Nonsense on Stilts – Part 1: The ICER 2020-2023 Value Assessment Framework for Constructing Imaginary Worlds

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

Abstract

Previous commentaries in the Formulary Evaluation section of INNOVATIONS in Pharmacy have pointed to the lack of credibility in modeled claims for cost-effectiveness and associated recommendations for pricing and access by the Institute for Clinical and Economic Review (ICER). The principal objection to ICER reports has been that their modeled claims fail the standards of normal science: they are best seen as pseudoscience. The purpose of this latest commentary is to provide a critique of the recently released ICER 2020 Value Assessment Framework (VAF). Although ICER has taken upon itself the pole position in health technology assessments and recommendations for product pricing in the US health care system, the incremental, lifetime cost-per-QALY modeling methodology should not be taken seriously. The creation of imaginary modeled worlds, built entirely from assumption, fails the demarcation test between science and pseudoscience. The ICER evidence reports are best seen as the health technology assessment equivalent of 'intelligent design' in counterpoint to 'natural selection'. It is surprising, therefore, that health care decision makers should take ICER's recommendations seriously as providing 'approximate information' for formulary decision making. What is not appreciated is that the claims made by ICER lack credibility, are impossible to evaluate and lack the ability to be replicated across treatment settings. Indeed, the models presented under the guise of a 'state of the art' value assessment were never intended to support evaluable claims. We have no idea and will never know if they are right or if they are wrong. ICER's position becomes even more untenable once the models presented are assessed in detail. Without in any way supporting the ICER methodology, it is worth noting that all too often ICER's claims for incremental QALYs in specific models are based upon what appears to be, from the limited evidence presented, a casual and ad hoc assemblage of utility scores from diverse constructs. This is a critical weakness given the role attributed by ICER to the modeled cost-per-QALY claims as central to ICERs imaginary value assessment. ICER also overlooks the fact that the utility scores it captures from the literature to populate its imaginary reference case world lack objectivity. They are ordinal rather than interval measures. To apply these manifest scores to time spent in a disease stage and then aggregate these over different disease stages is nonsensical. The critical issue is one of instrument development. The case made here is for the application of Rasch Measurement Theory (RMT) to construct a unidimensional instrument with interval properties, in this case from the needs fulfillment construct of quality of life (QoL). Unless an instrument meets RMT standards in its development, the logic of Rasch modeling to achieve fundamental measurement standards means that other scales are, by definition, ordinal. It is absurd to 'assume' they are interval. RMT is designed to create instruments to evaluate change and test hypotheses. In the absence of instruments that have RMT properties, the cost-per-QALY reference case modelling meme collapses. It is an analytical dead end. If we are to support a meaningful scientific program to discover new facts to support health care delivery and improve the lives of patients, caregivers and their families, then ICER should be put to one side.

Keywords: ICER, imaginary claims, Rasch model, measurement error, unnecessary distraction, nonsense QALYs, ignoring the patient voice

---

Introduction

The release of the 2020 version of the Institute for Clinical and Economic Review's (ICER) Value Assessment Framework (VAF) represents a millstone in health technology assessment [1]. It affirms ICER's ongoing commitment to the construction of imaginary worlds to support pricing and access recommendations. ICER sees its VAF as forming the *backbone of rigorous evidence reports that, within a broader mechanism of stakeholder and public engagement, will help the United States evolve towards a health system that provides fair pricing, fair access, and a sustainable platform for future innovation.*

Corresponding author: Paul C Langley, PhD
Adjunct Professor, College of Pharmacy,
University of Minnesota, Minneapolis MN
Director, Maimon Research LLC; Tucson, AZ
Email: langley@maimonresearch.com

While this no doubt laudatory objective must appeal to a wide audience, it is a backbone that fails the standards of normal science. To claim that ICER presents rigorous evidence is patently absurd when their modeled claims are deconstructed. While ICER sees its VAF as seeking to *place scientific methods and evidence analysis at the heart of a clearer and more transparent process*, the fact is that the modeled claims that drive recommendations for pricing ('fair prices') and access ('fair access') are imaginary and are the antithesis of 'scientific methods'. The ICER model is only one of a potential multiverse of competing imaginary worlds all with their own meaningless recommendations. The reference case model, as demonstrated in this commentary is not a 'sustainable platform'; it is an unnecessary distraction. More to the point: the so-called population perspective rests on generic utilities that fail the axioms of fundamental evidence; they are an ordinal manifest score which means that building quality adjusted life year (QALY) claims based on ordinal 'values' is nonsensical.

DOI: https://doi.org/10.24926/iip.v11i1.2444

Previous commentaries in *INNOVATIONS in Pharmacy* have, over the past 4 years, both reviewed ICER evidence reports as well as providing detailed critiques of the ICER methodology: in particular the failure of the application of the ICER reference case [2] to meet the standards of normal science [3 4 5 6 7 8 9 10 11]. The argument is straightforward. The ICER modeled reference case technology assessment fails the demarcation test and is best seen as pseudoscience (intelligent design) rather than normal science (natural selection). The lifetime reference case requirements fail to generate credible claims. By definition, the claims are not evaluable and, by extension, not replicable across treating populations. We have no idea if the claims are right or if they are wrong, we will never know and were never intended to know.

Certainly, the reference case methodology is seen as the 'state of the art' in health technology assessment which supports the construction of imaginary, simulated models projecting over the lifetime of a hypothetical patient cohort to generate incremental cost-per-QALY claims based on ordinal utilities. These claims are set against willingness to pay thresholds to convince an audience, who are typically non-technical, to take at face value recommendations for product pricing and access based on a hypothetical world. It is acknowledged by technology assessment groups that these are artificial (yet 'realistic') but that their redeeming feature, apparently, is that they generate 'approximate information' for decision makers; or, more precisely, 'imaginary' information (or disinformation) [12]. It is, perhaps, surprising, if not of concern, that a major focus of health economics is on the fabrication of imaginary worlds with a disregard of the axioms of fundamental measurement.

Even if attempts were made by a successor to ICER to use the reference case framework to create credible and evaluable claims, the result would still fail the demarcation test. This is because, apart from the lifetime perspective, the QALY construct fails to meet the required axioms of fundamental measurement. Generic utility 'values' and the majority of disease specific patient reported outcomes (PRO) instruments fail to meet, as detailed below, the fundamental measurement axioms of invariance of comparisons and sufficiency; the unidimensional properties of the Rasch model [13 14]. If health technology assessment has any hope of being taken seriously then it needs to generate claims that are disease specific, credible, evaluable and falsifiable; claims that are based on a coherent quality of life construct (QoL) and meet the standards, notably construct validity and order, for Rasch Measurement Theory (RMT) in instrument development. This has been recognized in papers presented in both the ISPOR house journal *Value in Health* (in 2004) [15 16 17] and more recently in the *Journal of Medical Economics* (2019) [18 19].

The purpose of this commentary is to review the 2020 Value Assessment Framework (VAF) which ICER proposes to use in future imaginary evidence reports and recommendations for pricing and affordability. The focus of this critique is to make the case that that ICER's VAF should not be taken seriously. Its modeled claims for value claims not only strain credulity, but judged against the standards of normal science are nonsensical.

While the ICER approach embraces the health technology assessment meme advanced by the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) and a number of single payer health system assessment agencies such as the National Institute for Health and Care Excellence (NICE) in the UK, the meme fails standards for fundamental measurement, discovery and the growth of knowledge that have been place for almost the past 400 years.

The manifest shortcomings in the ICER reference case modeling are considered from a number of perspectives:

- The construction, by assumption, of an imaginary world that is intended, not to test hypotheses, but to generate 'approximate information'
- A failure to point out that the ICER reference case can produce a multiverse of competing imaginary world all of which may have, or can be constructed to have, competing recommendations
- A focus on generic ordinal utility measures of 'benefit' that argue for 'measuring' health related quality of life (HRQoL) but which fail to provide a coherent patient-centric latent quality of life construct; as well as failing to meet the required standards of Rasch measurement theory (RMT) for instrument development
- Building a VAF around ordinal utility 'fabricated' QALYs which lack any meaningful interpretation in terms of the fundamental axioms of measurement theory
- Selecting an ordinal utility measure (e.g., EQ-5D-3L) to generate QALYs which is only one of a number of competing ordinal generic measures all of which fail to meet fundamental measurement axioms
- Putting to one side the 'voice of the patient' in ignoring value claims that are a direct measure of needs fulfillment

Building Instructions for Fabricating Imaginary Worlds

The 2020 VAF is quite clear: ICER is committed to its reference case to create imaginary assumption-driven lifetime models to track a hypothetical cohort of patients assumed to be treated with and responding to specific therapies. The objective of the economic evaluation is *to determine the incremental cost-effectiveness of the cost per unit of health benefit gained of one treatment over another*. The response to therapy, measured as imaginary QALYs, is a community preference weighted response. The building kit instructions to support any number of imaginary worlds are detailed in the reference case [2]. With few exceptions, the deterministic base case reference case is applied uniformly across all products and devices assessed by ICER. Evidence reports are presented and reviewed. Competing model structures are not considered.

DOI: https://doi.org/10.24926/iip.v11i1.2444

The common outcome measure, to support value claims and ICER recommendations, is the imaginary modeled incremental cost-per-QALY for a hypothetical patient population. This is, by assumption, a lifetime or course of chronic disease QALY framework; extending 10, 20 or 30 years into the unknown but modeled ICER-assumed future. The choice of utility metric, derived from a literature review, is a US preference based multi-attribute system. A specific measure (e.g., EQ-5D-5L) is not mandated although ICER's preference is apparently for the EQ-5D-3L. The imaginary lifetime incremental cost-per-QALY constructed measure is matched against willingness-to-pay thresholds ($50,000 to $200,000 per QALY) and recommendations made for possible price discounting together with budget impact assessments.

It is quite clear that there is no intention that the reference case modeled claims being presented should be in a form that allows empirical evaluation. Indeed, the structure of the reference case and its lifetime perspective means that any claims are impossible to evaluate empirically and were never intended to be evaluable. ICER defends its imaginary construct and claims in terms of validation against other imaginary worlds, the internal consistency of the model structure, discussions with the model builders and the choice of its assumptions. While this may seem an odd way to establish an 'imaginary' value, it is important to remember that this is in the realm of science fiction; adventures in imaginary VAF worlds. As detailed below, the failure to recognize that the utility measure fails to meet the standards of Rasch Measurement Theory (RMT) means that the ordinal lifetime QALY claims have no merit anyway.

