MAIMON RESEARCH LLC

RECONSTRUCTION OF HEALTH TECHNOLOGY ASSESSMENT



BLISSFUL IGNORANCE: THE ISPOR COMMITMENT TO MEASUREMENT FAILURE

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

University of Minnesota, Minneapolis, MN

LOGIT WORKING PAPER No 2049 NOVEMBER 2026

www.maimonresearch.com

Tucson AZ

ABSTRACT

Background: *The scientific legitimacy of health technology assessment (HTA) depends upon the measurement properties of the quantities used in quantitative evaluation. ISPOR occupies a globally influential position in establishing professional standards, methodological recommendations and educational expectations for health economics and outcomes research. Previous interrogations of ISPOR knowledge bases have identified measurement inversion as a critical failure. This paper extends that critique by examining whether the principal constructs endorsed within legacy HTA satisfy the axioms of representational measurement.*

Methods: *ISPOR's methodological framework is assessed against the requirements of representational measurement, including attribute definition, unidimensionality, empirical representation, uniqueness, admissible transformations, dimensional homogeneity and commensurability. The analysis examines health-state valuation, time trade-off (TTO), regression-based utility algorithms, decrements from unity, quality-adjusted life years (QALYs), reference-case simulation, incremental cost-effectiveness ratios (ICERs), patient-reported outcomes, mapping and university curricula. Particular attention is given to the distinction between lawful duration ratios and unestablished preference measures.*

Findings: *The central measurement failure occurs before QALY construction. TTO generates a lawful duration ratio, t/T, but its reinterpretation as a preference ratio lacks an independent demonstration of the required measurement axioms. Regression algorithms subsequently estimate coefficients associated with multidimensional health-state descriptions and subtract decrements from a conventionally assigned unity for full health. This arithmetic produces numerical valuation scores without establishing that they measure a defined, unidimensional utility attribute possessing ratio-scale properties. Negative scores further expose the absence of a defensible proportional interpretation. Multiplication of these scores by survival duration constructs QALYs without establishing the measurement legitimacy of the resulting quantities. Reference-case simulations propagate these unestablished constructs into ICERs and cost-per-QALY thresholds. Statistical sophistication, reproducibility and professional consensus cannot substitute for empirical measurement. ISPOR's endorsement of these procedures contributes to university curricula in which arithmetic has abandoned measurement.*

Conclusions: *Legacy HTA represents an institutionalized system of numerical storytelling in which quantitative operations are accepted without demonstrating the measurement properties required for their scientific interpretation. Marginal methodological refinement cannot remedy this foundational failure. Reconstruction requires curriculum replacement rather than supplementation, beginning with lawful measures of defined attributes and proceeding to empirically testable, reproducible and falsifiable therapy-impact claims. The existing nine-unit Maimon Research HTA Reconstruction Program provides a practical route for faculty development, graduate education and institutional transition. Reconstruction need not await ISPOR's approval. Universities and research institutions can initiate the replacement of arithmetic without measurement by an empirical science founded on representational measurement.*

INTRODUCTION

The purpose of this paper is to evaluate key aspects of ISPOR's continuing commitment to the false methodological standards of legacy health technology assessment (HTA). There is now substantial evidence, including some 17 previous interrogations of the ISPOR HTA knowledge bases, that measurement inversion is established within legacy HTA [1] . These findings point to a systematic failure to recognize representational measurement as the necessary foundation for quantitative assessment. What remains insufficiently examined is the scientific legitimacy of the principal analytical constructs that ISPOR teaches, recommends and endorses when judged against the standards for quantitative measurement and the axioms of representational measurement.

The issue is fundamental. Scientific measurement requires that quantitative assessment begin with the specification of a defined, unidimensional attribute, followed by the demonstration of its measurement properties. The distinction between nominal, ordinal, interval and ratio scales determines the arithmetic that can legitimately be performed. Equal units, meaningful absolute zeros, admissible transformations, dimensional homogeneity and commensurability are not optional refinements. They establish whether numerical operations preserve the empirical meaning of the quantities involved. Without these foundations, mathematical calculations may generate numerical results but cannot establish that those results represent measurable empirical relationships.

ISPOR occupies a particularly influential and global position in contemporary HTA. Through its Good Practices for Outcomes Research reports, methodological task forces, competency frameworks, professional education and scientific programs, it helps define the analytical standards adopted by universities, researchers, pharmaceutical manufacturers, consulting organizations and reimbursement agencies. Its influence extends well beyond its membership. The methods it promotes contribute to the international framework within which evidence is generated, economic evaluations are constructed and healthcare resource-allocation decisions are justified.

This influence carries a corresponding scientific responsibility. The central question is not whether ISPOR's recommended procedures are widely accepted. It is whether the quantities entering those procedures possess the measurement properties necessary to support the arithmetic performed upon them. Professional consensus cannot substitute for measurement axioms, and methodological standardization cannot transform an arbitrary numerical assignment into a lawful measure.

The present assessment moves beyond identifying the absence of measurement foundations in educational and methodological content. It examines the consequences of that absence for the principal constructs of legacy HTA: health-status valuation, time trade-off, utility algorithms, QALYs, cost-effectiveness ratios, reference-case simulation, mapping and patient-reported outcome measurement. Particular attention is given to the distinction between manifest and latent attributes, the requirements for lawful ratio measurement, and the attempted transfer of numerical coefficients between preference space and time space.

The analysis also considers ISPOR's relationship with university HTA curricula. Where professional guidance endorses quantitative procedures without requiring their measurement properties to be established, universities risk reproducing the same methodological deficiencies in

successive generations of students. The resulting educational framework emphasizes statistical estimation, economic modelling and numerical decision rules while neglecting the scientific conditions under which quantitative claims can be considered legitimate.

The central proposition of this paper is that ISPOR in promoting legacy HTA has institutionalized arithmetic while abandoning measurement [2] . Utility scores are generated without demonstrating that they represent lawful ratio measures; QALYs are constructed by applying these scores to survival duration without establishing the necessary empirical mapping; and cost-effectiveness claims are derived from modelled quantities whose measurement properties remain unproven. The resulting calculations may be reproducible within their assumed models, but reproducibility of arithmetic is not evidence of scientific measurement. All that is produced is numerical storytelling.

The purpose is not simply to identify deficiencies in individual methods. It is to determine whether ISPOR's recommended analytical framework satisfies the foundational requirements for meaningful quantitative assessment. Where those requirements are absent, the appropriate response is not further refinement of conventional modelling procedures but reconstruction from the principles of representational measurement. The distinction is between producing numbers and measuring attributes. Until that distinction is recognized, legacy HTA remains vulnerable to the charge that its principal quantitative claims constitute numerical storytelling rather than empirically defensible science.

UNDERSTANDING REPRESENTATIONAL MEASUREMENT

The modern classification of measurement scales was established by Stevens (1946), who distinguished four levels of measurement: nominal, ordinal, interval and ratio [3]. His contribution was fundamental because it demonstrated that assigning numbers to observations does not necessarily constitute quantitative measurement. Each scale possesses distinct mathematical properties, defined by the empirical relationships represented and the transformations that preserve those relationships. These properties determine which arithmetic operations are legitimate. The distinction between numbers as labels and numbers as measures is therefore foundational to scientific assessment [4] .

Nominal scales classify observations without establishing magnitude or order. Ordinal scales establish an ordering but do not demonstrate equal intervals between successive positions. Interval scales possess equal units and support meaningful differences, but lack an absolute zero. Their admissible transformations are:

$x' = a \times (x + b)$, where $a > 0$.

Consequently, ratios of interval-scale values are not invariant. Ratio scales possess equal units and a meaningful absolute zero, with admissible transformations restricted to:

$x' = a \times x$, where $a > 0$.

Ratios of measured magnitudes are therefore preserved. This distinction is critical because proportional comparisons and multiplication or division intended to represent empirical relationships require measurement properties appropriate to those operations.

