QALY models. This is not new; the failure to appreciate fundamental measurement has characterized health technology assessment for over 30 years. Indeed, ISPOR is apparently proud to point out that thousands of imaginary cost-per-QALY studies have been carried out and catalogued (e.g., the Tufts Utility Emporium), apparently without recognizing that they are based on flawed measurement assumptions and are essentially worthless [7].

Fundamental Measurement

Fundamental: affecting or relating to the essential nature of something or the crucial point about an issue.

Following the seminal contribution of Stevens, some 70 years ago, we define measurement as the assignment of numbers to objects or events according to a rule which determines the measurement scale [8]. Four measurement scales are recognized: nominal, ordinal, interval and ratio. Each of these scales satisfies one or more of the following properties:

- Identity: each value on a measurement scale has a unique meaning
- Magnitude: each value on a measurement scale has an ordered relationship t where some are larger or smaller than each other
- Invariance of comparisons: scale units along a number line are equal to each other, there is an invariance of comparisons where the a given difference means the same at all levels of the variable
- True Zero: the scale has a true zero point below which no values exist

With the properties for each scale:

- Nominal: only satisfies the identity property; a descriptive category with no numerical value with respect to magnitude
- Ordinal: satisfies the properties of identity and magnitude where each value on the scale has a unique meaning and an ordered relationship to every other value on the scale but lacks any concept of the distance between the values
- Interval: satisfies the properties of identity, magnitude and invariance of comparisons such that the distance between values is known
- Ratio: the ratio scale satisfies all four properties such that the distance from a minimum true zero is known

The properties of the scales are critical for the type of arithmetical operation they can support. For the ordinal scale the only 'measures' that can be supported are modal or median counts for the ordered values; the scale is non-linear. Like an elastic band, the 'ordinal interval' scale can support any number of distributions, depending on the placement on ordinal scale of the linear 'markers'. The interval scale can only support addition or subtraction. The absence of a true zero prohibits multiplication or division as we have no idea of the distance to zero. The ratio scale supports all four properties. It is the only scale that allows all four arithmetic operations. Descriptive statistical measures such as means and standard deviations are supported by the interval scale as these require only addition. These are invalid with ordinal scales, where many are bound to an artificial range (e.g., 0 – 1 for EQ-5D measures). This scale not only lacks equal interval units but being bound causes distortions of intervals towards the margins, independently of possible floor or ceiling effects, where the floor may take on negative values. When ordinal scales are manipulated mathematically (e.g., creating QALYs) the results are logically invalid [9]. The values on an ordinal scale are just manifest raw scores. Importantly, Rasch modeling of ordinal data, if meeting the required standards, can transform ordinal to interval scales [10]. Even then, the Rasch scale only allows claims to be made for differences in response: in terms of measurement axioms it supports invariance of comparisons and sufficiency (i.e., if no other statistic from a sample yields more information as to the value of a parameter; in Rasch measurement the number correct is the sufficient statistic for estimating item difficulty and person ability). The latent estimate derived is an interval scale.

Rasch modeling to translate ordinal to interval scales does not extend to the creation of ratio scales. Indeed, the history of measurement in the physical science points to the difficulty of creating ratio scales with a true zero. In thermometry, for example, we have interval scale thermometers adapted to different measurement tasks, where the focus is on response (addition and subtraction) rather than ratio scale requirements linked to a measure of absolute zero. There seems little point in arguing for the ability to multiply or divide degrees Celsius on a scale of 0 to 100; change is the critical requirement. This applies to Rasch models. If the latent, unidimensional construct is needs based quality of life, then the focus is on response to therapy interventions; have the needs of patients been met? To go beyond this to the mysteries of lifetime QALYs seems somewhat ridiculous.

