MEANINGLESS MAGICAL MEASUREMENTS: ICER's BIZARRE BELIEF IN THE EQ-5D RATIO SCALE

WORKING PAPER No. 20 SEPTEMBER 2020

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

ABSTRACT

The Institute for Clinical and Economic Review (ICER) holds to a bizarre belief that the EQ-5D instruments (EQ-5D-3L and EQ-5D-5L) have ratio scale properties. The reason for this false belief is that ICER, to support its imaginary value assessment business case, must be able to create quality adjusted life years (QALKYs). This can only be justified, within the alternative reality of the ICER value assessment framework, if the utility score can be applied to estimated time spent in a disease state to create perfect health equivalent years. ICER is clearly wrong. That is, the scale must have a multiplicative property. This requires a demonstrable true zero; which as there are negative utilities, is an impossible requirement. The purpose of this working paper is to set out the case for non-existent ratio utility scales. The argument is quite simple: if the utility scale was not designed to have ratio measurement properties then it won't have those properties. ICER also deludes itself that the time trade off (TTO) scale has interval properties. It does not for precisely the same reason. While ICER is perfectly entitled to indulge in creating value assessments in an alternative reality which allows utility scales to have ratio scales and, in a wider context, to be able to ignore the standards of normal science in creating by assumption simulated lifetime models, ICER's intended audience should be aware that the models and evidence reports presented are nothing more than magical mystery tours.

INTRODUCTION

Central to the institute for Clinical and Economic Review's (ICER's) value assessment methodology is the modelling of lifetime incremental cost-per-QALYs. This requires the QALY to be constructed for the various disease stages a hypothetical patient population may pass through. This is achieved by multiplying time spent in the disease state by a utility score, which is assumed to be on a 0 – 1 ratio scale (dead to perfect health). This discounts time spent to equivalent time spent in perfect health.

For those who are unaware of the axioms of fundamental measurement a brief digression may inform. There are four main types of measurement scale; putting to one side conjoint simultaneous measurement which underpins Rasch Measurement Theory (RMT)[1]. These are: nominal, ordinal, interval and ratio. Each satisfies one or more of the properties of: (i) identity, where each value has a unique meaning; (ii) magnitude, where each value has an ordered relationship to other values; (iii) interval, where scale units are equal to one another; and (iv) ratio, where there is a 'true zero' below which no value exists. Nominal scales are purely descriptive and have no inherent value in terms of magnitude. Ordinal scales have both identity and magnitude in an ordered relation but the unknown distances between the ranks means the scale is capable only of generating medians and modes, and the application of nonparametric statistics. The interval scale has identity, magnitude and equal intervals. It supports the mathematical operations of addition and subtraction, and range of parametric statistics. A ratio scale satisfies all properties, supporting the additional mathematical operations of multiplication and division. Recognition and adherence to these fundamental axioms of measurement theory is critical if an instrument, including those designed to capture patient outcomes

is to have any credibility [2]. In the physical sciences this has been long recognized; accurate measurement is the key to hypothesis testing and the discovery of new facts. The same arguments apply to the social sciences. Unfortunately, they appear all too often to be absent in health technology assessment. Indeed, this seems to be the only 'discipline', I use the term loosely, where claims rest on entirely imaginary constructs.

UNRAVELING THE EQ-5D

The utility score most frequently used by ICER is the EuroQoL EQ-5D. There are two variants: the EQ-5D-3L and the EQ-5D-5L [3]. Both instruments ask respondents to report on five health symptoms or attributes: mobility, self-care, usual activity, pain/discomfort and anxiety/depression. The difference is in the response levels. The EQ-5D-3L has three: no problem, some problem and major problems; the EQ-6D-5L has five response levels: no problem, slight problems, moderate problems, severe problems and unable/extreme. Each symptom response level reported is on an ordered ordinal scale. That is, we know one level is 'worse' than another but we don't know by how much. Ordinal scales are commonly used in patient reported health status. The limiting factor is that they can only be analyzed by reporting medians and modes and the application on nonparametric statistics. They cannot support other arithmetic operations. Hence the QALY as an impossible mathematical construct.

To generate a utility score coefficients are applied to these reported ordinal levels and combined with other coefficients to yield an aggregate score that is subtracted from unity. The maximum (negative) score for the EQ-5D-3L is -1.59 which means the scale can create negative utilities (referred to as states worse than death), if death is arbitrarily given a utility of zero.