Certainly, the ICER imaginary world model is hedged with mini-modeled scenarios to capture the effect of modifying imaginary assumptions, structural parameter assumptions and the application of sensitivity analyses. This does not change the fact that the model is an entirely imaginary construct which, as detailed below, fails the demarcation test, between science and pseudoscience. It is only one of a potential multiverse of imaginary constructs that may be applied to the specific disease area and therapies under review. There is no hint in the ICER guidelines for constructing imaginary world that empirical assessment might be considered or that any prospective audience member would be remotely interested in evaluable incremental cost-effectiveness claims.

ICER is not NICE

It is not clear where ICER's belief in its mandate both to perform as sole arbiter for health technology assessment in the US originates and to provide 'much needed' cost-per-QALY inputs to better manage resource allocation in health systems originates. There does not seem to have been a groundswell of opinion where US health system decision makers have approached ICER, pleading for mandated incremental cost-per-QALY modeled claims based on generic multi-attribute preference systems, as the national formulary decision criteria. Certainly the Academy of Managed Care Pharmacy (AMCP) has embraced the imaginary world meme in its formulary submission guidelines, but we also find instances of professional groups proposing alternative value metrics to support formulary decisions and guideline development [20].

Perhaps most importantly, there is scant evidence to show that health systems have the staff and the skills necessary to assess the merits, if any, of imaginary constructs or take these on board as a management tool. If it takes an 8 month gestation period for ICER to come to term in its clinical assessment of target therapies and its construction of imaginary cost-per-QALY worlds, it is difficult to see a comparable deconstructive effort from health systems. After all, if they take the view that the effort would be in pursuit of a pseudoscientific construct, the most reasonable response would be to put it to one side. Unfortunately, rather than deconstructing the imaginary ICER claims and recognizing their lack of scientific merits, insurers and health systems take them at face value.

ICER is in a quite different position from that of the National Institute for Health and Care Excellence (NICE) in the UK [21]. While ICER may be seen as a simulacrum of NICE (NICE-lite), the facts are that (i) it is not operating in a single payer health care system and (ii) it has no legislative role to provide guidance on the acceptance of technologies as NICE does within the English National Health Service (NHS). ICER's perceived and self-appointed role as the arbiter of value judgements for the US health market, supported by proclaimed processes of stakeholder involvement, clinical benefit assessment, model building and, ultimately, voting by an ICER appointed expert panel on the merits or otherwise of target therapies, should not obscure the fact that the end-results are imaginary value judgements.

NICE takes a reference case approach to establish model parameters, to generate incremental cost-per-QALY claims and apply willingness to pay thresholds. In this case, however, rather than an in-house model developed by its staff, manufacturers are asked by NICE to submit their own reference case model. This is typically a lifetime incremental imaginary cost-per-QALY model with the EQ-5D-3L generic HRQoL instrument as the standard utility measure. The point is that with NICE and other countries such as Canada (CADTH) Australia (PBAC), New Zealand (PHARMAC) and Ireland (HIQA) who have followed NICE's lead in mandating the construction of simulated or modeled worlds to support formulary submissions, the requirement has legislative and regulatory backing [22] [23] [24] [25]. Lifetime imaginary 'for approximate information' worlds are the required standard. While modeling reference case lifetime value judgements might seem odd and be objected to on grounds of scientific merit, there is an acceptance of this approach. As noted in a recent commentary: *The playing field is level and all parties know the rules of the 'game'. There are even imaginary world referees, typically in academic institutions, who will adjudicate the manufacturer's imaginary submission. They can pronounce whether it is*

DOI: https://doi.org/10.24926/iip.v11i1.2444

acceptable, modifiable or should be replaced by the referees own proposal for an imaginary world. NICE, as senior referee, is the judge [26] .

There is no reason why ICER should assume that value judgements based on constructed evidence from simulated worlds should have relevance to health care decision making in the US. The US is not a single payer health system. There is no legislated or regulatory across-the-board requirement for imaginary reference case modeling to support value judgements and formulary decisions. Certainly, ICER might believe in the sure and certain hope that incremental cost-per-QALY lifetime simulation models are the current and future 'state of the art' in health technology assessment, a position taken by professional groups such as ISPOR. This does not mean that the ICER business model standard is appropriate. Indeed, under the Affordable Care and Patient Protection Act (2010) it is made clear that the Patient Centered Outcomes Research Group (PCORI) should exclude discounted cost-per QALY or similar measures as threshold values for priority setting in health by the Centers for Medicare & Medicaid Services (CMS) [27]. While this exclusion may give pause to those advocating a role for imaginary QALYs in pricing and access, the debate overlooks a more substantive concern: the ICER business model lacks scientific merit. It is best seen, as detailed below, as pseudoscience.

Similar objections apply to the recent proposals by ICER for the modeled value assessments of transformative therapies (SSTs) [28] [29]. These have been reviewed in a recent commentary in *INNOVATIONS in Pharmacy* [30]. The proposed framework for SSTs will continue to use the reference case to construct imaginary worlds, subsumed within the overall VHF methodology, As such, they should be rejected.

Meeting the Standards of Normal Science

The requirement for testable hypotheses in the evaluation and provisional acceptance of claims made for products and devices is unexceptional. Since the 17th century, it has been accepted that if a research agenda is to advance, if there is to be an accretion of knowledge, there has to be a process of discovering new facts. Indeed, as early as the 16th century Leonardo da Vinci (1452 – 1519) in notes that appeared posthumously in 1540 for his *Treatise on Painting* (published in 1641) clearly anticipated the standards for the scientific method which were widely embraced a century later in rejecting thought experiments that fail the test of experience. By the 1660s, the scientific method, following the seminal contributions of Bacon, Galileo, Huygens and Boyle, had been clearly articulated by associations such as the Academia del Cimento in Florence (1657) and the Royal Society in England (founded 1660; Royal Charter 1662) with their respective mottos *Provando e Riprovando* (prove and again prove) and *nullius in verba* (take no man's word for it) [31].

By the early 20th century standards for empirical assessment were put on a sound methodological basis by Popper (Sir Karl Popper 1902-1994) in his advocacy of a process of 'conjecture and refutation [32] [33] . Hypotheses or claims must be capable of falsification; indeed they should be framed in such a way that makes falsification likely. Life becomes more interesting if claims are falsified because this forces us to reconsider our models and the assumptions built into those models. This leads to the obvious point that claims or models should not be judged on the realism or reasonableness of assumptions or on whether the model 'represents' for a public advocacy research group such as ICER their belief in lifetime comparative cost-per-QALY outcomes future fictional reality. A future reality that is unknown and unknowable, and is never intended to be known. This is an intellectual and analytical dead-end.

Although Popper's view on what demarcates science (e.g., natural selection) from pseudoscience (e.g., intelligent design) is now seen as an oversimplification involving more than just the criteria of falsification, the demarcation criteria remains [34]. Certainly, there are different ways of doing science but what all scientific inquiry has in common is the 'construction of empirically verifiable theories and hypotheses'. Empirical testability is the 'one major characteristic distinguishing science from pseudoscience'; theories must be tested against data. Indeed, paradoxically, while the development of pharmaceutical products and the evidence standards required by the Food and Drug Administration (FDA) for product evaluation and marketing approval are driven by adherence to the scientific method, once a product is launched and claims made for cost-effectiveness and, in the case of ICER, modeled for pricing and access recommendations, the scientific method is put to one side. Pseudoscience succeeds science. Darwin's (Charles Darwin 1809 -1882) on *The Origin of Species* [35] is succeeded by *Of Pandas and People* [36].

The rejection of a research program that meets the standards of normal science by groups such as ICER is best exemplified by the latest version of the Canadian health technology guidelines where it is stated: *Economic evaluations are designed to inform decisions. As such they are distinct from conventional research activities, which are designed to test hypotheses* [37]. While this position puts modeled health technology assessment in the category of pseudoscience, it is also what may be described as a relativist position. Rather than subscribing to the position that the standards of normal science are the only standards to apply in health care decisions and value claims, the relativist believes that all perspectives are equally valid. Health care decisions are to be understood sociologically. No one body of evidence is superior to another. Results of a lifetime modeled simulation are on an equal basis with those of a pivotal Phase 3 randomized clinical trial. For the relativist, the success of a scientific research program, in this case one built on hypothetical models and assumptions, rests not on its ability to generate new knowledge but on its ability to mobilize the support of the community.

DOI: https://doi.org/10.24926/iip.v11i1.2444

Basing decisions on models and simulations underpins the consensus view that evidence is constructed, never discovered. Instead of coming to grips with reality, science is from their perspective about rhetoric, persuasion and authority [31]. Truth is consensus.

The Health Technology Assessment Meme

If truth is consensus, how is this consensus, resting upon the construction of imaginary worlds, maintained; in this case for over 30 years of imaginary cost-effectiveness modeled claims. The ISPOR consensus, embraced by ICER, on health technology assessment has been characterized in previous commentaries as a meme. This is deliberate, as it underpins the interpretation of ICER's continued unqualified acceptance of the reference case as its core business model, as a sociological phenomenon.

After all, it is unusual to find the central pillar in an academically respectable social science the construction of fictional imaginary worlds; in this case to support non-evaluable cost-outcomes claims with the fabrication of 'approximate information'. In this context, the ISPOR/ICER cost-per-QALY reference case can be characterized as a unit of cultural transmission or unit of imitation; as an analog of gene pool propagation 'by leaping from body to body via sperm or eggs' [38].

One of the key health technology assessment meme tenets is the belief in the QALY. A venerated dogma which is central to value assessment. Human beings are good at imitation. The reference case meme, the faith in the QALY, appears to be adept in its infectivity, supported by an organizational infrastructure to defend it against competition in the technology assessment meme pool, ensuring survival through supporting propagation, longevity, fecundity (or acceptability) and, of particular note high copying fidelity. The control exercised over the meme ensures few mutations. As Dawkins notes, few individuals brought up in a certain faith switch to other faiths or reject the 'faith and mysteries' of their parent's belief system [39]. The widespread adoption and propagation of this meme in seen with literally thousands of imaginary world technology assessments published over the past 30 plus years. Add to this continued willingness of journal editors to publish imaginary claims, even if they are sponsored marketing exercises. In the case of this continued acceptance it is sufficient to point to the advocacy of the meme by organizations such as ISPOR with its global membership, its good practice guidelines for constructing imaginary worlds, training programs for newly arrived imaginary world apprentices, and conferences, together with endorsements from technology assessment agencies such as NICE, CADTH and the PBAC. Add to this its place in university post-graduate programs (including Colleges of Pharmacy) together with the contribution of textbooks that have rigorously supported the creation of imaginary worlds [40].