From Helmholtz (1887) [5] and Hölder (1901) [6], through Campbell (1920, 1928) [7] [8], Stevens (1946), Suppes and Zinnes (1963) [9], Luce and Tukey (1964) [10], Krantz and colleagues (1971, 1989, 1990) [11] [12] [13], and Rasch (1960/1980) [14], the foundations of quantitative measurement have been progressively developed through the specification of empirical attributes, axiomatic representation, uniqueness, admissible transformations and invariance. Numbers legitimately represent an attribute only when the relationships among empirical observations can be mapped into a numerical system while preserving the relevant relationships.

The axioms of representational measurement specify the conditions under which such mappings are possible. Depending upon the attribute and proposed measurement structure, these include ordering, transitivity, consistency, solvability, Archimedean conditions, independence and appropriate forms of cancellation. These requirements are not statistical conventions. They establish whether an empirical attribute admits a numerical representation and which transformations and operations preserve its meaning. A measurement scale cannot acquire properties merely because an analyst assigns numerical values or applies a statistical model.

Quantitative assessment must therefore begin with the identification of a defined, unidimensional attribute. The analyst must determine whether the attribute is manifest and directly observable or latent and inferred through an appropriate measurement model. The required measurement properties must then be established, including the scale type, meaningful units, invariance and, where ratio measurement is claimed, an absolute zero. For latent attributes, Rasch measurement provides the framework for testing unidimensionality, specific objectivity and person–item conjoint measurement, although the logit location scale itself has an arbitrary origin.

Two additional requirements govern subsequent arithmetic. Dimensional homogeneity requires mathematical operations to preserve the dimensional meaning of the quantities involved. Commensurability requires quantities entering a ratio intended to express a proportion to represent the same attribute. A dimensionless numerical result does not, by itself, establish that the ratio is empirically meaningful.

These principles are essential to health technology assessment because therapy-impact claims are quantitative scientific claims. Before scores can be combined, durations adjusted, outcomes aggregated or cost-effectiveness ratios calculated, the measurement status of every relevant variable must be established. Statistical significance, reliability, predictive validity and model sophistication cannot compensate for the absence of lawful measurement.

Representational measurement is therefore not an optional addition to HTA methodology. It is the scientific foundation that determines whether quantitative assessment is possible. Without it, arithmetic precedes measurement, numerical assignments are mistaken for empirical magnitudes, and increasingly elaborate calculations risk becoming numerical storytelling rather than falsifiable scientific claims.

THE IMPORTANCE OF ISPOR

ISPOR occupies a central position in the international development, dissemination and institutionalization of health economics and outcomes research (HEOR) and health technology assessment (HTA). As a global professional organization, it influences the methodological standards applied by academic institutions, pharmaceutical manufacturers, consulting organizations, regulatory and reimbursement agencies, and healthcare decision-makers. Its importance extends beyond providing a forum for scientific exchange. Through its Good Practices for Outcomes Research reports, methodological task forces, professional competencies, educational programs and scientific conferences, ISPOR helps determine which analytical approaches are recognized as legitimate and appropriate for contemporary HTA.

This influence is particularly evident in the continued acceptance of conventional cost-utility analysis. Health-state valuation, time trade-off, standard gamble, preference-based utility algorithms, quality-adjusted life years (QALYs), cost-effectiveness ratios and reference-case economic modelling are embedded in the methodological framework that ISPOR teaches and promotes. These procedures are presented as established instruments of quantitative assessment, supported by extensive guidance on their construction, statistical evaluation and application. Their widespread adoption has created an international professional consensus around a common set of analytical practices.

Yet professional consensus is not a substitute for scientific measurement. The legitimacy of quantitative assessment depends upon the properties of the attributes being measured and the admissibility of the arithmetic performed upon them. ISPOR's influence therefore creates a fundamental responsibility: to ensure that the methods it endorses satisfy the axioms of representational measurement.

The distinction is critical. ISPOR's recommendations frequently begin with numerical scores, statistical estimates or model inputs whose measurement properties have not been established. Utility algorithms generate coefficients that are treated as measures of health-state preference without demonstrating the requirements of unidimensionality, equal units, absolute zero and lawful ratio measurement. These coefficients are subsequently multiplied by survival duration to construct QALYs, despite the absence of an empirically established mapping between preference space and time space. The resulting quantities become denominators in incremental cost-effectiveness ratios and inputs to resource-allocation decisions. At each stage, the acceptance of conventional arithmetic substitutes for the prior demonstration of measurement legitimacy. This is unacceptable in a subject that claims professional status. ISPOR should not be promoting numerical storytelling.

ISPOR's importance is reinforced by its relationship with university education. Academic programs in health economics, pharmacoeconomics, outcomes research and HTA draw upon the same methodological literature, professional guidance and analytical conventions. Students are trained to construct economic models, estimate utilities, calculate QALYs and interpret cost-effectiveness thresholds. Where representational measurement is absent from the professional standards informing these programs, its absence risks being reproduced in university curricula. Successive generations of graduates may therefore acquire proficiency in numerical analysis

without understanding the axiomatic requirements that distinguish lawful measurement from numerical assignment. Again, university HTA courses should not be training graduate students in numerical storytelling.

This problem is compounded by the distinction between manifest and latent attributes. Manifest outcomes such as survival duration and resource quantities require appropriate linear ratio measurement. Latent outcomes such as pain, fatigue and physical functioning require measurement models capable of establishing unidimensionality, invariance and defensible quantitative representations. The role of Rasch measurement in constructing and testing latent measures is consequently fundamental. Yet these distinctions are not established as prerequisites throughout ISPOR's conventional methodological framework.

Previous assessments of measurement inversion have raised questions about the extent to which representational measurement is recognized by ISPOR within contemporary HTA [15]. The significance of these findings is heightened by ISPOR's institutional reach. A deficiency in the methodological standards promoted by a major professional organization is not confined to its publications. It can influence adversely educational priorities, research practices, modelling conventions and the criteria by which quantitative evidence is judged.

ISPOR is therefore an appropriate and consequential subject for a detailed deconstruction of legacy HTA. Its importance lies not simply in the number of methods it endorses but in the authority those endorsements confer. When professional standards permit arithmetic without requiring the underlying measurement properties to be demonstrated, numerical procedures acquire legitimacy through institutional acceptance rather than empirical justification.

ISPOR’s endorsement of numerical storytelling becomes a global epistemic phenomenon, sustained by an institutionalized epistemic memeplex in which mutually reinforcing assumptions, professional standards, methodological conventions and educational practices shelter false measurement from scientific scrutiny [16] [17] [18] . Within this closed system, the legitimacy of numerical constructs is established through professional consensus and repeated application rather than demonstration against the axioms of representational measurement. Utility scores, QALYs, cost-effectiveness ratios and reference-case simulations acquire authority because they are endorsed, taught, published and reproduced across successive generations of researchers and practitioners. Challenges to their measurement foundations are marginalized because they question the legitimacy of the framework itself. The result is a self-reinforcing system in which institutional acceptance substitutes for empirical justification, protecting numerical storytelling from the scientific standards it should be required to satisfy.

The central issue is whether ISPOR's leadership recognizes it has enforced a system of false quantitative assessment in which the production of numbers is mistaken for the measurement of attributes. If the principal constructs of legacy HTA cannot satisfy the requirements of representational measurement, their continued endorsement cannot be justified by professional convention, statistical refinement or widespread adoption. ISPOR's international influence makes this failure particularly consequential and the reconstruction of HTA around lawful measurement correspondingly urgent.

ISPOR AND HEALTH-STATE VALUATIONS: THE ABANDONMENT OF MEASUREMENT FOUNDATIONS

The first and most substantive objection to ISPOR's advocacy of legacy health technology assessment is its endorsement of a framework founded upon health-state valuations rather than the measurement of defined attributes. This approach effectively sets aside decades of established measurement theory concerning the relationship between measurement scales, admissible transformations and the arithmetic operations that preserve empirical meaning as evidenced by Stevens (1946) contribution.