Instrument Development
Instrument: a measuring device for determining the present value of a quantity under observation

Meeting fundamental measurement standards is essential in the physical sciences; an instrument has to be designed to meet these standards. Fundamental measurement is possible when units can be physically concatenated (e.g., weight). These additive measures are, however, in the minority. We have to discover the characteristics of other physical measures indirectly (e.g., relative density of a substance where density = mass/volume). Where we are dealing with intangible or non-physical attributes in the social sciences (e.g., needs based quality of life) then we need a technique for detecting measurement structures. The solution has been simultaneous conjoint measurement (SCM) developed in the early 1960s by Luce and Tukey, and by Rasch [11] [12]. Under SCM, the crucial indicator of an additive measurement structure in the data, interval levels scales, are the observable relationships between and among item matrices. These item matrices are defined by two observable attributes: the difficulty of a questionnaire item and the ability of respondents to affirm the item. This yields a matrix of expected response where the probability of success depends on the ability of the person and the difficulty of the item.

The Rasch principal is that interval-level measurement can be derived from ordinal responses when the levels of one attribute (e.g., probability of correct response) increase with the values of the other two attributes: item difficulty and person ability. The key point is that the Rasch instrument, such as an interval scale for quality of life response, is created from the 'ground up'. Items for the instrument are selected to meet or fit Rasch model requirements. The interval level property can therefore be confirmed. This is contrast to instruments typically developed in the social sciences where the instrument fits the data. This is the case with the EQ-5D-3L and EQ-5D-5L where application of an econometric model to sample responses defining health states yields, after a little 'tweaking', a scale calibrated from unity to some negative value depending on the algorithms preference weights or community values (EQ-5D-3L = -0.59). There was no attempt to consider SCM standards in instrument development. But technology assessment, the ISPOR paradigm, puts these minor concerns with measurement theory to one side; it may not be a conscious decision to support imaginary constructs and ignore fundamental measurement. It is more likely a tradition in the social sciences to base statistical analyses on a belief that we can put to one side the 'blatant ignorance that ordinal data do not constitute measurement' [13]. Is it possible that we can overthrow this paradigm of ignorance, of methodological thought disorder?

Examining the EQ-5D Entrails
Entrails: the innermost parts of something (Oxford Dictionaries)

As an example of the closely held belief in the ratio properties of the EQ-5D-3L and EQ-5D-5L, it is of interest to consider two studies which, ignoring any hint of measurement properties, have examined the distribution of EQ-5D-3L manifest scores or index values: the Parkin et al study of the distributional characteristics of the EQ-05D-3L and the Feng et al study of the distributional characteristics of the EQ-5D-5L [14] [15]. These are issues which are seldom addressed by studies which have simply reported on aggregate EQ-5D scores for target patient populations in disease states. The result is that ISPOR, ICER and other analysis groups have taken these scores at face value with no thought, not only to fundamental measurement, but to the odd and flexible distributional characteristics which make the application of summary manifest scores even more bizarre. This closer attention would point to the nonsense of ordinal utility scores in QALY creation.

The EQ-5D-3l instrument with its five symptoms or health dimensions (mobility, self-care, usual activities, pain and discomfort and anxiety and depression) and 3 response levels for each (no problems, some problems and extreme problems) yields 243 possible health states (=3^5) plus. These are anchored to a scale 1 = perfect health to 0 = death and values less than death having a value less than zero. In terms of a number line with variable distances between manifest scores (a variance of comparisons) we have a scale ranging from 1.0 to -0.59. Each of the 15 health responses are weighted with those for perfect health having a value of zero and the others representing community preference weights generate by sample survey. To calculate a value for each health state the preference weights are summed and subtracted from unity. Where the response is for all 'no problem' responses the value is unity. For all other health states a constant is added to be subtracted. If any health state involves a 'severe problem' response then a further constant amount is added. The result is a single manifest score for each health state on an imaginary number line with the range from unity to –O.59. The number line minimum (negative) value is, of course, only 'fixed' in terms of an agreed set of preference weights for the 15 response levels. If you calculate a utility score from ordinal responses by attaching weights then you end up with a multiattribute ordinal scale.