It is of interest to speculate, given the receptive audience for imaginary technology assessment claims, together with the 'technical' belief structure that underpins them (e.g., probability sensitivity analysis), whether or not we are receptive to pseudoscientific claims; is there a response bias toward accepting pseudoscientific claims as true? Is there an asymmetry between belief and unbelief? Is additional processing required if this bias is to be overcome? Is there an asymmetry that reinforces acceptance of the technology assessment meme and the acceptance of 'imaginary approximate information' even though it is a 'mystery' as to what this actually means? Perhaps we just accept it on 'faith'?

A further possibility is our inability to detect pseudoscientific constructs. Do we judge something as profound because we have failed to understand it? Are there measurable differences in the ability of individuals to discern or detect pseudoscientific statements including the more complex (and often obscure) modeling constructs supporting ICER imaginary worlds and attendant scenarios? Can we engage in analytic thinking? Do we understand the axioms of fundamental measurement? To what extent is our ability to reflect on, rather than reflexively accept at face value, offset by our acceptance of a belief system that is central to our professional standing?

Perhaps we should not be surprised that the nature of the scientific method is not appreciated. After all, some 27% of Americans don't accept heliocentrism, 48% don't accept common ancestry (natural selection) and 61% don't accept the big bang [41] . Even so, we should not be unduly pessimistic as a recent survey indicated that probably less than 2% of Americans believe in a flat earth, although globally traveling flat-earthers seem active on the conference front.[42]. There are, of course those who require visual evidence. One respondent remarked that he did not believe in gravity because he could not see it. Presumably he did not believe in imaginary worlds either.

Approximate Information

Central to the ICER business model is the need to express value judgements on constructed estimates of lifetime incremental costs-per-QALY. This is an article of faith for those supporting the health technology assessment meme. As stated by Neuman et al in the 2018 ISPOR Task Force Report on health economics in value assessment: *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 interventions in terms of incremental cost-per-QALY gained* (emphases added) [43]. It is not clear, in constructing imaginary worlds, how we are to distinguish the 'approximate' from the 'non-approximate' information content and how this 'approximate information' factors into formulary decisions.

If we accept the primacy of the scientific method, as a tool for discovery, over the recycling of 'evidence based' assumptions to create imaginary modeled claims, then any defense of the reference case as fabricating 'approximate information' to 'support' formulary decisions, pricing and access, seems odd. After all, we could equally well talk about 'approximate

DOI: https://doi.org/10.24926/iip.v11i1.2444

disinformation'. When does 'approximate information' mutate to 'approximate disinformation'? Presumably, the role of academic referee centers for imaginary worlds is to render judgement for a Jesuitical 'housekeeping seal of approval' on the 'chosen' imaginary world. After some 338 years Galileo would, no doubt, appreciate the irony (Galileo Galilei 1564-1642). Although the criteria are not entirely obvious, the preference would be presumably for one set of assumptions to drive a hypothetical world 30 years into the future, with imaginary value claims that are 'nearly precise or correct' when matched to an imaginary (yet unknowable) lifetime claim. Crystal ball or tea leaves?

Models and Assumptions

It is accepted that knowledge is provisional and permanently so. This stems from the obvious point that we can at no stage prove that what we 'know' is true. Attempting to believe or justify our belief in a theory is logically impossible. What we can do, by empirical assessment, is to try and demonstrate our preference for one theory over another (and apply it to the best of our knowledge) [32].

Constructing imaginary worlds which were never intended to generate potentially falsifiable claims cannot, therefore, be defended by an appeal to the 'truth' of their assumptions. If a health technology assessment claim is built upon a series of assumptions, a reasonable question is to ask what is the status of the various assumptions? Are they to be viewed as 'reasonable' or 'realistic' metrics for an unknown future reality? Can we just assume that utilities have interval scale properties? Have the utilities been selected from the literature because they seem appropriate? Are they the 'best available' from limited data? Are they all that are available?

More to the point, there is a belief in the fact that when the selected assumptions are based, where feasible, on an empirical study, this validates the choice of assumption. If the model is intended to incorporate utilities that have been reported in one or two studies (usually as few as that) for progression and time spent in the stages of a disease over a hypothetical future lifetime, then there is an immediate methodological issue. To claim that an assumption is valid is to revisit Hume's induction problem (David Hume 1711-1776): an appeal to facts to support a scientific statement. Unfortunately, as Hume pointed out, no number of singular observations can logically entail an unrestricted general statement. Certainly, there may be comfort in reporting that 'so far' the claim that all swans are white has not been contradicted (until that Qantas vacation in Western Australia) so that one fully expects the next swan to be white. But as Hume pointed out, this is a fact of psychology and does not entail any general statement. From a utility perspective, the fact that one hundred papers have agreed (within limited bounds) generic ordinal utilities from the same instrument for a target population in a disease state stage is immaterial. We cannot secure this assumption: it cannot be '*established by logical argument, since from the fact that all past futures have resembled past pasts, it does not follow that all future futures will resemble future pasts*' [44]. Claims, for the relevance of a constructed imaginary world built on the assumption that the model elements have been validated by observation is simply nonsensical.

Despite ICER's continued embrace, logical positivism is dead. It died some 80 years ago. All knowledge is provisional. Popper's contribution was to make clear that Hume's problem with induction can be resolved. We cannot prove the truth of a theory, or justify our belief in a theory by attendant assumptions, since this is to attempt the logically impossible. We can only justify our preference for a theory by continued evaluation and replication of claims. Constructing imaginary worlds, even if the justification is that they are for 'approximate information' is, to use Bentham's (Jeremy Bentham (1748-1832) memorable phrase 'nonsense on stilts'. If there is a belief, as subscribed to by ICER, in the sure and certain hope of the relevance of approximate information created by imaginary worlds, a belief to drive formulary and pricing decisions, then it needs to be made clear that this is a belief that lacks scientific merit.

Certainly, assumptions can be a critical element of models; the difference is the models should support testable hypotheses. This is echoed by Newton (Isaac Newton 1642-1727) with Descartes' as his target (René Descartes 1596-1650) in saying '*hypotheses non fingo*' (I do not feign hypotheses). Descartes in Newton's view had 'produced fantastic and untestable ideas, then assumed them to be true and used them as building blocks of his philosophy ' [45].

Measurement: The Rasch Model

Measurement is critical for the advance of science, through hypothesis testing and the discovery of new facts. Even so, the appearance of fundamental measurement scales is rare. The majority of measurement scales, if they are to meet the axioms for invariance and sufficiency have to be constructed, standardized and agreed. The 17th century, with the invention of science, witnessed a focus on constructing instruments to meet fundamental measurement standards. In many cases it took decades for agreement on the appropriate tools; consider thermometry, the invention of temperature [46] .

Where the object to be measured is a psychological or non-physical construct (e.g., quality of life) the situation in the social sciences is more complex. It was not until the early 1960s that recognition of the fundamental axioms of conjoint simultaneous measurement provided a framework for going beyond the notion of simple interval and ratio scales, to propose a framework for detecting, if they exist, measurement structures in non-physical attributes with interval and possibly ratio properties. That is, unlike ordinal scales where only the order of values matters and we can say nothing about the difference in the values, the interval or cardinal scale is one where we know the order and the exact difference. Unlike the ordinal scale which allows only statements regarding the mode

DOI: https://doi.org/10.24926/iip.v11i1.2444

or median, interval scales allow addition and subtraction. This permits calculation of measures of central tendency and dispersion (e.g., effect size). However, as the interval scale does not have a true zero, we cannot compute ratios (i.e., multiplication and division). It is only with ratio scales that we have a true zero. The seminal contributions are those by Luce and Tukey, and Rasch [47 48].

The critical step is to recognize the contribution of Rasch Measurement Theory (RMT) to constructing outcomes instruments in health technology assessment [14]. As noted below, the criteria for designating a scale as meeting the axioms of fundamental measurement, is to develop the instrument by application of Rasch model standards. Otherwise, the instrument will generate nothing but ordinal manifest scores.

The Rasch contribution is to recognize the need, if we are to develop the analog to measurement in the physical science, to produce the data (items in a questionnaire) to fit the Rasch model, not in, for example Item Response Theory (IRT) to fit the model to the data. As an example, for a mathematics test, a matrix may be defined by the ability of examination candidates (row elements) and the difficulty level of the various items in the test (column elements). Patterns of relationships between the cells, where each cell gives the probability of an outcome (Yes/No) as the difference between the difficulty of an item and the ability of the student can be determined by applying the axioms of conjoint simultaneous measurement.

The Rasch model, although developed independently of Luce and Tukey, utilizes a modified form of the axioms of conjoint simultaneous measurement, to assess patterns in a matrix of expected response probabilities; again as a function of differences between ability and difficulty [17]. The unidimensional Rasch model, a focus on a single attribute or dimension captured in a latent construct, rests on two 'order' premises:

- The easier the item, the more likely it is to be affirmed; and
- The more able the respondent, the more likely are they to affirm an item

Data inputs have to fit the Rasch model [14]. This is in contrast to classical test theory (CTT) where the model is applied to the data. If the data items fit the Rasch model, they are translated from ordinal scores to interval scores where the unit of measurement is the logit or logs odd unit. The Rasch model rejects raw scores. Rather, a log-odds transformation is applied to these ordinal attribute measures to create a Rasch relative distance or interval measurement scale. This scale avoids the 'clumping' of raw scores around the middle scores and enhances the contrast in results for, in the case of ability, for those at the extreme values of the scale. The purpose of the Rasch model is to build a measurement tool (a list of items, tasks, questions) that will make a meaningful assessment of a latent construct. Difficulty is relative to the other items in the scale. Each item on a unidimensional scale should contribute meaningfully to the construct being evaluated. Hence the importance of a data matrix that relates respondents and items coherently, is one that is more likely to represent the construct; hence the importance of fit statistics in developing the item, respondent choice and order.