There is no comparable theory of admissible arithmetic within conventional health-state valuation. Numerical utility scores are generated through time trade-off, standard gamble, discrete-choice experiments and regression-based valuation algorithms, but no corresponding classification establishes which mathematical operations are justified by their empirical measurement properties. Scores are subtracted, multiplied, divided and aggregated because these operations are required by economic evaluation, not because an axiomatic measurement structure has demonstrated their legitimacy.

The omission extends beyond Stevens' classification. The development of representational measurement, established the axiomatic foundations required to represent empirical attributes numerically. These foundations concern the existence of a numerical representation, the preservation of empirical relationships and the conditions under which the resulting scale is unique up to specified admissible transformations. Measurement is not established merely by generating a numerical score. It requires an empirical relational system whose structure can be represented within an appropriate numerical system.

The distinction between representation and uniqueness is especially important. Representation establishes whether the empirical relationships defining an attribute can be mapped into numbers while preserving those relationships. Uniqueness establishes which transformations of the numerical representation remain admissible without changing its empirical meaning. Together, these determine the measurement scale and the arithmetic that can be justified.

The axiomatic approach also requires explicit specification of the attribute being measured. Depending upon the proposed measurement structure, relevant conditions may include ordering, transitivity, solvability, Archimedean properties, independence and cancellation. These are empirical requirements that must be examined rather than assumed. Their purpose is to establish whether a numerical representation exists, whether its properties are invariant and whether the proposed arithmetic preserves the underlying empirical relationships.

Conventional health-state valuation provides no equivalent demonstration. Instruments such as the EQ-5D describe combinations of mobility, self-care, usual activities, pain or discomfort, and anxiety or depression. These are distinct attributes assembled into multidimensional descriptions. Respondents are subsequently asked to express preferences between health-state descriptions, sometimes combined with survival duration. The resulting observations are used to construct numerical valuation algorithms. Yet neither the multidimensional description nor the elicited preference establishes a single quantitative attribute with demonstrated interval or ratio properties.

The problem is not resolved by anchoring utility scores at zero for dead and unity for full health. These are conventions defining the numerical valuation system. They do not demonstrate that zero represents the empirical absence of a unidimensional preference attribute, that successive units are equal, or that a score of 0.6 represents 60% of a measurable preference magnitude. Nor does the statistical estimation of valuation coefficients establish the uniqueness properties required for meaningful proportional comparisons.

ISPOR's Good Practices recommendations continue this tradition by emphasizing valuation procedures, statistical estimation, validity, reliability and the application of utility scores in economic models [19] [20] [21] . The common failure across all three ISPOR Good Practices reports is the absence of an explicit axiomatic foundation for measurement. They emphasize generating, validating, mapping and applying utility scores without first demonstrating that these scores represent a defined, unidimensional attribute, satisfy the representation and uniqueness requirements of measurement theory, or possess the scale properties necessary to justify the arithmetic.

The fundamental objection is therefore not that health-state preferences cannot be elicited or numerically represented. It is that numerical representation is being treated as quantitative measurement without demonstrating the empirical axioms that would justify that interpretation. A preference ordering may be established, but an ordering is not an interval measure. An interval measure, even if demonstrated, is not a ratio measure. And a ratio measure of preference, even if established, would not automatically justify its use as a proportional adjustment to survival duration.

ISPOR's endorsement of health-state valuation therefore represents a departure from the foundations of scientific measurement. It accepts numerical operations without requiring the empirical structure that would make those operations meaningful. There is no demonstrated counterpart within legacy HTA to Stevens' scale classification, the axiomatic representation requirements of Krantz and colleagues, or the uniqueness conditions governing admissible transformations.

ISPOR AND THE TIME TRADE-OFF: THE UNJUSTIFIED TRANSFER FROM TIME TO PREFERENCE

The fundamental problem with the time trade-off (TTO) is not the arithmetic used to identify an indifference point. It is the interpretation subsequently imposed upon that point. The TTO procedure identifies a relationship between two survival durations and then treats the resulting numerical ratio as though it were a measure of preference for a multidimensional health-state description [22] . This transition from time space to preference space is the critical step. It is precisely where the requirements of representational measurement should be demonstrated, yet conventional HTA accepts the reinterpretation without establishing the necessary measurement foundations.

The conventional TTO asks a respondent to consider two alternatives: living for a specified duration, T, in a described health state, or living for a shorter duration, t, in full health. The duration t is adjusted until the respondent expresses indifference between the alternatives. At this

equilibrium point, the respondent regards the two specified combinations of health status and survival duration as equally preferable.

The numerical result is expressed as:

h = t / T

Both t and T are durations. They belong to time space, where duration is a manifest attribute measured on a linear ratio scale. The ratio is dimensionally homogeneous and commensurable because both quantities represent the same attribute. Their dimensions cancel:

[Time] / [Time] = 1 (dimensionless)

For example, suppose a respondent expresses indifference between ten years in a specified health state and six years in full health. The resulting ratio is:

h = 6 / 10 = 0.6

This calculation is legitimate. It establishes that the shorter duration is 60% of the longer duration. It also identifies the position of indifference within the particular choice exercise. What it does not establish is that the respondent possesses 60% of a measurable preference attribute.

The indifference point is a relationship between two alternatives, each combining a health-state description and a duration. It is not an observation of a preference magnitude independently measured on a preference scale. The numerical ratio is constructed entirely from durations. Its ratio properties are inherited from the measurement of time, not from the respondent's preference.

The decisive error occurs when conventional HTA reinterprets this result as a health-state utility. The duration ratio is effectively relabeled as a preference ratio:

h = t / T = 0.6

is treated as equivalent to:

r = p / P = 0.6

Here, p and P would have to represent two independently established ratio measures of the same unidimensional preference attribute, with P specifying a reference magnitude. The dimensional expression would be:

[Preference] / [Preference] = 1 (dimensionless)

However, neither p nor P is measured by the TTO procedure. No preference measurement scale is independently constructed, and no demonstration establishes that the numerical relationship between preferences possesses the same properties as the duration ratio. The fact that both expressions are dimensionless does not make them equivalent.

This is not a minor technical objection. It concerns the identity of the attribute being measured. Time and preference are distinct empirical domains. The properties of one cannot be transferred to the other through a change of terminology. A coefficient of 0.6 in time space remains a duration proportion unless an independently justified mapping establishes its relationship to a corresponding proportion in preference space.

Such a mapping would require several demonstrations. First, the proposed preference attribute must be explicitly defined and shown to be unidimensional. Second, its empirical relationships must satisfy the axioms required for ratio measurement, including a meaningful absolute zero, equal invariant units, meaningful proportional comparisons and admissible similarity transformations. Third, the proposed correspondence between the preference ratio and the duration ratio must be independently established and empirically testable.

The TTO indifference response does not provide these demonstrations**.** It establishes a choice relationship under specified conditions. Utility-theoretic assumptions may provide a model for representing those choices, but they do not by themselves establish a ratio measure of a preference attribute in the representational-measurement sense.

The original TTO work of Torrance and colleagues did not establish an independent ratio measurement structure for health-state preferences. Torrance's subsequent consideration of conditional interval utility likewise did not demonstrate that the observed duration ratio could be interpreted as a ratio of two measured preference magnitudes [23] . An interval-scale interpretation would, in any event, be insufficient to justify proportional comparisons of preference scores.

ISPOR's continued endorsement of TTO-based valuation perpetuates this unresolved distinction. Guidance on elicitation procedures, study design, statistical analysis and valuation algorithms addresses how numerical scores should be generated and used. It does not establish the axiomatic justification for converting a lawful duration ratio into a lawful preference ratio.

The problem is subsequently compounded when TTO-derived utility scores are assigned to health-state descriptions, incorporated into regression algorithms and applied as proportional adjustments to survival duration in QALY calculations. At that stage, a coefficient originally constructed from time is first interpreted as preference and then returned to time space as a supposed duration-adjustment factor. Neither transition has been independently justified through representational measurement.