The key feature of the distribution of utility scores for the EQ-5D s that there are no distributional characteristics. Distributions cannot be created from ordinal rankings. The reason is obvious: we have no idea of the distance between the EQ-5D utilities. The intervals are unknown. The so-called scores are ranked on an imaginary non-number line. Parkin et al and Feng et al are apparently blissfully unaware of this characteristic of ordinal scores. In this state of measurement bliss they treat these 'non-numbers' as if they had ratio properties. They provide no justification for this assumption; they are presumably wedded to the technology assessment QALY meme. This leads to a ridiculous discussion on the apparent clustering of EQ-5D-3L scores into two distributions: a low cluster of responses and a high number of responses. These are entirely imaginary constructs. Clustering by assumption, the existence of ordinal non-numbers that have interval properties, says nothing about the underlying latent construct. It is a will o'the wisp.

As a matter of interest, assuming the instrument's number line has interval properties, is a clustering of utility vales into two distributions: a low score distribution and a high score distribution. The conclusion by Parkin et al is that this bi-modal clustering is due to the EQ-5D-3L classification system which generates differences between patients with the same condition in respect of dimensions that are mainly observed at level 2 (some problems) and level 3 (extreme problems). The weights assigned to extreme problems exacerbate this by placing a larger weight on extreme problems (and hence lower vales). To add to this distributional feature of responses to EQ-5D-3L only a few of the 243 profiles are noted with any frequency. In one data set 22 profiles covered 90% of all health states, with 161 not observed at all.

Apart from the clusters reflecting only a small proportion of possible health states, a further problem is whether or not a single index or average of the individual respondent indexes makes any sense. These distributions are flexible; there is no requirement for invariance of comparisons. We can generate any distributional properties we want by simply changing the artificial interval scale intervals; these might be 'tight' at each end of the scale, yet 'wide' in the middle. Choose your scale and choose your distribution.

Mean utility 'values' are regularly reported in the few studies that ICER relies on to create QALYs; typically no mention is made of the underlying distribution of responses and whether they are unimodal or bi-modal, elastic or otherwise. If, as assumed by the Parkin et al analysis the scale has invariance properties (including invariant negative utilities) then under this assumption they are likely to be bi-modal then we face the issue that there is no agreed summary statistic (or statistics) to quantify the parameters of a general bimodal distribution. Consider the distributions presented by Parkin et al for asthma: Low cluster mean = 0.086 (SD 0.159); range -0.484 to 0.383 while the high cluster mean = 0.722 (SD 0.159); range = 0.414 to 0.883. Given we have essentially separate distributions, presenting a 'utility' as a weighted average of the two means does not make much sense? What is the interpretation to place upon it? Do we have two manifest score distributions to describe the EQ-5D-3L health related quality of life for the asthma population? How would we interpret QALYs for the asthma group? Would we have to assess the actual distribution and possible clustering of individual EQ-5D-3L manifest index scores before assuming a single 'mean' index score from an underlying single distribution? Finally it is worth noting that for all the disease groups, as might be expected, there are negative index scores. Remember also that as these are ordinal manifest scores, the summary distributional characteristics described above for asthma are an artifact; these types of statistical operations are precluded by the ordinal nature of the data. Specifying a range and then noting that the range of low and high cluster scores are 'separate' adds little to our understanding of the data as we have no idea, in the absence of an interval scale, of the 'distance' between any two scores. Parkin et al have just assumed that the number line for manifest scores has interval properties; we can only rank the scores. Estimating distributional characteristics and claiming well defined clusters is simply nonsense.