As the Rasch model, as part of its development, establishes construct validity as well as the application of other CTT assessments to the items provisionally selected prior to fitting, and if necessary rejecting items that do not give a good fit, the resulting item–based instrument has good psychometric properties. If a scale is to provide fundamental measurement, this is to be established prior to the evaluation of psychometric properties. This points to the importance of ensuring that respondents are a random sample of the target patient population with an assumed distribution of abilities that matches that of the target population. Potential items for incision in the Rasch model are created from qualitative interviews with the sample.

Quality of Life as a Construct

The foundation for a Rasch model is to agree a coherent single attribute or construct of what is to be measured. These are not clinically determined health status dimensions of symptoms with ordinal response levels as in HRQoL measures (which are not constructs but a collection of clinical responses which are operational not conceptual). Rather the focus in QoL should be on a patient-centric needs fulfillment construct that was proposed over 20 years ago [15 16]. The application of constructs in science is common as they order observations. In psychology, to include QoL, the same objective holds. A coherent construct focuses on *attributes of people, situations or treatments that are reflected in responses to scales or other observations* [49] . The construct theory defines a variable (QoL) in terms of a model with a limited set of predictor variables (items). *The validity of a construct theory reflects the extent to which it predicts variations in item values and person scores* [16]. The litmus test is objective measurement: going beyond a simple ordering (ordinal scales) to data-based calibrations in a common calibration that have interval (and possibly, ratio) scale properties. If a QoL instrument, or other outcomes instrument, fits the Rasch model then the required level of calibration has been achieved. Otherwise, it should be discarded as the response or functional levels are, by definition, on an ordinal scale. This may reflect in each symptom item responded to by the patient indicating a change from one 'level' to another (response profile) or, for the more adventurous, some attempt to aggregate over the item level responses. The result is still a comparison of manifest ordinal scores.

Needless to say, few PROs are developed to capture the unidimensional requirements of the Rasch model. Rather, we have a top-down approach where physicians or expert groups, with typically minimal patient input, decide on the symptoms (health dimensions) and ordinal (e.g., Likert scale) responses

DOI: https://doi.org/10.24926/iip.v11i1.2444

within symptoms. The patient voice is effectively ignored. The axioms of fundamental measurement are ignored (or never considered). The primary function of the instrument is to provide input to the treating physician for those symptoms and responses judged to be of clinical interest.

Assuming Interval Measurement

The application of RMT is to ensure that the resulting instrument meets the axioms of fundamental measurement. There is no debate over this [13 14]. RMT was designed and applied to generate unidimensional measurement for a single latent attribute. It is consistent with the standards for instrument development in the physical sciences. Claims that a generic or disease specific PRO has unidimensional properties can only be sustained if it has been developed from day one within the Rasch framework or if, with possible item elimination, it can be demonstrated to have Rasch properties. If an instrument fails to meet Rasch standards, then it is, by definition, an ordinal instrument generating manifest scores.

Ordinal measures are manifest scores; they cannot be applied to arithmetic operations. They cannot be used to multiply modeled time spent in a disease state by a utility (any utility will do apparently) to produce a QALY score. If time spent is modeled at 2 years then multiplying it by an SF-5D-3L utility of 0.5 to yield 1 QALY is complete nonsense. Just because we label a utility as 0.5 by the application of a preference algorithm to create a number line of 0 to 1 (and putting the oddities of negative utilities aside), does not mean that the number line guarantees interval properties. It is just a space for placing manifest adjusted scores. We have no idea of what the distance between manifest scores means; 0.5 to O.55 does not 'mean' the same distance as 0.65 to 0.70. The number line could equally well have been anchored between 110 and 200. We could just as well have labeled the manifest scores A, B, C and D. There is no true zero (which means multiplication and division are disallowed).

Otherwise we have the absurd position, exemplified in the latest edition of probably the most widely used textbook in health care program evaluation, where it is simply assumed that utilities have interval properties on the 0 – 1 scale. This is not acceptable even as a 'simplifying assumption' [40].

Failing to apply Rasch standards has a major impact on the ICER business model. While previous commentaries have pointed to the lack of scientific merit in the creation of imaginary worlds, the more recent commentaries have brought fundamental measurement to the fore. The construction of QALYs and lifetime QALY estimates are a complete nonsense for one reason: the utility scores that they depend on fail to meet standards for fundamental measurement. Constrained, in the case of utility scores, to an arbitrary range 0 – 1 (which allows mythical QALYs to be created), distortion can occurs at the margins with bunching toward the extremes. Mathematical manipulations are not logically valid. *They are incompatible with the construction of fundamental measurement* [13] .

Attempts to apply Rasch analysis to consider the possible unidimensional character of utility systems do not augur well for blanket assumptions of interval properties. The data requirements are demanding because access is needed to individual responses. A recent study that assessed the EQ-5D-3L and Eq-5D-5L in persons with back and neck pain in Sweden receiving physiotherapy in a primary care context, while finding goodness-of-fit evidence for unidimensionality, found little other evidence to meet RMT standards, including item selection, limited response differentiation, differential item functioning and differential test functioning [50].

Also of interest is an analysis undertaken some 10 years ago of what the authors describe as the EQ-5D VAS measure [14]. This study combined the EQ-5D with the EQ-VAS (visual analog scale) to see if together they could form a valid interval scored measure of HRQoL in a US representative sample with the most prevalent chronic diseases. The VAS responses were collapsed to form a 9-category item. The Rasch rating scale model was used to calibrate the responses on the EQ-5D-3L items and the Rasch partial credit model for the 9-category VAS scores. The EQ-5D item anxiety/depression consistently showed misfit. This was improved with the addition of the VAS item. The findings suggested: (i) The EQ-5D and EQ-VAS can be combined to provide an overall measure of HRQoL; (ii) they might serve as a suitable measurement framework for deriving population preference weights; and (iii) important gender-specific reporting differences created measurement disturbances for the anxiety/depression item and for the anxiety and depression disease groups. These results were only reported as posters and summarized by Bond and Cox [14]. No other developments can be found. While of interest, these results are not sufficient to challenge the ordinal nature of the EQ-5D-3L and EQ-5D-5L. As a bolt-on item, the EQ-VAS is not present in reported EQ-5D scores. It is unlikely that this will be followed up.

The evidence for a blanket assumption that all generic utility systems yield, inadvertently, interval properties across all disease states, to meet Rasch standards is hardly compelling. Evidence for the other utility systems was absent (at least from a systematic review). A key point to note is that the Rasch model is disease and target patient population specific. If ICER wishes to assume that in a target disease state and patient population the chosen utility system (e.g., EQ-5D-3L) has interval properties than this has to be demonstrated in each case. This is an impossible task. Respondent data would be required for the target population. The analysis may suggest item elimination or even possible bolt-on items (e.g., VAS scores). Then the resulting instrument would have to be recalibrated for preference scores and utility algorithms applied to yield a score on a 0 – 1 scale. It is easier to assume interval scoring and hope nobody notices.

Given ICER's commitment to the EQ-5D-3L as the 'utility measure' of choice, the absence of a commitment to Rasch standards means that all ICER evidence reports to date are

DOI: https://doi.org/10.24926/iip.v11i1.2444

based on false measurement assumptions. It is not a question of ICER making a 'reasonable assumption' (among many other model assumptions) but of the reference case methodology failing to engage with outcomes instruments that meet, by construction, the axioms of fundamental measurement.

At risk of repetition, if an instrument does not utilize RMT in its construction then it is considered ordinal. Instruments that have relied on CTT for their development will always be ordinal. RMT is designed to create instruments with fundamental measurement properties. While this is an obvious point, it is critical. If our understanding of the impact of new therapies is to be understood then, as in the physical sciences, measurement is key. Non-physical attributes such as needs-based QoL present a challenge, but one that was met some 60 years ago. This is in obvious contrast to the construction of imaginary or fantasy lifetime reference case worlds which lack any pretense to set the stage for hypothesis testing, relying instead on the weak defense that they provide 'approximate information' (or disinformation) which might possibly be of interest to decision makers, or the more credulous, as is the case with CVS and ICER recommendations [51].

Nonsense with Ordinal Utilities

Clearly, to maintain face, ISPOR, ICER and their leaders could continue to argue the case for ordinal utilities as obvious surrogates for interval scales. The argument here is that this is untenable. Unless an instrument meets demonstrable Rash standards for unidimensionality, to reflect a single latent construct, it will always be ordinal. Claiming that the instrument was developed using CTT is insufficient.

Once we acknowledge the absence of interval measurement properties for the EQ-5D-3L, due to its development neglecting RMT requirements (assuming that its HRQoL characteristics defined a meaningful construct), then QALY modeled imaginary claims collapse. We also have to put to one side virtually all PRO measures that have been promoted and applied over the past 30 years or more using CTT. Indeed, to emphasize the point, a basic assumption of rating scales is that they measure a common underlying construct. We have to be able to demonstrate unequivocally that the instrument is based only on a coherent single construct and meets Rasch standards for, unidimensionality. Put simply, the implications of meeting the fundamental axioms of invariance in comparisons and sufficiency are essential.

In short, for those who subscribe to the technology assessment meme, we have to put to one side models that fail to meet the RMT standards. These include multi-attribute utility systems such as the EQ-5D, the Nottingham Health Profile, Item Response Theory (IRT) measures and the PROMIS system. To these would be added the hundreds of HRQoL instruments and others claiming to capture broader concepts of QoL.

Rasch Confirmatory Analysis

Over the past 20 years there have been many examples where Rasch analysis has been used to evaluate a PRO instrument for potential interval properties and the creation of summary scores. In some cases the extent to which the original PRO ordinal scales fit the Rasch model has involved minimum item reduction. One example is the Gibbons et al report on a Rasch analysis of the Motor Neurone Disease (MND) Social Withdrawal Scale (SMS) [52]. Recognizing that the original instrument developed with classical test theory (CTT) would always, by definition, be ordinal, the 24 items in the original were assessed for their factor structure and evaluated for model fit, category threshold analysis, differential item functioning, dimensionality and local dependency. The four factor solution of the original instrument was confirmed with Mokken scale analysis suggesting the removal of one item and Rasch analysis a further three. Following this, each of the four scales exhibited excellent Rasch model fit. A 14-item summary scale was shown to fit the Rasch model after dropping one of the sub-scales. This provided a total measure of social withdrawal.