The central measurement error in the TTO is the assumption that the properties of a lawful duration ratio can be transferred to preference merely by relabeling the numerical result. The ratio $\mathbf{h = t / T}$ is legitimate because its numerator and denominator are independently measured durations possessing ratio-scale properties. However, interpreting the same numerical value as $\mathbf{p / P}$, a proportion of preference, introduces an entirely different empirical attribute. The measurement axioms supporting the duration ratio do not establish the corresponding axioms for preference. Even if the relabeling were accepted as a proposed interpretation, it would remain necessary to demonstrate independently that preference is a defined, unidimensional attribute possessing a meaningful zero, invariant units and a quantitative structure supporting ratio comparisons. No such properties follow from the respondent's TTO indifference judgment. The TTO therefore cannot

manufacture a preference ratio by transferring the measurement properties of time to an unmeasured preference attribute.

The conclusion is unavoidable. Although **h = t / T** is a lawful ratio of measured durations, its. The TTO process therefore reaches a measurement dead end: it produces a lawful ratio of time but fails to establish the preference measure that subsequent economic evaluation in legacy HTA requires.

ISPOR AND ALGORITHMS WITH DECREMENTS

The measurement failure underlying health-state utilities begins before the final utility score is calculated. It arises in the statistical algorithms used to transform responses to health-state valuation exercises into numerical coefficients and, subsequently, decrements from full health. ISPOR's methodological recommendations address the estimation, selection, statistical performance and application of these algorithms, but the prior question is whether their inputs and outputs possess the measurement properties required for the arithmetic performed.

In the EQ-5D tradition, health states are described through combinations of levels across distinct attributes, including mobility, self-care, usual activities, pain or discomfort, and anxiety or depression. Responses obtained through time trade-off, standard gamble or other valuation procedures are used to estimate statistical models relating these health-state descriptions to numerical valuation scores. The resulting coefficients are then incorporated into algorithms for assigning values to health states.

The first question concerns the dependent variable. What attribute is being measured by the numerical responses entered into the regression? In the case of TTO valuation, the underlying observation is an indifference judgment involving alternative durations and health states. The resulting duration ratio, **h = t / T**, is lawful within time space. Its interpretation as a measure of preference, however, requires an independent demonstration that the proposed preference attribute satisfies the axioms of representational measurement. Regression analysis cannot supply that missing demonstration. It operates on the numerical values provided to it; it does not establish that those values measure the attribute they are claimed to represent.

The second question concerns the regression coefficients. Are these coefficients themselves ratio measures? The answer is not established by their statistical estimation. A regression coefficient expresses an estimated relationship between the dependent variable and a specified predictor, conditional on the model. For a binary health-state indicator, its coefficient represents a fitted difference in the dependent-variable score associated with that indicator, holding other model terms constant. Its measurement interpretation therefore depends on the measurement status of the dependent variable and the model specification. A statistically significant coefficient, a good model fit or a reproducible prediction does not establish an independent ratio measure of utility or preference.

The third question concerns arithmetic. In additive tariff algorithms, coefficients associated with reported health problems are combined to estimate a decrement from the full-health reference value. The resulting health-state score may be expressed as:

$$U = 1 - (d_1 + d_2 + ... + d_n)$$

where the d terms represent estimated decrements associated with the relevant health-state characteristics. Some algorithms include interaction terms, constants or alternative functional specifications, but the measurement question remains unchanged.

Subtraction requires the quantities entering the operation to possess a compatible quantitative interpretation. The value of unity is assigned by convention to full health. The estimated decrements are derived from a regression on numerical valuation responses. What is the scientific justification for claiming that subtracting regression-estimated decrements from an arbitrarily assigned value of unity produces a measure of utility? The calculation generates a numerical health-state valuation score, but it does not establish that the score represents a defined, unidimensional attribute with a meaningful zero, invariant units and the ratio-scale properties required for proportional interpretation.

The problem becomes particularly visible when accumulated decrements exceed unity, producing negative health-state scores. These values are not merely inconvenient outcomes of a statistical procedure. They expose the absence of a demonstrated ratio-scale interpretation in which zero represents the absence of a defined attribute and numerical magnitudes support meaningful proportional comparisons. Changing the lower bound, rescaling the scores or reconsidering the death anchor may alter the numerical results, but none of these operations independently establishes the missing measurement structure.

The distinction between statistical estimation and scientific measurement is decisive. A regression algorithm can estimate coefficients, predict responses and reproduce patterns in valuation data. It cannot, by those achievements alone, establish the representational measurement properties of the dependent variable or transform its predictions into lawful ratio measures. Nor does the conventional subtraction of estimated decrements from unity demonstrate that the resulting health-state scores possess the properties required for proportional arithmetic.

The algorithm therefore does not resolve the measurement problem; it carries the unestablished measurement status of the original valuation responses into a new numerical scoring system. Before these scores can be treated as utilities, their attribute, measurement structure and admissible arithmetic must be demonstrated independently. The foundational measurement failure in health-state utility valuation occurs before the construction of the QALY. It arises when statistically estimated decrements associated with multidimensional health-state descriptions are subtracted from a conventionally assigned value of unity for full health. The resulting numerical score is designated a utility, but no empirical demonstration establishes that it measures a defined, unidimensional attribute possessing ratio-scale properties. The regression coefficients are estimates derived from valuation responses, not independently established magnitudes of utility. Their subtraction from unity generates a numerical index; it does not establish a meaningful zero, invariant units or the legitimacy of proportional comparisons. A score of 0.6 cannot therefore be interpreted as 60% of a measured utility attribute merely because it lies between zero and one. Negative scores, produced when decrements exceed unity, make the absence of a defensible proportional interpretation particularly apparent.

ISPOR AND UTILITIES: ARITHMETIC WITHOUT MEASUREMENT

ISPOR's Good Practices recommendations provide extensive instructions for creating, estimating, selecting and assessing health-state utility scores for economic evaluation. As noted, reports by Wolowacz et al. (2016), Wailoo et al. (2017) and Brazier et al. (2019) address preference elicitation, statistical mapping, utility estimation, model selection, validity, uncertainty and the incorporation of utility scores into cost-effectiveness models. What is conspicuously absent is a prior requirement to establish whether these numerical scores possess the measurement properties necessary for their intended arithmetic.

The conventional starting point is a utility scale anchored at 1 for full health and 0 for dead, with values assigned to health-state descriptions through time trade-off, standard gamble or preference-based valuation algorithms. In the EQ-5D tradition, regression algorithms estimate coefficients associated with different levels of reported health problems. These coefficients are commonly expressed as decrements from the full-health anchor of 1, producing numerical scores that are subsequently treated as utilities and which can yield negative values or so-called states worse than death. Yet subtracting statistically estimated decrements from unity does not establish that the resulting values measure a single utility or preference attribute. Nor does it demonstrate equal units, a meaningful zero, admissible transformations or the ratio properties required for proportional arithmetic.

The difficulties created by negative health-state utility scores have prompted attempts to reconsider the conventional anchoring of utility at zero for death and unity for full health. Sampson, Parkin and Devlin (2024) acknowledge that negative health-state values create fundamental measurement problems and question whether death should remain the zero anchor [24]. Their proposed solution is to reverse the conventional construction of the QALY. Rather than deriving QALYs from health-state utility scores, they propose treating the QALY as the foundational unit of measurement, with health-state values derived from quantities of QALYs. They introduce the *value of a healthy year* (VHY) to describe the proposed quantity.

This proposal illustrates the consequences of attempting to resolve measurement failures without first defining the attribute being measured. One year in full health is assigned a value of one QALY, but no independent measurement procedure establishes how one year in an impaired health state becomes 0.6 or 0.4 QALYs. The proposed reversal assumes the existence of the very quantitative relationship that requires demonstration. Nor does redefining a negative utility as a shortfall exceeding one VHY establish an empirical measurement structure supporting that shortfall. It merely provides another interpretation of a numerical decrement.