The coefficients, described as Time Trade Off (TTO) tariffs or table of weights for the EQ-5D-3L are:

Symptom	No problem	Some problem	Major problem
Mobility	0	0.069	0.314
Self-care	0	0.104	0.214
Usual activity	0	0.036	0.094
Pain/discomfort	0	0.123	0.386
Anxiety/depression	0	0.071	0.236

The TTO tariff (the ordinal EQ-5D-3L scale) is created by the following procedure:

(1) Full health = 1
(2) Constant term for any dysfunctional item (i.e., non-zero tariff by response level) = -0.081
(3) Mobility: subtract TTO tariff for some problem/major problem
(4) Self-care, usual activity, pain/discomfort, anxiety/depression [as for (3)]
(5) N3: if any major problem cited for any symptom = -0.269
(6) Create score = add up (2) thru (5) and subtract from unity.

As an example consider a respondent with the following patient symptom levels: mobility = no problem; self-care = major problems; usual activity = some problems; pain/discomfort = major problems; and anxiety depression = major problems. This is described as one of 243 health states and given the numeric code 13233. The calculation for the utility score is:

Utility (-0.220) = 1 – (0.081 + 0 + 0.214 + 0.036 + 0.386 + 0.236 +0.269)

The utility score is negative, defining a state worse than death (or at least an arbitrary point we call death). For those with a wet Sunday afternoon at the beach house, the number of possible negative utility states could be calculated. The number of negative utility scores for the general population and specific disease states could be estimated. The point is that irrespective of the distributions and the respective percentages of worse than death states, the fact is that the EQ-5D-3L (and EQ-5D-5L) can generate negative utilities. This means it is not a ratio scale (assuming we could even show it had interval properties). Yet belief trumps facts; in the ICER alternative universe it is assumed to have ratio properties.

Putting these utility scores on a number line with interval invariance properties (-0.59 to 1.0) gives the impression that the utility scale has interval properties. This is a false impression. We could have any arbitrary number line with ordered scores placed with random intervals or distances between the scores; this would better represent the problem. Unless the instrument is designed to capture interval properties we cannot make any claim to its distance: how distant is 13233 from 13223? Certainly we can compute a score (0.041) where the difference in these scores is, obviously, due to the value coefficients. But these are not designed to create an interval scale; they are regression or similar coefficients. We can create an ordered ranking but can say nothing about the equivalence of distances between the scores. We might just as well label 13233 = A and 13223 = B. We can say B is a 'better' health state than A as defined by the EQ-5D-3L but we don't know by how much. We can only make sense of response claims if the scores are interval scaled. The same issues apply to visual analogue scales.

THE ORDINAL EQ-5D SCORE

Thus is created the EQ-5D-3L ordinal utility score. The coefficients are those deemed to 'best fit' the data but the scale is not designed to have interval or even ratio properties. The issues with this approach were covered by simultaneous conjoint measurement introduced in the early 1960s and the emergence of Rasch Measurement Theory (RMT). While health economists and ICER seem singularly unaware of this it is worth paraphrasing Bond and Cox. Conjoint simultaneous measurement was proposed by Luce and Tukey in the early 1960s as a new type of fundamental measurement, which subsumed the other types [4]. A specific application was to provide a framework for detecting measurement structures in non-physical attributes (e.g., quality of life). In a slightly modified form this is now applied as RMT. Rasch measurement standards, following those of the physical sciences are designed to create instruments that have interval measurement properties (but not ratio properties; that is a difficult next step). 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* [5]*.* It is these rules that comprise RMT. To reiterate: RMT is designed to construct fundamental measures. CTT and IRT focus on the observed data, these data have primacy and the results describe those data. 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.**

A further issue that is ignored in those advocating the pre-eminent role of the EQ-5D in eliciting preferences is the absence of dimensional homogeneity [6]. In the physical sciences instruments are designed to capture and report on a single attribute. This avoids confusion in attempting to unscramble aggregate scores that are the result of combing different attributes as well as being, from the perspective of measurement theory, inconsistent with fundamental axioms. If attribute scores are to be combined then they must exhibit dimensional homogeneity. Otherwise we are left with a ratbag of the sum of ordinal scales that says little if anything about response to therapy; a multidimensional composite index. In the case of the EQ-5D-3L we are, with additional flourishes, combining five ordinal scales that are presumably attempting to capture some underlying attribute that defines the individual symptoms.