The Feng et al manifest score distribution assessment for the EQ-5D-5L presents a somewhat different picture in an analysis of 3 patient groups. The five symptom dimensions remain but with five response levels (no, slight, moderate, severe or extreme problems) yielding 3,125 possible health profiles (= 5^5). The analysis involves application of two value sets: the Devlin et al English Value Set (EVS) and the van Hout value set (MVS) [16 17]. Both are used to compare to the EQ-5D-3L index values. The results, although suggestive, are seen as preliminary with a caution from the authors about application of the EQ-5D-5L index in creating QALYs; the assessment assumes (as with Parkin et al) that the index scores, rather than being ordinal manifest scores, can support standard arithmetical operations. In any event, the application of the EVS provides no 'strong or consistent evidence' for clustering while there is clearer evidence of clustering using the MVS. It goes without saying that there are negative scores for both the EVS (-0.285) and the MVS (-0.594). This analysis again, for ordinal data, is nonsensical. Again, we can only rank scores; further statistical operations have no interpretable meaning.

If there are, given this fantasy construct, truly bimodal clusters in the health related quality of life of patients by stage of disease then these should be recognized in modeling. Can we define, believing in the ratio properties of ordinal scales, that as part of a trial or observational study protocol, target patients that inhabit disease state should be defined by the symptom and response levels of the EQ-5D-3L instrument? Should we insist that the EQ-5D-3L be a required element in a phase 3 protocol if we intend to utilize the EQ-5D-3L index score to support an incremental lifetime cost per QALY claim? Or are we chasing a will o'the wisp of constructed idiocy? If any EQ-5D utility will do to construct the QALY then we have a smorgasbord of options. Choose your manifest score, assume it has ratio properties (ignoring the negative vales) and convince those with little understanding of fundamental measurement that this imaginary world is the promised land of ICER cost-per-QALY pronouncements.

Negative QALYs and Negative Time

Negative: consisting in or characterized by the absence rather than the presence of distinguishing features (Oxford Dictionaries).

If we ignore the axioms of fundamental measurement and continue to assume, along with the leaders of ISPOR, that the EQ-5D instruments are, in effect, ratio scales (ignoring the obvious fact that there is no true zero) then we have the intriguing fact that as people can have negative utilities then, by extension, they can have negative QALYs. Indeed, if we consider how model builders add and subtract QALYs for events associated with disease states, then patients can hop in and out of negative QALY states over the lifetime course of their disease. While a negative utility is considered a state worse than death (and subjects could even report they are dead with a utility of zero and then 'recover' or resurrect) it is not clear where that 'state' is spent? Possibly some crypt in a Transylvanian castle with a retainer called Igor? Where the value weights are community determined, as with the EQ-5D instruments, it is presumably the community that can inform the patient that in their opinion they are in a state worse than death. The policy implications for, say, continued use of health care resources is unclear. Are they intended to bring patients back from this negative state to a positive utility or QALY state? Do health care administrators advise patients that, from a QALY perspective, they would be better off dead? Presumably offering euthanasia as a covered service by insurers? A possible ICER recommendation? A true zero, where by construction (if possible) there are no negative values would solve this conundrum. This is impossible. Or we could assume that negative values are another way of defining an absorbing state?

One of the more intriguing aspects of negative QALYs is the implication of a new concept in fundamental physics: negative time. Perhaps inadvertently, this possibility has been overlooked by pharmacoeconomists in their modeling of imaginary worlds. We now have the prospects of a true science fiction fantasy: a world with negative time - or at least a world where members of a target patient group can experience negative time over the course of their hypothetical disease state.

The QALY has an intrinsic appeal; to characterize time spent in a disease state adjusted by a preference value for health status (deviation from perfect health) has an aura of authenticity. In the case of the EQ-5D measures, the preferences are not those of the patient but of a community sample, most of whom have no experience, either directly, or through family members of the disease state. The EQ-5D measures are multiattribute, rather than focusing on one health attribute. Certainly, we can appreciate the complexity of health experience, but this should be captured one attribute at a time from the patient perspective or by an instrument (e.g., pain) that meets unidimensional and interval measurement standards. The complexity of patient response or status cannot be expressed by one score. Attempting to define the multitude of different disease states experience by patients (and caregivers) in terms of five symptoms and 3 (or 5) response levels seems absurd. Combining attributes, which may be latent constructs in their own right, degrade the 'precision' of the measure with predictions more hazardous.