The fundamental problem is therefore not the choice of death as an anchor. It is the absence of a defined, unidimensional utility attribute possessing demonstrated measurement properties. Removing death from the valuation scale does not establish equal units, a meaningful zero, admissible transformations or the ratio properties required for proportional arithmetic. Neither does renaming the QALY as a VHY demonstrate a lawful mapping between preference and survival duration.

ISPOR does not establish a definition of what a utility score must be as a scientific measure. A utility score intended to support proportional adjustment must represent a defined, unidimensional attribute and satisfy the relevant axioms of representational measurement. These requirements begin with unidimensionality, followed by an empirically established ordering of magnitudes, equal and invariant measurement units, an absolute and non-arbitrary zero representing the absence of the attribute, and a measurement structure supporting meaningful ratios. The empirical relations must be preserved by the numerical representation, with admissible transformations restricted to positive similarity transformations:

$u' = a \times u$, where $a > 0$

The measurement procedure must also demonstrate invariance and reproducibility under its specified conditions. Assigning zero to death and unity to full health by convention does not establish these properties. The endpoints of an algorithmic scoring system are not substitutes for an empirically demonstrated ratio scale.

The requirement becomes decisive when utility scores are interpreted as proportional adjustments. Consider a utility coefficient of 0.6. For this to represent a lawful proportion within utility space, it must first be expressible as:

$c = u / U = 0.6$

Here, u and U must be ratio measures of the same defined utility attribute, obtained on a common scale, with U representing the specified reference magnitude. Dimensional homogeneity requires compatible dimensions, while commensurability requires that both quantities measure the same attribute. The dimensional expression is:

[Utility] / [Utility] = 1 (dimensionless)

The dimensions cancel to produce a dimensionless coefficient. However, this coefficient has been established exclusively within utility space. Its numerical value expresses a relationship between two magnitudes of the same utility attribute. It has no demonstrated interpretation as a proportion of survival duration. Dimensional cancellation does not eliminate the empirical identity of the attribute from which the coefficient was constructed.

Time, by contrast, occupies its own measurement space. Survival duration is a manifest attribute measured on a linear ratio scale, with a meaningful zero, equal units and admissible proportional comparisons. A duration ratio is constructed entirely from measures of time:

$r = t1 / t2$

Here, t1 and t2 are independently measured durations. Its dimensional expression is:

[Time] / [Time] = 1 (dimensionless)

This ratio is also dimensionless, but it is a proportion defined in time space, not utility space. The fact that u / U and t1 / t2 are both dimensionless does not establish that they measure the same relationship or that their numerical values are interchangeable.

A utility proportion of 0.6 is not a duration proportion of 0.6. Each belongs to its own empirically defined measurement structure.

Any proposed correspondence between utility proportions and duration proportions would require a separate empirical mapping:

u / U = 0.6 in utility space

t1 / t2 = 0.6 in time space

These are two distinct relationships. Establishing the first does not establish the second. To claim that the same numerical proportion applies in both spaces, an independently justified and empirically testable mapping must be demonstrated. Numerical equality alone cannot provide that justification.

This exposes the foundational problem with utility algorithms. An algorithm can assign numbers to multidimensional health-state descriptions, but numerical assignment is not measurement. Regression coefficients describe statistical relationships within valuation data; they cannot independently establish the measurement status of the dependent variable. If the underlying valuation procedure has not demonstrated a lawful preference measure, subsequent modelling cannot manufacture one. A score of 0.6 cannot simply be assumed to represent 60% of a measurable utility attribute. Even if that interpretation were established, it would remain confined to utility space.

If death creates a problem for health-state valuation, abandoning the death anchor does not resolve the underlying failure of measurement. The scientific requirement remains unchanged: define the attribute, establish its axiomatic measurement properties and demonstrate the legitimacy of the proposed arithmetic. Reversing the relationship between utilities and QALYs cannot substitute for these requirements. It simply relocates the measurement problem from one unestablished numerical construct to another.

ISPOR's recommendations nevertheless proceed to the selection, combination and adjustment of these scores. Brazier et al. (2019), for example, discuss multiplicative utility adjustments without establishing that the quantities entering the calculation are ratio measures. The central question remains unanswered: what empirical measurement structure authorizes division, multiplication or proportional interpretation of utility scores within utility space?

This is the critical failure of the utility tradition within ISPOR's practice standards. Guidance is provided for generating scores, assessing their statistical performance and incorporating them into economic evaluation, but the prior question of representational measurement is left unresolved. Neither the conventional anchors of zero and unity nor the statistical properties of a valuation algorithm establish a ratio measure of utility.

If utility were established as a lawful ratio measure, multiplication by survival duration would not, in itself, violate the rules of arithmetic or dimensional homogeneity. The fundamental objection is that conventional health-state utility scores have not been demonstrated to possess those measurement properties. Consequently, their multiplication by duration cannot be justified as the construction of a lawful derived measure."

The conclusion is therefore limited but decisive. ISPOR provides detailed instructions for creating and manipulating numerical utility scores without demonstrating that these scores possess the properties required for lawful ratio measurement**.** This failure arises within utility space itself, before any consideration of survival duration. It cannot be remedied by statistical validation, algorithmic refinement or the conventional acceptance of utility tariffs.

ISPOR AND THE QALY: THE IMPOSSIBLE ADJUSTMENT OF TIME

The QALY introduces a second and more fundamental measurement problem. Unlike health-state utility scores, survival duration is an established ratio measure. Time possesses a meaningful absolute zero, equal units, an empirically defined ordering and admissible proportional comparisons. It satisfies the requirements of unidimensionality, dimensional consistency, invariance and the relevant axioms of representational measurement. Its admissible transformations preserve ratios. Consequently, durations can legitimately be added, divided and compared within time space.

A proportional adjustment to duration must itself be justified within that measurement space. Consider two measured durations, t1 and t2. Their ratio is:

r = t1 / t2 = 0.6

Both quantities are measures of the same attribute. They are dimensionally homogeneous and commensurable. Their dimensions cancel:

[Time] / [Time] = 1 (dimensionless)

The resulting coefficient expresses a lawful proportion of duration. For example, six years divided by ten years gives 0.6. Applying this proportion to a specified duration is justified by the established relationship between measured quantities in time space.

The conventional QALY instead introduces a numerical utility coefficient and multiplies it by duration:

Q = U × T

For example:

Q = 0.6 × 10 years = 6 QALYs

This calculation assumes that the utility score of 0.6 can legitimately serve as a proportional adjustment to ten years of survival. Yet the coefficient originates in an entirely different domain: preferences associated with multidimensional health-state descriptions. It is not constructed from two measured durations and has not been demonstrated to represent a proportion of time.

The difficulty cannot be resolved merely by claiming that utility is dimensionless. Even if a utility ratio could be established as **u / U = 0.6**, its dimensional expression would be:

[Utility] / [Utility] = 1 (dimensionless)

This remains a relationship within utility space. It does not establish that **t1 / t2 = 0.6** in time space. A scientifically defensible QALY would require an independently demonstrated mapping between these distinct attribute spaces.

ISPOR's Good Practices recommendations proceed with QALY calculations without establishing either the required ratio properties of utility scores or the empirical mapping necessary to transform utility proportions into duration proportions. The arithmetic is accepted because it is conventional, not because its measurement foundations have been demonstrated. The QALY therefore represents an unsupported adjustment of a lawful time measure by a coefficient whose measurement status and relationship to time remain unestablished. This is the central failure of dimensional homogeneity and commensurability in conventional cost-utility analysis. ISPOR's endorsement of the QALY perpetuates arithmetic without measurement.

ISPOR AND RASCH MEASUREMENT: THE NEGLECT OF LATENT ATTRIBUTES

The failure of ISPOR to recognize Rasch measurement reflects a more fundamental omission in conventional HTA: the distinction between manifest and latent quantitative attributes. Once the separate elements of a health-status description have been bundled into a composite score, there appears to be little concern for whether the resulting number represents a measurable attribute. The underlying assumption is that numerical aggregation is sufficient to establish a quantitative outcome. This assumption disregards the requirements of representational measurement.