Dimensional homogeneity is critical to instruments that meet the standards of fundamental measurement. Variables can only be combined if they have the same dimension. If they fail, then they lack construct validity. It is invalid to add together variables that lack a common dimension. Hence the EQ-5D-3L lacks dimensional homogeneity. In mathematics all components of an equation must have the same degree of value or quantities of the same base units on both sides; only quantities having the same dimension may be compared, equated, added or subtracted. The EQ-5D-3L and other generic utility instruments and the majority of PROs fall, therefore at the first hurdle. They are a 'mashup' of ordinal scales. This is seen when we ask what the 'distance' is between the various symptom 'problem levels'. Does 'some problem' mean the same to different respondents? Does the application of community preference strictly define the distance between 'some problems' and extreme problems'? Are the distances the 'same' for each symptom level?' Are these questions meaningful?

IGNORANTIA SIT BEAUTITUDO: THE PEARSONIAN RATIO SCALE

It is one thing to ignore fundamental measurement requirements; it is another to apply this ignorance in situations where claims based on this ignorance have practical or policy implications. Perhaps we can draw an analogy with the motorist who consistently ignores STOP signs? This is the position ICER is in with its embrace of utility scores as ratio scales to support creating imaginary cost per QALY value assessments for pricing recommendations and product access when it has been informed repeatedly that it is creating imaginary worlds which fail measurement standards. Unfortunately, people may take you seriously; a classic example is the ICER belief in the impossible or I-QALY.

To illustrate the absurd position ICER is in, with its defense of what we may describe as the Pearsonian Ratio Scale (the ratio scale you have when there is no ratio scale), consider the following statement by ICER in the evidence review of TIMs in ulcerative colitis:

> *We (and most health economists)* ***have the understanding*** *(emphasis added) that the EQ-5D (and other multiattribute instruments) do have ratio properties. The EQ-5D value sets are based on time trade-off assessments (which are interval level) with preference weights assigned to different attributes. We fail to see why this should be considered as an ordinal (ranked) scale. ICER believes that the dead state represents a natural zero point on a scale of health related quality of life. Negative utility values on the EQ-5D scale represent states considered worse than dead* [7].

Apart from the obvious response that this is nonsense, let's deconstruct this alternative reality [8]**:**

- **It is not clear what 'have the understanding' means? Does it mean they know it is not a ratio scale but are prepared to believe it is because it would be embarrassing to deny it? Or do they truly believe it? Is this belief based on ignorance of the axioms of fundamental measurement? Or is the belief held strongly because the object of the belief, the ratio scale EQ-5D-3L is so improbable? The fact that thousands of health economists do believe It (or are the last word in cynicism) is evidenced by the thousands of papers modeling cost-per-QALY claims.**
- **If you apply preference weights to ordinal rankings you still get ordinal rankings; the ranking may change but the ordinal property is retained. You can't add value coefficients together and then assume that the utility score is on a ratio scale (without a true zero).**
- **Clearly ICER holds tightly to the belief (they quite frankly have no other option) that these utility scales have ratio properties. That is they have a true zero (no negative numbers) and can support all arithmetic operations: addition, subtraction, multiplication and division (the interval property is implicit). However, ICER cannot provide a proof of this assertion (because there isn't one).**
- **In one respect ICER might be commended, we await the proof, for a major step forward (for the first time since conjoint simultaneous measurement in the early 1960s) in measurement theory. The belief that there can be a ratio scale without a true zero and negative numbers! A ratio scale with utilities capped at unity and a floor at -0.59 that can support multiplication and the creation of QALYs. This rejection of over 100 years of measurement theory represents a major step forward. This unique development, which we can label the Pearsonian Ratio Scale, supports multiplication without a true zero.**
- **If it is a new ratio scale then we have the possibility of negative QALYs and the calculation of negative incremental cost-per-QALY claims. Patients, over their hypothetical existence in the lifetime simulation can then hop between negative and positive (and even zero) QALY states and then end up with an aggregate lifetime zero QALY count. It's not clear if, having died and entered the ultimate absorbing state (zero utility) how, Lazarus like, they can be resuscitated to enjoy another time in either a negative or positive QALY state.**
- **The fact that 0 = death in all utility scales (as they are capped with 1 = perfect health) means that they will have different proportions of states worse than death. A respondent may be in a state worse than death on one scale yet surprisingly alive and enjoying a positive health experience on another.**
- **Time trade off assessments are ordinal (health state preferences are ranked; we have no idea what the distance between them means (e.g., 23333 vs 33333) as well as lacking dimensional homogeneity** [9]
- **ICER fails to see why the EQ-5D should be considered an ordinal scale. Perhaps ICER should become acquainted with measurement theory and instrument development.**