Non-Numbers

Number: Numbers are strings of digits used to indicate magnitude. They measure size - how big or small a quantity is. In mathematics there are several types of numbers, but they fall into two main classes, the counting numbers, and scalars.

A part of the confusion, if not a large part, with the acceptance of the EQ-5D-3L and EQ-05D-5L as ordinal rankings, is to treat the manifest scores as numbers; they are not numbers; they are 'scores' that could equally well be designated with alpha notations (e.g., 11233 or AABCC). The latter avoids attributing the manipulation of numbers with interval or ratio properties to the EQ-5D-3L and EQ-5D-05L manifest scores to categorical data (e.g., can be conceptualize a mean value from a distribution of non-numbers defined as alpha set). Certainly, one property of the EQ-05D scoring is that we can rank the 'scores'; but we still have no idea of the difference between them. Different preference weights will create different EQ-5D scores and may even change the overall ranking of these non-numbers.

A Manifest Dream

Manifest: a document giving comprehensive details of a ship and its cargo and other contents, passengers, and crew for the use of customs officers.
Dream: a state of mind marked by abstraction or release from reality

Without wishing to be unreasonably curmudgeonly, it does seem to be somewhat of a waste of time to devote resources to this assessment of EQ-5D entrails. After all, the EQ-5D instruments were never designed to meet fundamental measurement standards. In Rasch measurement terms, there was no intent to create an interval index of response; let alone a ratio scale with a true zero. What appears to have been overlooked, if the holy grail is a utility index to create QALYs, then that is a hopeless quest, an impossible dream [18 19].

Continuing our odyssey into the assumed ratio world of the EQ-5D, even with eventide falling, we must address the abiding mystery of the QALY. To say that the QALY is central to the memetic paradigm of health technology assessment is like affirming belief in original sin. Remove the QALY, remove original sin, and the belief system collapses. Memes can endure for centuries as a unit of cultural transmission (e.g., Rosicrucians). Organizations attempt to ensure memetic inter-generational copying fidelity, with the odd misstep as in the case of Luther and the *Schlosskirche* doors; children seldom challenge their parent's belief systems. ISPOR appears no different. Newly minted Ph.D's emerge, wedded to a Jesuitical belief in the construction of reference case imaginary worlds for approximate information, with no thought or exposure to the axioms of fundamental measurement. If the lack of response to attempts to introduce Rasch modeling over the past 20 years is any guide, this is more than likely to continue [20]. The most widely used textbook in the field makes no reference to Rasch standards. It perseveres with the belief that generic utility measures have interval properties but with no understanding that the support for EQ-5D and similar instruments points to an assumed belief in ratio properties [21].

Playtime Disclaimer

Disclaimer: a statement that denies something, especially responsibility (Oxford Dictionaries).

It is not the intention here to discourage examinations of the profiles, values and distributions of index scores from either the EQ-5D-3L or the EQ-5D-3L. Certainly, analysts can put to one side consideration of limits imposed by the failure of these two instruments to meet the standards of normal science. A meme, as noted in previous commentaries, can be tenaciously held with its followers ensuring generational copying fidelity to support continuing mysteries. After all, on a wet Sunday afternoon, evaluating distributions of multiattribute manifest index scores by target patient groups can pass the time; as a welcome respite to the 2000 piece jigsaw puzzle (with no applicable measurement properties). This should not discourage the creation of QALY reference case models; these have been a memetic centerpiece to technology assessment over the past 30 years and, as an exercise in imaginary constructs, will continue to have a therapeutic appeal. Finally, it should not discourage manufacturers in their support, presumably for marketing purposes, for the creation of imaginary worlds as an exercise. After all, in a classic example of the blind leading the blind, those supporting such models in marketing and public affairs departments will have no idea of the details of the model. Their aim is to present, hopefully favorable comparative claims (if they are not suppressed) for pricing and access to health system decision makers, who are equally ill equipped to understand how the recommendations were created.