Other examples include the Hospital Anxiety and Depression scale (HADS) in MND where Rasch analysis led to minimum item reduction for the two constituent scales [53]; an analysis of the Mini-Mental Health Adjustment to Cancer Scale (mini-MAC) which required more extensive item reduction [54]; the minimum item reduction for the Depression Anxiety and Stress Scale [55]; and, against these, a Rasch analysis of anxiety scales in Parkinson's disease where it was concluded that none of the currently used anxiety scales had satisfactory measurement properties [56].

At the same time there have been a number of cases where an instrument has been developed *de novo* utilizing the Rasch development model. These include a number of needs-based, disease specific, quality of life instruments developed by Galen Research [57]. A previous commentary detailed these instruments for rheumatological diseases [10].

ICER and the Indefensible

It is of passing interest to speculate on how ICER might respond to this commentary where the charge is, to use a well-worn metaphor, that the 'emperor has no clothes'. The obvious response is to claim that the 'belief' that utility number lines have interval properties is a 'state of the art' assumption or dogma central to the health technology assessment meme. It does not have to be demonstrated, we simply accept the 'mystery' of the meme. To challenge it would be an affront. Our belief is that much stronger if it is not challenged. As Kant (Immanuel Kant 1724 – 1804) wrote in the preface to the second edition of the *Critique of Pure Reason*: *I have therefore found it necessary to deny knowledge, in order to make room for faith* [58]. Kant was always committed to scientific knowledge, but knowledge limited to experience and not metaphysical ideas.

DOI: https://doi.org/10.24926/iip.v11i1.2444

A further response might be that as ICER has opted for a population focus, that generic HRQoL is a common framework. Acceptance by agencies such as NICE supports the proposition that, putting concerns with interval measurement to one side, the construction of imaginary incremental cost-per-QALY worlds for pricing and formulary decisions is a useful imaginary creation. After all, the claims can never be challenged. Perhaps it is just a game. We know the rules and assumptions, even though they may ignore the standards of normal science, but we play the game. To the extent that there are rules and referees to manage proposals and propose the most 'realistic' construction of future worlds where ordinal scales become interval, we endeavor to persevere and transmit our meme to future generations.

The value that individuals might attach to change in health status is ignored. ICER would put these concerns to one side and argue that the 'state of the art' in modeling imaginary worlds demands population preferences or weights to drive ordinal utility scores, create QALYs, and propose meaningless measures of incremental 'change'. If we are to introduce some consideration for the 'patient voice', meeting the needs of patients, then this is seen by ICER as an afterthought. Certainly concerns might be raised, such as the scope of the health dimensions, the relevance of response level sensitivity, the input from caregivers and issues such as access to care, but these are secondary to the central role of community preferences in driving clinically focused therapy response captured in a handful of ordinal health manifest scores to generate nonsensical QALY estimates and pricing recommendations.

ICERs insistence on its multi-attribute generic utility score to support its imaginary QALY claims means that it sees no role for disease specific measures that meet RMT criteria in assessing competing products. ICER's contribution is to trawl the literature for likely ordinal utility candidates to populate its modeled imaginary worlds. ICER has no intention of discovering new facts; it has no commitment to the contribution of normal science, rather it is concerned to provide 'approximate information' that fails the requirements of fundamental measurement.

ICER might insist on a 'population' perspective for the imaginary VAF model arguing that: *Taking a population perspective implies that the ICER value framework seeks to analyze evidence in a way that supports population-level decisions and policies, such as broad guidelines on appropriate care, pricing, insurance coverage determination, and payment mechanisms*. Again while these are no doubt commendable objectives in health care planning the population focus suffers from two limitations: (i) the value metrics that are presumably intended to support policies and guidelines fail the standards for fundamental measurement; and (ii) the insistence on clinically operational symptoms and responses rather than measures which are patient centric and which, again, fail the required measurement standards. Happily, though, the ICER model VAF *creates an explicit place and role for consideration of elements of value that are important to individual patients but that fall outside traditional clinical measures.*

It would have been more useful for ICER to recognize, not only the constraints of ordinal measurement, but that we have the Rasch model for taking explicit account of patient needs with RMT standard instruments available across a range of disease areas. Abandoning a population focus and ordinal utilities would open the doors for ICER to focus on the assessment of new and competing therapies, to discover new facts, with models designed to generate credible claims. Not only would the instruments have interval properties, but their potential focus on a needs-fulfilment QoL construct would put functional status to one side in favor of the patient voice.

Which QALY Metric?

If the embrace of the health technology assessment meme is to embrace an indefensible standard, then the ICER business model collapses. However, for the purpose of argument, if we put the lack of scientific merit to one side, then there are issues which ICER has yet to address if it persists in its dogged belief in imaginary ordinal worlds. A major concern must be the indiscriminate use of the term 'QALY' in the ICER evidence reports. The impression is given, perhaps inadvertently, that there is some objective ordinal QALY 'gold standard metric'. Unfortunately, different utility metrics and different models will create different QALYs even for the same target population and therapy comparisons. What ICER appears to have overlooked is that the various generic ordinal utility scales are different. As Drummond et al point out, there is no simple answer to the question of which preference based multi-attribute health status system to use, or whether to opt out : (i) the decision does matter as the systems are far from identical, they differ in the health dimensions and levels assigned to each dimension, in the description of those levels and in the severity of the most severe level; (ii) they differ in the population surveyed in the construction of the system and the instruments used to determine the preference based scoring; and (iii) they differ in the theoretical approach taken to modeling the preference data into a scoring formula [59]. Even so, the issue of fundamental measurement is not raised.

Although a review of previous ICER evidence reports might have suggested that ICER has, in practice, adopted the EQ-5D-3L system as the preferred ordinal metric, this is put to one side if a literature review is not successful in locating the preferred metric for target patient populations. In this event ICER has on occasion brought in another utility metric. This is seen in the ICER modeling of oral semaglutide for Type 2 Diabetes (T2DM) where it is stated: *The utility values for events modeled from the risk equations were drawn from two sources due to a lack of a single comprehensive source of health-related quality of life inputs. It is also important to point out that the two sources used different preference-weighted measures (EQ-5D and*

DOI: https://doi.org/10.24926/iip.v11i1.2444

HUI3), and these two instruments are known to produce slightly different utility estimates (emphasis added) [60]. Not only is the case for cardinal measurement overlooked, but for ICER the choice of utility metric appears to be incidental to capturing any utility assumptions for modeling.

For an organization which sees itself in pole position for excellence in reference case health technology assessment modeling in the US, as the arbiter for 'state of the art' standards in the modeling of imaginary worlds, this is a most unfortunate statement. There are no references given for this claim, specifically for references supporting this claim for the target T2DM target population which ICER is attempting to model. Indeed, if ICER is to make unsupported claims for ordinal utility 'equivalence' (i.e., they may be different constructs but we assume they are pretty much the same) then it should have provided a systematic review of ordinal utility metrics in the target T2DM population. Only then, for those who believe in the construction of imaginary worlds, could this assumption have been justified. If ICER is to provide justification for its utility choice then ICER should apply the review standards proposed by the ISPOR SPRUCE checklist (Minimum Reporting Standards of Systematic Review of Utilities for Cost-Effectiveness Models) [61]. These standards might, in retrospect, be re-labelled to include the term 'ordinal'.

It is also of interest to note that, for those trawling the literature for ordinal utilities to populate imaginary worlds there is what might be described as an 'ordinal' utility emporium, the Tufts University Cost-effectiveness Analysis (CEA) Registry, which since 1976 has assembled a database of over 8,000 cost-utility analyses [62]. Apparently, for those utility models selected for inclusion some 40 data points are extracted. These include utility values and cost-per-QALY claims. There does not seem, however any check on whether the utility scales meet the required axioms of fundamental measurement. While one would hesitate to describe this confection of cost-per-QALY studies as a redundant undertaking, this assemblage would surely rank alongside the relic collection of Federick III, Elector of Saxony (1463-1525)[63].

A further issue which deserves attention is the application of mapping to generate ordinal utility metrics from clinical markers. It is unusual to find protocols for Phase 3 clinical trials mandating the use of a multi-attribute utility instrument as an endpoint (either primary or secondary). As a result, as the ICER reference case notes, a utility metric can be created from a selected clinical marker. There is an extensive literature for this as the incentive, at least outside of the US, is to respond to guidelines by NICE and others to populate their ordinal reference case models with a generic multi-attribute metric for formulary submission. As a result considerable effort has gone into developing mapping algorithms to populate imaginary constructs. Putting questions of fundamental measurement to one side, two issues are important: (i) the choice of mapping algorithm and (ii) the choice of clinical marker. As to the latter point, there may be a number of relevant clinical markers that may capture the stage of disease and the response to therapy. There has to be justification for the one that is used. As a result objections may be raised that this is not the most relevant, even though the 'ordinal' exercise is pointless.

Unfortunately, ICER provides no guidance as to the choice of indirect or mapped ordinal utility metric? Are the mapping algorithms relevant to the target reference case hypothetical population? Given these uncertainties, it would be appropriate for ICER, not only to report, as noted above, on the utility metrics reported in the literature for the target disease state and population, but to provide a detailed justification for the choice of mapping function and metric. In this respect ICER should follow the ISPOR good practice guidelines for mapping from non-preference based outcomes measures [64]. This, once again, may be of academic interest yet from a practice perspective is pointless.

A Multiverse of Imaginary Worlds

If we accept the belief that the central role of health technology assessment is to construct 'approximate information' ordinal imaginary worlds, then it is reasonable to point out that there is a potential multiverse of imaginary ordinal worlds generating a potential tsunami of conflicting ordinal and illogical cost-per-QALY claims. While it might not be clear as to whose value (physician, patient, insurer, health system) a model is considered to represent, the model builder can press forward in the sure and certain hope that the claims made will escape any scrutiny. The claims are 'for approximate information only' and are not intended, as detailed above, to meet standards for empirical credibility, evaluation and replication in treating environments. Claims will not be deconstructed; they will be taken at face value. After all, considerations of the axioms of fundamental measurement are unlikely to resonate with formulary committee members and even media representatives.