The distinction between manifest and latent attributes is fundamental. Manifest attributes are directly observable and can be measured on linear ratio scales. Survival duration, hospital days, treatment episodes and resource quantities possess empirically identifiable units and meaningful absolute zeros. Their measurement properties permit lawful proportional comparisons and arithmetic consistent with dimensional homogeneity and commensurability.

Latent attributes, by contrast, are not directly observable. Pain severity, fatigue, physical functioning and other patient-reported outcomes represent constructs whose magnitudes must be inferred from responses to appropriately designed items. Assigning numbers to response categories does not establish equal measurement units, an absolute zero or meaningful ratios. The responses are observations requiring a measurement model, not measures simply because numerical codes have been attached to them.

Conventional patient-reported outcome assessment frequently proceeds by summing item scores to create a total. This assumes that the response categories and items can be combined as though they contribute interchangeable units of the same attribute. Yet the items may differ in difficulty, severity, discrimination and dimensional content. Without empirical evidence supporting the required measurement structure, the resulting total cannot be interpreted as a lawful quantitative measure. At best, a summed score may provide an ordinal ordering of respondents. At worst, particularly where heterogeneous attributes have been combined, the total possesses no defensible axiomatic measurement interpretation.

Rasch measurement enters precisely because latent attributes cannot be measured by numerical assignment or summation alone**.** It provides a probabilistic framework for testing whether item responses support the construction of a unidimensional measurement scale. Rather than assuming that a questionnaire produces a measure, Rasch analysis examines whether the empirical response structure satisfies the requirements of the measurement model.

Central to this approach are unidimensionality, person–item conjoint measurement, specific objectivity, local independence, ordered response thresholds, appropriate item and person fit, and invariance across relevant populations and conditions. Differential item functioning must be investigated to determine whether items retain their measurement properties across groups. Where these requirements are satisfied, person locations can be estimated on a common logit scale, with invariant units supporting meaningful comparisons of the latent attribute.

The distinction between the Rasch logit scale and a conventional questionnaire total is decisive. A raw score is an aggregation of assigned category values. A Rasch estimate is derived from an explicitly specified measurement model whose assumptions can be tested against empirical observations. The logit expresses the logarithm of an odds relationship, so a difference of one logit corresponds to a factor of approximately 2.718 in the associated odds. This does not mean that an arbitrary raw score has acquired ratio properties. Nor does it make the logit location scale itself a conventional ratio scale with an absolute zero. Any proportional interpretation of the latent attribute requires a separately justified measurement representation.

ISPOR's methodological guidance and university HTA curricula have not established Rasch measurement as a foundational requirement for quantitative assessment of latent attributes. Instead, attention is commonly directed toward questionnaire development, reliability, construct validity, responsiveness and statistical analysis of scores. These procedures do not substitute for demonstrating the axiomatic properties required for measurement.

The omission has important consequences for health-status valuation. EQ-5D descriptions combine distinct dimensions of health, while conventional valuation procedures assign preference scores to their combinations. Rasch analysis cannot convert such a multidimensional bundle into a single lawful attribute merely by applying a statistical model. The attribute must first be defined, and the items must be shown to support its measurement. Separate attributes may require separate measures, with therapy-impact claims expressed in the units appropriate to each.

The failure is therefore both methodological and educational. Students are taught to generate, summarize and analyze patient-reported scores without necessarily being introduced to the

distinction between manifest and latent attributes or the requirements of conjoint measurement. The result is another instance of arithmetic preceding measurement.

Rasch measurement is not simply an additional statistical technique that ISPOR might recommend. It represents an essential challenge to the assumption that numerical questionnaire scores are already measures. Its importance lies in requiring the analyst to demonstrate, rather than presume, that a defined latent attribute can be represented quantitatively. Without that demonstration, patient-reported outcome scores remain numerical assignments whose admissible arithmetic and scientific interpretation have not been established.

ISPOR AND INSTRUMENT MAPPING: THE IMPOSSIBILITY OF CROSS-SCALE MEASUREMENT

Representational measurement establishes two distinct approaches to the quantitative assessment of therapy impact. Manifest attributes, such as survival duration, hospital days and resource utilization, are measured on linear ratio scales with equal units and a meaningful absolute zero. Latent attributes, such as pain, fatigue, physical functioning and ability, require fundamental measurement through Rasch models. Rasch measurement establishes the possession of a defined, unidimensional latent attribute, expressed in logits. The logit scale has equal intervals but an arbitrary origin. It is not a linear ratio scale, and ratios of logit locations cannot represent proportions of the latent attribute possessed. These differences are fundamental to the permissible arithmetic and interpretation of the resulting measures.

Mapping between Rasch logit measures and linear ratio measures is not a legitimate scale conversion in either direction. A Rasch measure of possession of a latent attribute cannot simply be transformed into a linear ratio measure, nor can a linear ratio measure be converted into a Rasch measure of possession. The two measurement structures have different properties, and their numerical values do not represent interchangeable quantities. Statistical regression cannot change those properties, establish a missing absolute zero or confer fundamental measurement status on a predicted score. Mapping in this sense is impossible as a measurement operation. A numerical prediction must not be confused with a lawful conversion between measurement scales.

Even mapping between established linear ratio measures of different attributes presents strict requirements. The attributes must be independently defined, their measurement properties demonstrated, and the proposed relationship between their measurement spaces empirically justified. Possession of ratio-scale properties does not establish commensurability or authorize proportional substitution. This is the same problem encountered in constructing QALYs. Even if health-state utility were established as a linear ratio measure, a separate demonstration would still be required to show that a ratio in utility space could represent a proportional adjustment in time space. The mere existence of two ratio measures does not establish a lawful relationship between them.

ISPOR's mapping practice standards reveal a fundamental failure to recognize the importance of Rasch measurement. Its Good Practices for Outcomes Research Task Force recommends statistical procedures for predicting preference-based health-state utility scores from non-preference-based patient-reported outcome (PRO) instruments [25] . These procedures commonly involve predicting

EQ-5D-3L utility tariffs from disease-specific questionnaire scores. ISPOR emphasizes regression specification, predictive accuracy, uncertainty and model performance. Yet its mapping recommendations do not make fundamental Rasch measurement a prerequisite for the quantitative interpretation of latent attributes. They consequently fail to confront the distinction between a Rasch logit measure of possession and a purported linear ratio measure of health-state utility.

Consider a PRO instrument intended to assess physical functioning. Conventional practice assigns numbers to ordered response categories and sums them to generate a total score. Such a summation is not automatically a measure of physical functioning. Fundamental Rasch measurement requires demonstration that the responses support measurement of a defined, unidimensional attribute. Where the requirements are satisfied, the resulting measure expresses possession of physical functioning in logits. This is the scientific starting point that ISPOR's mapping procedures bypass. Rather than establishing fundamental measurement, ISPOR permits conventional questionnaire scores to serve as predictors of EQ-5D utility tariffs, with statistical performance treated as evidence of a successful mapping.

The proposed target introduces a further measurement failure. EQ-5D-3L utility tariffs are derived from preferences over multidimensional health-state descriptions through regression-based valuation algorithms. They have not been demonstrated to possess the properties of a unidimensional linear ratio measure. A regression predicting these tariffs from PRO scores therefore begins with an unestablished source measure and ends with an unestablished target measure. Even replacing the conventional PRO summation with a valid Rasch logit measure would not resolve the problem. A Rasch measure of possession of physical functioning cannot be treated as a proportional utility measure, and no regression algorithm can make it one. The predicted tariff remains a numerical estimate, not an established measure of health-state utility.

The implications for ISPOR's professional standards are substantial. Instrument mapping has become an accepted practice in economic evaluation despite the absence of the measurement foundations required to interpret its outputs as lawful quantitative measures. The failure is not simply inadequate statistical specification. It is the substitution of statistical prediction for fundamental measurement. ISPOR's mapping framework proceeds as though numerical relationships between instrument scores can establish the equivalence of measurement spaces. It ignores the prior requirement to establish the attributes, scales and admissible operations involved.