- **ICER believes that the zero point represents a natural zero? Why? There is a substantial literature on the dead concept in health measurement. Questions have been raised regarding death playing a central role in valuation; can it be seen as a manifestation of health status?**
- **ICER recognizes that there are states worse than death (hence negative QALYs) but still maintains the EQ-5D is a ratio scale. Do we need to anchor on death? As this is an ordinal scale perhaps the answer is that anchoring is irrelevant.**

This is the fatal error. When a series of coefficients are estimated from an econometric modelling exercise, where the model is fitted to the data, considerations of fundamental measurement theory are put to one side. Why? Because if we are to meet these standards the instrument has to be designed from the get-go to have the required properties. Assuming that by some magical process the resulting scoring algorithm will have interval, let alone ratio properties, is just daft. In this case where a sample of TTO scores are the dependent variable the model is being fitted to an ordered ordinal sample. Why ordinal? Because in developing the TTO techniques for 'valuing' preferences for health states no concern was given to its properties. In fact, as has been pointed out the TTO gives no consideration to invariance of comparisons (required for interval scales) or unidimensionality. In this latter case the TTO, in attempting to capture five health status attributes ensures that it lacks dimensional homogeneity. Added to this is a long-standing concern with how states worse than death are captured in health preference scales and what interpretation is to be given to the zero death state. There appears to be a growing consensus that the TTO is past its use-by date and with it the ordinal EQ-5D. Disease specific instruments that meet required measurement axioms are moving to center stage to evaluate response to therapy, abandoning imaginary QALYs at the same time [10].

AN ALTERNATIVE REALITY

ICER has always existed in an alternative value assessment reality to one where claims are empirically evaluable [11]. The ICER reference case both defies logic as well as the standards of normal science. Understanding that the EQ-5D utility scale has ratio measurement properties is just one manifestation of this rejection. But ICER is not alone; it joins thousands of health economists over the past 30 years who are similarly misinformed. Accepting the importance, as is commonplace in the physical sciences of developing instruments to meet the axioms of fundamental measurement is, apparently, an alien concept. Unfortunately, if an instrument is not designed to have these attributes then, oddly enough, it won't have them. It is no good assuming after you have fitted a model with econometric techniques to a data set, as is the case of the EQ-5D instruments, that some magic sprinkling of ICER fairy dust will create a ratio or even an interval scale.

But we must not deny ICER the opportunity to employ nonsense assumptions to drive value assessments; an ICER alternative reality for pricing and access recommendations. We all need our place in the sun. After all, the notion of lifetime cost-per-QALY simulations driven by assumptions is itself nothing more than re-imagined intelligent design; given the widespread support for intelligent design it should come as no surprise that the ratio scale EQ-5D is accepted. While we might compliment ICER on the imaginary Pearsonian ratio scale as a significant contribution to measurement theory in an alternative universe, it is possibly time in this 21st century universe to

create value assessment frameworks that are not only logically coherent but meet the standards of normal science.

REFERENCES

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

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

[3] 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

[4] Luce RD , Tukey JW. Simultaneous conjoint measurement. *J Math Psychol*, 1964; 1:1-27

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

[6] 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

[7] ICER. Ulcerative Colitis: Response to Public Comments. September 2020. https://icer-review.org/material/ulcerative-colitis-response-to-public-comments/

[8] 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). *Inov Pharm*. 2020;11(3): in press

[9] 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

[10] 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

[11] Langley P. Nonsense on Stilts – Part 1: The ICER 2020-20234 value assessment framework for constructing imaginary worlds. *Inov Pharm*. 2020;11(1): No. 12 https://pubs.lib.umn.edu/index.php/innovations/article/view/2444/2348