Choose your Disease Stage

Assuming that our belief in the ICER reference case is not yet completely undermined, there is a further issue to consider: the structure of the modeled imaginary world. The model structure will determine the time spent by the hypothetical patient population in each disease stage as their assumed lifetime experience of the disease unrolls. Different model structures and assumptions regarding transition probabilities between modeled disease states will provide different estimates of time spent. This may be further augmented by assumptions as to significant adverse events within disease stages and the corrections made to capture their impact on the ordinal utility score (not possible with ordinal manifest scores). Clearly, with the range of model frameworks to choose from, ICER should be in a position to justify its decision on the number of disease states, the time spent in each state and the basis for the transition probabilities compared to other models in the literature.

DOI: https://doi.org/10.24926/iip.v11i1.2444

Choose your Costs

The final element in the construction of an imaginary incremental ordinal cost-per-QALY claim is the composition of the numerator: how has ICER defined the direct medical costs to populate the model over its hypothetical distant lifetime? Should the base case costs reflect assumed societal costs as the EQ-5D represents societal preference? Are there cost elements that have been put to one side? What assumptions have been made regarding how costs may increase over time? Obviously there is considerable flexibility which will impact the cost claims (which, of course, are discounted) and any threshold criteria in the potential multiverse of imaginary social engineering models.

Choose your Thresholds

There is an ongoing debate in the technology assessment literature over the past decade on the relevance of willingness to pay, cost-per QALY thresholds. Cleemput et al, writing in 2011, take the view that incremental cost-effectiveness ratios (ICERs) and ICER threshold values are insufficient for assessing interventions' value for money and should be considered as only one element in the decision making process, although the weight that might be placed on them is unclear [65]. Since then, a vague consensus has emerged that recommends that a mix of factors should accompany any threshold recommendations in formulary decisions. Recommendations, it should be noted, that are redundant once we accept the impossibility of constructing lifetime cost-per-QALY claims with interval properties to match to thresholds. A cost-per-QALY threshold value itself has no meaning as a 'cost' cannot be 'attached' to a manifest ordinal utility score creating QALYs. Again, thresholds assume interval utility properties.

A smorgasbord of value frameworks has been proposed, none of which has received more than 'local' approval by a professional group or consultants proposing a marketable package. ICER has proposed in its VAF that it will explore the possibility of quantifying 'broader' measures of value, postponing any decision until some consensus is achieved between ICER, stakeholders and other parties (i.e., in the fullness of time or never). Barring such a communion, the threshold based ordinal imaginary recommendations will continue to take center stage. This has led commentators to suggest that the various thresholds should be built into early modeling of the feasibility of a compound achieving an acceptable 'judgment' for cost-effectiveness. This is nonsense; after all, which model do we choose?

The 2020 ICER VAF continues to utilize in reporting on its imaginary worlds a standardized set of cost-effectiveness or willingness to pay thresholds ranging from $50,000 to $200,000 per imaginary yet logically invalid QALY 'count'. The dollar threshold is an arbitrarily assumed amount that the US community (i.e., health systems) might be willing to pay for additional QALYS across all disease states. The 'value' of a product is the relationship between the nonsensical incremental cost-per-QALY claim and the various hypothetical thresholds.

For value based pricing benchmarks ICER will continue to use the range $100,000 to $150,000 per QALY. In maintaining a common set of thresholds, the impression is given, presumably mistakenly, that these standardized thresholds will yield comparable estimates for price discounting and affordability across model options including choice of utility metric. As noted above, an imaginary cost-per-QALY claim will depend on assumptions regarding: (i) the ordinal utility metric assigned to disease stages; (ii) the number and duration of disease stages and (iii) the assumed direct medical costs for each disease stage. If a $100,000 threshold, applied to a specific model yields, for example, a recommended 25% price discount, a model that differs in any one or a combination of these assumptions will yield a different discount recommendation. The threshold, in other words, is specific to the model structure and assumptions.

The conclusion must be, given the vagaries of the ICER reference case, that thresholds are misleading (and redundant). They represent a construct based on a lifetime imaginary world that lacks scientific merit. There is a nagging feeling that what has occupied analysts in health technology assessment for over 30 years has no social value. One is reminded of Wilde's (Oscar Wilde 1854 – 1900) observation on fox hunting: *the indescribable in pursuit of the inedible*.

Needs and the Patient Voice

If health technology assessments are to become more than the construction of HRQoL imaginary worlds to support threshold claims, specific to the reference model, and which fail to meet the standards of normal science, then we have to consider alternative outcome frameworks. Central to any meaningful assessment framework is the patient voice in therapy response. This is not easily achieved. ICER's embrace of generic multi-attribute ordinal measures, even if it involves employing different HRQoL measures in the same model, ignores the QoL of patients in that disease state.

It is important to make the point that the majority of HRQoL measures were not intended, nor designed to determine the value to patients of alternative health states, both generic and disease specific. A necessary starting point is to make a distinction between patient reported outcome (PRO) measures and patient-centric outcome (PCO) measures [66] [49]. PROs encompass a range of outcomes, including clinical status, treatment satisfaction, quality of life and utility. One PRO definition, focusing on HRQoL, narrows the scope to how respondents feel and function: 'how they feel or function in relation to a health condition' [49]. This, from the patient's perspective may be irrelevant, failing to take account of other factors that may impact QoL in the lives of patients such as access to care, financial resources, education, caregiver support and even the personality of the patient.

DOI: https://doi.org/10.24926/iip.v11i1.2444

PCOs are, by definition, disease specific; patients' needs can only be evaluated with reference to their disease state (or a combination of states if comorbidities are present) [49]. The needs of patients, where rare diseases which impact the patient, caregiver and wider family are a case on point, can only be understood from the patient's perspective. The underlying needs-based construct focuses not on the measurement of symptoms and activity limitations (HRQoL) but on the impact of therapy interventions on the lives and needs of the patient [49].

Patient centric, disease specific, outcomes instruments start from a needs-based construct. The needs model hypothesizes that the *value of individual lives is dependent on the extent to which their human needs are fulfilled. Value is low when few needs are met* [49] . The major factors in needs fulfillment are the presence and treatment of disease. Clinical HRQoL ordinal manifest scores for the symptoms covered may 'improve' through treatment, without necessarily impacting patient value. The numerical value of any improvement (change in response level) is, of course, unknown.

The starting point for a needs-based instrument is in-depth interviews with patients [49]. These identify common issues and are the basis for creating an item set for application to the Rasch model. The objective is to create a unidimensional interval scale to capture the latent needs construct. The measure is an index that determines the extent to which needs are met and the impact of therapy options. PCOs are a direct reflection of the patient voice; there is no need for an artificial distinction between EQ-5D ordinal utilities to drive (unacceptable) population health focused value claims and 'issues of interest to patients'. If there are clinical attributes that are important in a disease state then these will be captured in the PCO.

Exeunt Ordinal QALYs

If the only reason for focusing on ordinal QALY scores, with the assumption that they have interval and even ratio properties, is to justify the role of imaginary worlds in value assessment, then we might well ask why? Is it because, with a perceived need to make an upfront value assessment of a product, we can put the standards of normal science to one side and convince a, possibly ill-informed audience, that this is the gold standard? Is this the only way that ICER can demonstrate the worth of its reference case?

To assume, without any justification, that any utility score created by a system such as the EQ-5D-3L on a zero to unity number line must have interval scoring properties for any target patient group of interest is absurd. These various generic systems (SF-6D, HUI Mk3, etc.) were not designed to meet RMT standards. They are a selection of symptoms with ordinal responses.

On the other hand we could admit that the incremental population-focused incremental cost per QALY reference case is an analytical dead end. ICER may claim that their model is the 'one to watch' or, rather, avoid. If so, it hardly gives decision makers a justifiable evidence base. Formulary decisions should not be based on 'approximate information' for QALYs over the next 30 years driven by dubious assumptions. In a previous commentary, following a review of the practical impact of modeled cost-per-QALY claims it was concluded that: *In retrospect, it is doubtful, that the great expectations for QALYs could ever be realized outside of reference case imaginary worlds, or the willingness of decision makers to suspend belief in the standards of normal science, and accept lifetime cost-per-QALY claims as decision criteria. Unless, therefore, a case can be made for short-term and evaluable QALY claims, there seems little scope for QALYS, and associated cost-per-QALY claims, as inputs to formulary decision making. Perhaps, as Pip says to Estella, it has been 'a vain hope and an idle pursuit'* [67]*. After over 30 years perhaps we can put QALYs to one side and return to clinically and quality specific endpoints in comparative claims for pharmaceutical products in disease and therapeutic areas* [68].

Value Claims: Meeting Acceptable Evidence Standards

The last few years have seen a number of organizations propose value claim frameworks for evaluating products. Unlike the one-size-fits all ICER reference case, these frameworks are disease specific. Most recently, the National Pharmaceutical Council (NPC) has released recommendations for value assessment frameworks [69]. Putting to one side the question of who is going to implement these recommendations, sections on (I) methodology and (ii) benefits are of interest as they point, in implicitly accepting the technology assessment meme, to the failure to recognize the importance of meeting the standards of normal science in the application of 'established' health economic methodologies and the claims made. For the NPC 'methods should be based on established health economic methodologies, consistent with established standards'. Following these standards, continue the NPC, is necessary 'to produce a meaningful and credible assessment of value'. While not mentioning the role of reference case guidelines, NPC is clearly in support of the creation of ordinal imaginary worlds and pseudoscientific opinions; claims that fail the demarcation test. Whether they recognize this is an open question. Presumably the term 'credible' in the case of the imaginary centerpiece is not meant to be interpreted as amenable to empirical assessment (e.g., hypothesis testing of a value claim).

More to the point, the NPC continues, base case assumptions (presumably the ICER reference case) 'must represent reality'; they must be 'realistic and accurate.' It is unclear how NPC would judge assumed claims for 'realty', 'realistic' or 'accurate'. The question of whether or not assumptions regarding imaginary worlds 10, 20 or 30 years into the future, the ICER reference case, can ever be supported as 'realistic' is absurd. Although the NPC seems willing to accept that ICER can represent reality 30 years ahead, even going a few years ahead raise the same objections, let alone an agreement even now that a model assumption' is realistic'. After all we do not judge

DOI: https://doi.org/10.24926/iip.v11i1.2444

models on the realism of the assumptions (Hume's induction problem). The NPC recognizes that different assumptions will yield different outcomes, or more accurately, different imaginary outcomes. But if there are different assumptions, how do we determine which assumption is 'more realistic' than another?