Instrument mapping is therefore another manifestation of ISPOR’s measurement failure. The numerical output demanded by cost-utility analysis is specified first, and statistical procedures are developed to reproduce it. Fundamental Rasch measurement of latent attributes is bypassed, the distinction between logit and linear ratio scales is ignored, and the measurement legitimacy of the target utility score is assumed rather than demonstrated. Mapping cannot rescue EQ-5D tariffs or provide a lawful foundation for QALY construction. The appropriate scientific alternative is to measure therapy impact directly through established linear ratio measures of manifest attributes and Rasch logit measures of possession of defined latent attributes, retaining the interpretation and arithmetic appropriate to each measurement structure.

ISPOR AND THE REFERENCE CASE: SIMULATION WITHOUT MEASUREMENT

The ISPOR-supported reference-case tradition represents the institutionalization of arithmetic without measurement. Its purpose is to establish a standardized framework for economic evaluation, typically combining simulated survival, health-state utility scores, resource costs and uncertainty analysis to generate incremental cost-effectiveness ratios. Yet standardization does not establish scientific legitimacy. A reference case that fails to specify the measurement requirements governing its inputs, calculations and outputs cannot claim the status of a scientifically validated quantitative framework.

The foundational error is the treatment of health-state utility scores as legitimate inputs to simulation models without first demonstrating their measurement properties. In conventional Markov and related state-transition models, patients occupy health states defined by clinical characteristics or multidimensional descriptions. Numerical utility scores are assigned to these states and multiplied by the duration spent in each state. The model then aggregates the resulting products across states and cycles to generate QALYs.

QALYs are not generally entered as model inputs. They are constructed within the simulation from utility scores and survival duration. This distinction is critical. Duration is a lawful ratio measure in time space. The assigned utility score, however, has not been demonstrated to represent a unidimensional preference attribute possessing an absolute zero, equal invariant units and meaningful ratios. Nor has any empirically established mapping justified its interpretation as a proportional adjustment to duration. The model therefore creates QALYs through an arithmetic operation whose measurement legitimacy has not been established.

The problem extends beyond utilities. Every quantitative input to a simulation must be evaluated against the measurement properties required for its intended use. Nominal classifications, ordinal ratings, interval measures and ratio measures cannot be treated as interchangeable numerical quantities. The admissibility of addition, subtraction, multiplication and division depends upon the scale properties of the variables involved, together with dimensional homogeneity and commensurability. Where model inputs represent latent attributes, their measurement status must be demonstrated rather than inferred from numerical scoring conventions. Transition probabilities, event rates, treatment effects, symptom scores, resource quantities and other inputs require explicit definitions and justification of the operations performed upon them.

A reference case should therefore require a measurement audit before simulation begins. Each input must be linked to a defined attribute or legitimate mathematical quantity, its measurement scale established, and its proposed arithmetic justified. The model must preserve the empirical meaning and dimensions of the quantities entering its calculations. An input should not enter a model in a quantitative role unless its measurement properties support that role**.** Statistical estimation, calibration and sensitivity analysis cannot repair a failure to establish those properties.

The cost component introduces a further distinction. Modelled expenditures are assembled from heterogeneous resource quantities multiplied by their respective unit prices. These monetary aggregates may be legitimate accounting totals, but they do not constitute a single unidimensional measure of resource use. Resource consequences should remain identifiable in their original units, allowing claims such as hospital days avoided, consultations reduced or treatment days gained to be evaluated independently of monetary valuation.

The reference case compounds these problems by projecting outcomes over extended, frequently lifetime, horizons. Simulated quantities are presented with numerical precision, accompanied by probabilistic sensitivity analyses and cost-effectiveness acceptability curves. Yet uncertainty analysis addresses variation in model inputs and assumptions; it does not demonstrate that the variables are lawful measures or that the operations combining them are admissible. Nor does internal model consistency establish empirical falsifiability.

The reference case is therefore not a substitute for scientific measurement. Its reliance on utility-generated QALYs, together with the absence of a systematic assessment of measurement properties across model inputs, leaves its central quantitative claims without a demonstrated representational foundation. The appropriate alternative is not a more elaborate simulation but a reconstruction of the analytical sequence: define the attributes, establish lawful measures, determine admissible arithmetic, and advance therapy-impact claims in empirically testable units.

ISPOR's reference-case tradition reverses this sequence. It privileges standardized numerical procedures over the prior establishment of measurement. The consequence is a framework capable of generating increasingly sophisticated simulations without demonstrating that their principal outputs represent measurable empirical quantities. A reference case founded on unestablished measurement cannot confer scientific legitimacy upon the arithmetic it standardizes.

ISPOR AND COST-EFFECTIVENESS CLAIMS: NUMERICAL STORYTELLING WITHOUT MEASUREMENT

The culmination of ISPOR's conventional HTA methodology is the incremental cost-effectiveness ratio (ICER), presented as a quantitative basis for resource allocation. Yet its scientific legitimacy depends upon the measurement properties of the quantities entering the calculation. Division is not justified simply because two numerical values are available. The operands must possess measurement properties appropriate to the proposed operation, and the resulting quantity must have a defined empirical interpretation. **For division to yield a meaningful quantitative ratio, both operands must be ratio measures.** Without these requirements, cost-effectiveness analysis becomes an exercise in numerical storytelling rather than scientific measurement.

The conventional ICER is expressed as:

ICER = Incremental Cost / Incremental QALYs

Neither component possesses the required measurement foundation.

The numerator is not a single, unidimensional ratio measure of a resource attribute. It is an accounting aggregate constructed by multiplying heterogeneous resource quantities by their respective unit prices and summing the monetary products. Hospital days, consultations, pharmaceuticals and diagnostic procedures represent different resource attributes. Their aggregation into monetary costs does not establish a single underlying resource-utilization attribute with a true zero and unidimensional properties plus the measurement properties required for interpreting the resulting total as a homogeneous physical magnitude.

The denominator presents an even more fundamental problem. Incremental QALYs are calculated from utility-adjusted survival durations, commonly aggregated across health states and model cycles. Yet the QALY has no demonstrated measurement structure. Utility scores have not been established as lawful ratio measures, and no empirically validated mapping demonstrates that a utility coefficient can represent a proportion of survival duration. The QALY therefore cannot be assumed to constitute a measured attribute merely because its calculation produces a numerical result.

The ICER compounds these failures. Dividing a monetary cost difference by a constructed QALY difference produces a numerical quotient conventionally labelled cost per QALY. The quotient does not represent a demonstrated empirical relationship between two appropriately measured attributes. Its apparent precision is supplied by arithmetic, not by measurement. To claim for cost-effectiveness is merely to make clear a lack of understanding of representational measurement. The ratio is meaningless.

The subsequent comparison with cost-per-QALY thresholds introduces another layer of numerical convention. Thresholds such as £20,000–£30,000 per QALY do not establish the measurement legitimacy of the denominator or rescue the ICER from its foundational defects. They merely provide decision rules for interpreting numerical results generated within the same unsupported framework. A cost-per-QALY result has no legitimacy as a quantitative measure when the QALY denominator fails the requirements of representational measurement. Division cannot confer measurement properties on its operands. The resulting ICER is merely a numerical quotient, not an empirically meaningful measure of cost per unit of therapeutic benefit. Neither statistical precision, sensitivity analysis nor comparison with a willingness-to-pay threshold can remedy this foundational failure.

ISPOR AND THE INSTITUTE FOR CLINICAL AND ECONOMIC REVIEW: PUBLIC ARBITRATION WITHOUT LAWFUL MEASUREMENT

The Institute for Clinical and Economic Review (ICER) occupies a distinctive position in American health technology assessment. Through its value assessment framework, reference-case economic models and health-benefit price benchmarks, ICER seeks to inform judgments about the comparative value of competing therapies. Its assessments are intended to influence pricing, coverage and resource allocation decisions. The authority of these rests on the assumption that their quantitative outputs represent meaningful measures of treatment benefit and economic value [26] [27] . That assumption requires examination against the established axioms of representational measurement.