Conclusions: Converting the Illuminati?

Illuminati: people claiming to possess special enlightenment or knowledge of something (e.g. an elite comprising the Bavarian illuminati).

The purpose of this commentary has been to ask decision makers to abandon imaginary constructs; to put imaginary information, the mainstay of the ICER business model, to one side in favor of a program to support formulary claims driven by real word evidence.

Presumably, by definition, the illuminati are not to be converted. They possess a secret knowledge, passed down by the technology assessment meme from generation to generation of model builders. Their Boston chapter, ICER, is supported by manufacturers and other interest groups. The ICER business model is the imaginary world. They hold tenaciously to QALYs as the only product. Absent QALYs, or at least a belief in the mystery of the ratio manifest utility, the business model collapses. As leaders in technology assessment, attested to by ISPOR in their recent releaser of guidelines for imaginary worlds, the illuminati hold as an article of faith (a deeper mystery) that their role as *leaders in the field of economic evaluation in health care have long recommended that analysts seeking to inform resource allocation decisions approximate the value of information in terms of incremental cost per QALY gained*.

This is an impossible dream. Imaginary worlds looking 30 years into the future (but known to ICER and the illuminati) fail the standards of normal science. Building imaginary claims on nonsensical QALYs merely exacerbates the problem; exposing the imaginary modeling to even more ridicule. EQ-5D utilities are manifest scores; ordinal measures than cannot support the creation of QALYs. We have an analytical dead end; one that could have been avoided with only a nodding understanding of measurement theory by the illuminati 30 years ago. More likely is that we will revisit in 10 years, recognize that the leaders have prevailed and that ICER leads the next generation of imaginary cost-per-QALY imaginary worlds.

References

[1] Langley PC. 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

[2] Piglucci M. Nonsense on Stilts: How to tell Science from Bunk. Chicago: University of Chicago Press, 2010

[3] 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;21:119-25

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

[5] Svensson E. Guidelines to statistical evaluation of data from rating scales and questionnaires. *J Rehab Med*. 2001;33:47-48

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

[7] Tufts University, Center for the Evaluation of Value and Risk in Health. The Cost-Effectiveness analysis Registry https://www.tuftsmedicalcenter.org/research-clinical-trials/institutes-centers-labs/center-for-evaluation-of-value-and-risk-in-health

[8] Stevens S. On the theory of scales of measurement. *Science*. 1946;1(3):677-80

[9] Grimby G, Tennant A, Tesio. The use of raw scores from ordinal scales: Time to end malpractice. *J Rehabil Med*. 2012;44:97-098

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

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

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

[13] Stone G. The emperor has no clothes: What makes a criterion-referenced standard valid? Presented at the fifth fifth annual international Objective Measurement Workshop. 2002: New Orleans.

[14] Parkin D, Devlin N, Feng Y. What determines the Shape of an EQ-5D Index Distribution. *Med Decis Making*. 2016; 36:941-51

[15] Feng Y, Devlin N, Bateman A et al. Distribution of the EQ-5D-5L profiles and values in three patient groups. *Value Health*. 2019;22:355-61

[16] Devlin N, Shah K, Feng Y et al. Valuing health related quality of life: an EQ-5D value set for England. *Health Econ*. 2018;27(1):7-22

[17] Van Hout B, Janssen M, Feng Y et al. Interim scoring for the EQ-5D-5 mapping the Eq-5D-5L to EQ-5D-3L value sets. *Value Health*. 2013;15:708-15

[18] 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

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

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

[21] 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