NPC then goes on to recommend that weights should be included in any assessment to reflect different user preferences, although how this is to be achieved or how the weights are to be assigned to the various health dimensions and preferences is unclear. Is the NPC suggesting we bundle the various piece or imaginary and other evidence into a multiple criteria decision framework? The limitations of multiple criteria decision analysis (MCDA) are well known[70]. Are we proposing, for example, to redistribute multi-attribute weights and create new ordinal utility scoring algorithms from preference scoring by different target groups? Users should be allowed, presumably within this imaginary construct, to adjust assumptions and parameters to accommodate individual preferences for different outcomes and factors, and make adjustments to represent different scenarios. A wealth or imaginary ordinal constructs. Presumably, formulary committees will be presented with these numerous imaginary worlds tailored to the needs of different patient and payer interests, whether expressed as incremental generic cost-per QALY or some other PRO construct, presented as meeting their interpretation of the standards of health technology assessment in constructing multiple imaginary worlds for target disease states. As Lear remarks to Kent, reflecting on Goneril and Reagan's filial ingratitude: 'O, that way madness lies; Let me shun that'[71]. Or, in more prosaic terms, if there are credible NPC recommendations then it needs to be clearly stated as to how these are to be implemented.

Rather than putting to one side lifetime model standards, tailored to the preferences of the interested parties, the NPC accepts the reference case lifetime perspective to support imaginary claims: *The time horizon for value should be long term, ideally lifetime …. Many of the benefits …. Show up in the longer term …. The time horizon (to capture these) …. Should be long enough to capture those benefits, ideally covering a patient's lifetime.* As detailed in this commentary, this is pseudoscience; there is no apparent appreciation by the NPC of the role of discovery, of the requirement for claims to be credible, evaluable and replicable. The question of scientific method has not even surfaced, let alone fundamental measurement and RMT. Certainly, different models that generate credible and evaluable claims can be proposed. The key point is that these models should be judged on their empirical merits; on the assessment of their claims. Not on the realism or otherwise of their preferred assumptions. Perhaps the NPC might reconsider its support for imaginary reference case worlds.

Conclusions: ICER - The Unnecessary Distraction

Media releases following the release of ICER reports focus on the recommendations made for pricing and affordability. Health system decision makers are asked to take the ICER recommendations 'at face value'. There seems little debate over the lack of scientific merit in constructing ICER imaginary worlds. Any notion of fundamental measurement to drive claims is a foreign country.

Health care decision makers deserve better. Rather than continually applying model standards for imaginary constructs that do nothing to uncover new facts, groups such as ICER (and the supporting university-based consultants) would be better advised to reconsider their role in providing support for evidence-based formulary decisions. This may not be as appealing as creating fantasy scenarios and basking in the resulting media attention, but at least it would give ICER some claim to meeting the demarcation test.

If our standard is that claims made for therapy outcomes should pass the demarcation test, then the patient-centric approach is on firm ground. Rather than reflecting the interests of professionals, clinicians and even expert panels, it is grounded in the view that QoL therapy assessments should be needs-based and disease specific. Instrument development has to meet RMT standards.

Given the arguments presented here against the ICER reference case, it is clear that from the standards of normal science the ICER evidence reports are, from the modeling perspective, a waste of time. Failing to meet fundamental measurement standards means that ICER must repudiate all previous evidence reports. Whether manufacturers and other 'stakeholders' should continue to engage in this charade has an obvious answer.

Most importantly, given the apparent lack of appreciation of its manifest failures, ICER is an unnecessary distraction. Manufacturers and patient interest groups have to divert resources to relay their concerns over the ICER clinical claims as well as modeled claims. Unfortunately, given the perceived complexities in reference case modeling, manufacturers, health system decision makers and insurers take the ICER claims at face value. This cannot continue, but probably will as there is too much vested in the ICER imaginary world business model.

The purpose of this commentary has been to point to the manifest flaws in the ICER VAF approach. Rather than talking about more complex value assessment frameworks, the introduction of qualitative consideration and even multi-criteria decision analysis (MCDA), we should step back and ask the fundamental question: if we accept the relevance of the scientific revolution of the 17th century, the invention of science, then we should require that claims for comparative effectiveness to be credible, evaluable and replicable. Continuing to take notice of ICER modeled claims, which are

DOI: https://doi.org/10.24926/iip.v11i1.2444

best seen as the analog of intelligent design seems pointless. Memes certainly have a life of their own, but there is no reason why we can't reject them (a memetic reformation). One reason ICER and ISPOR talk about the 'need' to bring in, on an ad hoc basis, 'considerations' that are relevant to patients and caregivers is that the reference case model fails to capture these in the first place. Preference based, multi-attribute preference instruments are irrelevant; we need PCOs not PROs.

We will also have to reconsider the role of QALYs. They have been subject to considerable criticism apart from their unacceptable measurement properties [72]. Taken together with their reliance on ordinal utility scores, the case can be made that it is pointless to continue to argue for generic cost-per-QALY constructs. QALYs should be abandoned. There is no evidence that the EQ-5D-3L or other utility systems have interval scoring properties. To assume that is absurd.

Utilities must be put to one side. With the focus on disease specific PCO scores as our outcome measure, translating these into utilities, which is possible, seems unnecessary [73]. The lost opportunity in abandoning the HRQoL meme is made the more pointed by the fact that these arguments were made some 15 years ago in the ISPOR house journal *Value in Health* [15 16 17]. The fact that they had to be repeated some 15 years later in the *Journal of Medical Economics* merely reinforces the point that the HRQoL meme is well entrenched in technology assessment belief systems [18 19 49]. Awkward issues of meeting RMT standards for instrument development can be thankfully put to one side; after all, truth is consensus.

Formulary decisions can equally well be based on disease specific evaluable cost-per-PCO claims in a time frame that allows these to be evaluated and reported back to formulary committees, not in a ridiculous imaginary lifetime framework. This point was made some 15 years ago in developing draft guidelines for WellPoint (now Anthem) and in the proposed Minnesota formulary guidelines released in 2016 [74 75]. If there is a PCO with overlapping items for each disease state, then we have a metric that can be used to support ongoing therapy assessments and provide comparisons between therapies in different disease states [76]. If this is accepted, then we put to one side many of the hundreds of PROs that have emerged over the past 30 years; many of them developed to support just one or two studies. If the patient voice is to resonate in formulary decision making, then it can be through the development of psychometrically sound, RMT based, PCO instruments.

In summary, if we are concerned with the discovery of 'new facts', even though these are necessarily provisional, then the ICER reference case modeling of imaginary worlds is an intellectual and analytical dead end. ICER is clinging to a meme that has been a distraction for 30 years. More to the point, it is an unnecessary distraction. It is based on the assumption that ordinal utility manifest scores have interval properties. They do not. They were never intended to have interval properties because those developing the various scales had no conception of the potential contribution of RMT. This is an oversight that should not have happened. There was ample evidence for the application, at the time of their development, of RMT to ensure interval properties.

This arguments presented here will no doubt be challenged. After all, memes have considerable staying power. Mysteries reinforce this; perhaps we should call the accepted 'transformation' of ordinal to interval scales a mystery. As Dawkins notes faith can be strong despite not being based upon evidence [39]. Solving mysteries can be inimical to 'the spread of a mind virus' and perhaps 'it is not a virtue to solve mysteries'.

If ICER wishes to persevere in this Homeric odyssey through imaginary worlds then, to extend the metaphor, before embarkation, it will need to demonstrate that the proposed utility system has interval scaling properties for the target patient population in the disease group. This would save on time and effort in building modeled cost-per-QALY claims which were then rejected by decision makers on fundamental measurement grounds.

The question we should ask is: If ICER ceased to exist, would it be missed? If the standards of normal science are applied, the answer is 'no'. Constructing a base case imaginary reference standards world, where the modeled claims are only one of a possible multiverse of competing claims, allied to a threshold willingness to pay recommendation for pricing that is specific to the assumptions driving the model, is nonsense on stilts.

Conflicts of Interest PCL is an Advisory Board Member and Consultant to the Institute for Patient Access and Affordability, a program of Patients Rising.

DOI: https://doi.org/10.24926/iip.v11i1.2444

References

[1] ICER. 2020-2023 Value Assessment Framework. 31 January 2020 https://icer-review.org/wp-content/uploads/2019/05/ICER_2020_2023_VAF_013120-1.pdf

[2] ICER. Reference Case for Economic Evaluations: Principles and Rationale Current as of July 16, 2018 http://icer-review.org/wp-content/uploads/2018/07/ICER_Reference_Case_July-2018.pdf

[3] Langley PC. Cost-Effectiveness and Formulary Evaluation: Imaginary Worlds and Entresto Claims in Heart Failure. *Inov Pharm*. 2016;7(3): No. 6 https://pubs.lib.umn.edu/index.php/innovations/article/view/449

[4] Langley PC. Multiple Sclerosis and the Comparative Value Disease Modifying Therapy Report of the Institute for Clinical and Economic Review (ICER). *Inov Pharm.* 2017;8(1): No. 12 https://pubs.lib.umn.edu/index.php/innovations/article/view/492

[5] Langley PC. Imaginary Worlds and the Institute for Clinical and Economic Review (ICER) Evidence Report: Targeted Immune Modulators for Rheumatoid Arthritis. *Inov Pharm.* 2017;8(2): No. 10. https://pubs.lib.umn.edu/index.php/innovations/article/view/515

[6] Langley PC. Rush to Judgement: Imaginary Worlds and Cost-Outcomes Claims for PCSK9 Inhibitors. *Inov Pharm.* 2017;8(2): No. 11 https://pubs.lib.umn.edu/index.php/innovations/article/view/516

[7] Langley PC. Another Imaginary World: The ICER Claims for the Long-Term Cost-Effectiveness and Pricing of Vesicular Monoamine Transporter 2 (VMAT2) Inhibitors in Tardive Dyskinesia. *Inov Pharm*. 2017;8(4): No 12 https://pubs.lib.umn.edu/index.php/innovations/article/view/927

[8] Langley PC. Another Rush to Judgement: The Imaginary Worlds of ICER and Recommendations in Duchenne Muscular Dystrophy. *InovPharm*. 2019;10(3):No. 11 https://pubs.lib.umn.edu/index.php/innovations/article/view/2152

[9] Langley PC. Yet another Ersatz World: The ICER Final Evidence Report for Additive Cardiovascular Therapies. *InovPharm*. 2019;10(4): No 22 https://pubs.lib.umn.edu/index.php/innovations/article/view/2337

[10] Langley PC. More Unnecessary Imaginary Worlds - Part 1: The Institute for Clinical and Economic Review's Evidence Report on Janus Kinase (JAK) Inhibitors in Rheumatoid Arthritis. *InovPharm*. 2020;11(1):No. 2 https://pubs.lib.umn.edu/index.php/innovations/article/view/2402

[11] Langley PC. More Unnecessary Imaginary Worlds – Part 2: The ICER Evidence Report on Modeling Oral Semaglutide for Type 2 Diabetes. *InovPharm*. 2020;11(1):No. 10 https://pubs.lib.umn.edu/index.php/innovations/article/view/2420

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

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

[14] Bond T, Fox C. Applying the Rasch Model (3rd Ed.) New York: Routledge, 2015

[15] Doward L, McKenna S. Defining patient reported outcomes. *Value Health*. 2004;7(1 Suppl 1):S4 – S8

[16] McKenna S, Doward L, Niero M et al. Development of needs-based quality of life instruments. *Value Health*. 2004;7(1 Suppl 1):S17-21

[17] 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):S22-26

[18] McKenna S, Heaney A, Wilburn J et al. Measurement of patient reported outcomes 1: The search for the Holy Grail. *J Med Econ*. 2019;22(6):516-522

DOI: https://doi.org/10.24926/iip.v11i1.2444

[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] Langley P. Modeling Imaginary Worlds: Version 4 of the AMCP Format for Formulary Submissions. *InovPharm*. 2016;7(2):Article 11 https://pubs.lib.umn.edu/index.php/innovations/article/view/434

[21] Langley P. Sunlit uplands: the genius of the NICE reference case. *Inov Pharm*. 2016;7(2): Article 12 https://pubs.lib.umn.edu/index.php/innovations/article/view/435

[22] Langley P. True North: Building Imaginary Worlds with the Revised Canadian (CADTH) Guidelines for Health Technology Assessment. *InovPharm*. 2017;8(2):Article 9 https://pubs.lib.umn.edu/index.php/innovations/article/view/514

[23] Langley P. Dreamtime: Version 5.0 of the Australian Guidelines for Preparing Submissions to the Pharmaceutical Benefits Advisory Committee (PBAC). *InovPharm.* 2017;8(1): Article. 5 https://pubs.lib.umn.edu/index.php/innovations/article/view/485

[24] Langley P. He ao pohewa: The PHARMAC Prescription for Pharmacoeconomic Analysis in New Zealand and the standards of normal science. *InovPharm*. 2016;7(2): No. 13. https://pubs.lib.umn.edu/index.php/innovations/article/view/436

[25] Langley P. Na domhain shamhlaíochta: formulary submission guidelines in Ireland and the standards of normal science. *Curr Med Res Opin*. 2016;32(5): DOI:10.1080/03007995.2016.1190699

[26] Langley P. ICER, ISPOR and QALYs: Tales of Imaginary Worlds. *InovPharm*. 2019;10(4): No. 10 https://pubs.lib.umn.edu/index.php/innovations/article/view/2266

[27] Persad G. Priority setting, cost -effectiveness and The Affordable Care Act. *Am J Law Med* . 2015;41:119-166

[28] ICER. Value Assessment Methods for "Single or Short-Term Transformative Therapies" (SSTs): Proposed adaptations to the ICER Value Assessment Framework. August 6, 2019

[29] ICER. Value Assessment Methods and Pricing Recommendations for Potential Cures: A Technical Brief. August 6, 2019

[30] Langley P. The Imaginary Worlds of Cure Proportion Modeling: Survivorship and Reference Case Pricing of Transformative Therapies. *InovPharm*. 2019;10(3): No. 17 https://pubs.lib.umn.edu/index.php/innovations/article/view/2241

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

[32] Popper KR., The logic of scientific discovery. New York: Harper, 1959.

[33] Lakatos I, Musgrave A (eds.). Criticism and the growth of knowledge. Cambridge: University Press, 1970.

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

[35] Darwin C. The Origin of Species, London 1859

[36] Wells J, Dembski W. Of Pandas and People. Foundation for Thought and Ethics, 1989

[37] Canadian Agency for Drugs and Technologies in Health (CADTH). Guidelines for the economic evaluation of health technologies: Canada. Ottawa: CADTH, 2017

[38] Dawkins R. The Selfish Gene (30th Anniversary Ed). Oxford: University Press, 2006

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

[40] Drummond M, Sculpher M, Claxton K et al. Methods for the Economic Evaluation of Health Care Programmes (4th Ed.). Oxford; Oxford University Press, 2015.

[41] National Science Board. Science and Engineering Indicators, 2018. https://www.nsf.gov/statistics/2018/nsb20181/report/sections/science-and-technology-public-attitudes-and-understanding/public-knowledge-about-s-t

[42] Yes, Flat-Earthers Really do Exist. *Scientific American*. 18 October 2018 https://blogs.scientificamerican.com/observations/yes-flat-earthers-really-do-exist/?print=true

[43] 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

[44] Magee B. Popper. London; Fontana, 1973

[45] Briggs R. The Scientific Revolution of the seventeenth century. Longman, 1971.

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

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

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

[49] McKenna S, Wilburn J. Patient value: its nature, measurement, and role in real world evidence studies and outcomes-based reimbursement. *J Med Econ*. 2018;21(5):474-80

[50] Wahlberg M, Zingmark M, Stenberg G et al. Rasch analysis of the EQ-5D-3L and the EQ-5D-5L in persons with back and neck pain receiving physiotherapy in a primary care context. *Eur J Physio*. 2019.

[51] Langley P. CVS Health and the imaginary worlds of the Institute for Clinical and Economic Review (ICER). *Inov Pharm*. 2018; 9(4):No. 4 https://pubs.lib.umn.edu/index.php/innovations/article/view/1461

[52] Gibbons C, Thornton E, Ealing J et al. Assessing social isolation in motor neurone disease: A Rasch analysis of the MND Social Withdrawal Scale. *J Neuro Sci*. 2013;334:112-118

[53] Lambert S, Pallant J, Boyes A et al. A Rasch analysis of the Hospital Anxiety and Depression Scale (HADS) among cancer survivors. *Psychol Assess*. 2013;25(2):379-90

[54] Zucca A, Lambert S, Boyes A et al. Rasch analysis of the Mini-Mental Adjustment to Cancer Scale (mini-MAC) among a heterogeneous sample of long-term cancer survivors: a cross-sectional study. *Health Qual Life Outcomes*. 2012;10:55

[55]Shea T, Tennant A, Pallant J. Rasch model analysis of the Depression, Anxiety and Stress Scale (DASS). *BMC Psychiatry*. 2009;9:21

[56] Fiorjaz M, Martinez-Martin P, Dujardin K et al. Rasch analysis of anxiety scales in Parkinson's disease. *J Psychosom Res*. 2013;74(5):414-9

[57] Galen Research, Manchester, UK http://www.galen-research.com/measures-database/

[58] Kant I, Critique of Pure Reason, trans. Norman Kemp Smith (called *Critique*). New York: St. Martin's Press, 1965.

[59] Drummond M, Sculpher M, Torrance G et al. Methods for the Economic Evaluation of Health Care Programmes. 3rd Ed. New York: Oxford University Press, 2005.

[60] ICER. Evidence Report – Oral Semaglutide for Type 2 Diabetes. 1 November 2019. https://icer-review.org/wp-content/uploads/2019/09/ICER_Diabetes_Evidence-Report_110119.pdf

[61] Brazier, 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;22:267-75

[62] Tufts University. Center for the Evaluation of Value and Risk in Health https://cevr.tuftsmedicalcenter.org/databases/cea-registry

[63] Buckley C. The Relic Disaster. New York: Simon and Schuster, 2016

[64] Wailoo A, Hernandez-Alava M, Manca A et al. Mapping to estimate health-state utility from non-preference-based outcomes measures; An ISPOR Good Practices for Outcomes Research Task Force. *Value Health*. 2017;20(1):18-27

[65] Cleemput I, Neyt M, Thiry N et al. Using threshold values for cost per quality adjusted life-year gained in healthcare decisions. *Int J Technol Assess Health Care*. 2011;27(1):71-6

[66] McKenna S, The limitations of patient reported outcome measurement in oncology. *J Clin Pathways*. 2016;2(7):37-46

[67] Dickens C. Great Expectations. London, 1861

[68] Langley PC. Great Expectations: Cost-utility models as decision criteria. *Inov Pharm*. 2016:7(2); No. 14 https://pubs.lib.umn.edu/index.php/innovations/article/view/437

[69] National Pharmaceutical Council. Guiding Practices for Patient Centered Value Assessments. June 2019 https://www.npcnow.org/guidingpractices

[70] Baltussen R, March K, Thokala P et al. Multicriteria Decision Analysis to support health technology assessment agencies: Benefits, limitations and the way forward. *Value Health*. 2019;22(11):1283-88

[71] Shakespeare W. *King Lear*. Act III, Scene IV (folio Ed.) Wells S, Taylor G. (Eds.). The Oxford Shakespeare (2nd Ed.). Oxford: University Press, 2005

[72] Pettit D, Raza S, Naughton B et al. The Limitations of QALY: A literature review. *J Stem Cell Res Ther*.2016;6:4

[73] McKenna S, Ratcliffe J, Meads D. Development and validation of a preference based measure derived from the Cambridge Pulmonary Hypertension Outcome Review (CAMPHOR) for use in cost utility analyses. *Health Qual Life Outcomes*. 2008;6:65

[74] Langley PC, Recent Developments in the Health Technology Assessment Process in Fulda T and Wertheimer A. *Handbook of Pharmaceutical Public Policy.* New York, Haworth Press, 2007, pp. 457-477

[75] Langley PC. Guidelines for Formulary Evaluation [Proposed]. Program in Social and Administrative Pharmacy. College of Pharmacy. University of Minnesota. Version 2.0. December 2016

[76] Rouse M, Twiss J, McKenna SP. Co-calibrating quality-of-life scores from three pulmonary disorders: implications for comparative-effectiveness research. *J Med Econ*. 2016;19(6):596-603

DOI: https://doi.org/10.24926/iip.v11i1.2444