The first requirement of any simulation model is that its inputs possess the measurement properties necessary for the operations performed upon them. ICER's reference-case approach fails to establish this requirement for health-state utility scores. These scores are derived from preferences over multidimensional health-state descriptions without demonstrating a defined, unidimensional attribute possessing the required ratio-scale properties. Within the simulation, utility scores are combined with survival durations to construct QALYs. Yet the numerical multiplication does not establish that utility represents a lawful proportional adjustment to time. The resulting QALY

therefore lacks the demonstrated measurement properties required to support its interpretation as a quantity of therapeutic benefit.

The failure carries directly into incremental cost-effectiveness ratios. The numerator represents differences between composite monetary totals constructed from heterogeneous resource quantities and their unit prices. Such totals may be meaningful for accounting purposes, but they do not constitute a single measure of a defined resource attribute. The denominator, incremental QALYs, lacks an established measurement foundation. Dividing the former by the latter produces a numerical quotient, but arithmetic cannot confer measurement legitimacy upon either operand. A cost-per-QALY result cannot be claimed as a scientifically meaningful quantitative measure when the QALY itself has no demonstrated status as a measure.

ICER compounds this problem by applying cost-per-QALY thresholds to generate health-benefit price benchmarks. Thresholds of $100,000 or $150,000 per QALY cannot remedy an invalid a failure to recognize that division requires ratio measures as numerator and denominator. Sensitivity analyses, probabilistic simulations and alternative assumptions may describe how numerical outputs vary within the model, but they cannot establish the measurement legitimacy of the quantities being manipulated. The problem precedes uncertainty analysis. It concerns whether the proposed quantities are measures at all. This was made clear by Stevens in 1946.

The implications for ICER's role as a public arbiter of therapeutic value are substantial. A framework that fails the foundational requirements of representational measurement cannot legitimately claim that its cost-per-QALY outputs establish the relative economic value of competing therapies. This does not invalidate the clinical evidence ICER reviews, nor does it prevent legitimate analysis of drug prices, resource expenditures or measured clinical outcomes. It does, however, challenge the authority assigned to model-generated cost-effectiveness rankings and price benchmarks. The appropriate scientific alternative is to report therapy impact in the units of defined attributes, using lawful linear ratio measures for manifest outcomes and Rasch measurement for latent attributes, with claims subject to empirical testing, replication and falsification.

ISPOR AND UNIVERSITY CURRICULUM FAILURE: TRAINING IN NUMERICAL STORYTELLING

The failure of university education in health technology assessment is evident at multiple levels, but all can be traced to a common omission: the absence of representational measurement as the scientific foundation for quantitative assessment. Whether this reflects an institutional unwillingness to recognize the requirements of measurement or, more seriously, a lack of awareness that these requirements exist, the consequence is the same. Students are introduced to increasingly sophisticated analytical procedures without first being taught the conditions under which numerical assignments constitute legitimate measures.

The omission is particularly serious because representational measurement is not an optional methodological specialization. It establishes the relationship between empirical attributes and their numerical representation. Every quantitative assessment must begin with the identification of a defined, unidimensional attribute, followed by demonstration of the measurement properties

required for the intended analysis. The distinctions between nominal, ordinal, interval and ratio scales are fundamental. They determine which numerical operations are admissible, whether differences and proportions are meaningful, and whether multiplication and division can be justified. These requirements are reinforced by the axioms of representational measurement, dimensional homogeneity and commensurability.

Yet conventional HTA education reverses this scientific sequence. Students are taught health-state valuation, time trade-off, standard gamble, utility algorithms, QALY construction, cost-effectiveness analysis and reference-case simulation as established quantitative methodologies. The central question is rarely posed: what attribute is being measured, what are its measurement properties, and what arithmetic do those properties permit? Instead, numerical scores are accepted as legitimate inputs because they are generated by recognized instruments, statistical algorithms or established professional conventions.

The consequences extend across the curriculum. Students learn to interpret duration ratios as preference utilities without demonstrating a corresponding ratio measure in preference space. They learn to multiply utility scores by survival duration without establishing the empirical mapping required to justify a proportional adjustment in time space. They learn to construct QALYs, aggregate simulated outcomes and calculate cost-per-QALY ratios without first demonstrating the measurement legitimacy of the quantities involved. They are subsequently taught to interpret these calculations against cost-effectiveness thresholds as though numerical precision established scientific validity.

The distinction between manifest and latent attributes is similarly neglected. Manifest attributes such as survival duration and resource quantities require lawful linear ratio measurement. Latent attributes require explicit measurement models capable of establishing unidimensionality, invariant units and the relevant axiomatic properties. Within the proposed reconstruction of HTA, Rasch measurement provides the framework for constructing and testing measures of latent attributes. Conventional statistical procedures, psychometric reliability and regression-based valuation algorithms cannot be assumed to establish these properties.

ISPOR occupies an unfortunate influential position in perpetuating this educational framework through its Good Practices recommendations, competency specifications, professional courses and methodological guidance. These materials help define what universities regard as appropriate HEOR and HTA knowledge. The failure to place representational measurement at the foundation of professional training therefore extends beyond individual institutions. It becomes a systemic curriculum failure transmitted across successive generations of faculty, researchers and students.

The consequences are not confined to theoretical misunderstanding. Graduates enter academic research, pharmaceutical companies, consulting organizations and HTA agencies trained to produce numerical estimates without necessarily possessing the knowledge required to establish whether those estimates represent measurable empirical quantities. They become proficient in simulation, uncertainty analysis and statistical modelling while remaining unfamiliar with the axiomatic requirements that determine whether the underlying arithmetic is legitimate. Professional competence is consequently defined by mastery of established procedures rather than demonstrated understanding of scientific measurement.

The result is training in numerical storytelling**.** Students are taught to construct internally consistent quantitative narratives from numerical inputs whose measurement properties have not been established. Increasing mathematical sophistication gives these narratives an appearance of scientific authority, but complexity cannot compensate for the absence of lawful measurement. A simulation does not become scientifically meaningful because it is elaborate, nor does a numerical result become an empirical measure because it is reproducible within a computer model.

The necessary response is not the addition of a brief introductory lecture on measurement scales. It requires a fundamental reconstruction of the HTA curriculum, beginning with attributes, representational measurement, scale properties, admissible arithmetic and empirical falsifiability. These foundations must precede instruction in statistical analysis, economic evaluation and simulation. The existing nine-unit Maimon Research HTA Reconstruction Program demonstrates that such a transition can be organized and delivered now .

The central failure of university HTA education is therefore not a shortage of quantitative techniques. It is the teaching of arithmetic before measurement. Until this sequence is reversed, universities will continue to graduate students trained in the manipulation of numbers rather than the scientific measurement of therapy impact.

REPLACING ISPOR: THE RECONSTRUCTION OF HTA

ISPOR's advocacy of legacy HTA must be abandoned. The accumulated evidence of measurement failure leaves no credible basis for suggesting that marginal methodological changes, revised statistical algorithms or additional reporting standards can rescue a curriculum founded on unestablished measurement claims. The issue is not whether existing methods might be improved, but whether the quantities on which they depend satisfy the requirements of scientific measurement. There is no justification for preserving a false curriculum simply to protect established professional practices or institutional reputations.

The failure begins with health-state valuation. Numerical scores generated through time trade-off, standard gamble and regression-based utility algorithms have not been demonstrated to constitute ratio measures of a defined, unidimensional attribute. Decrements from an arbitrarily assigned unity do not establish utility measurement. Multiplication by survival duration does not establish the QALY as a lawful measure. Lifetime simulation models, incremental cost-effectiveness ratios and willingness-to-pay thresholds merely extend the consequences of this initial failure. Additional arithmetic cannot establish the missing measurement properties.

The educational consequences are profound. Generations of students have been taught to accept utility scores, QALYs, ICERs and reference-case simulations without first establishing the measurement status of their inputs. Arithmetic has replaced measurement**.** Statistical estimation has substituted for empirical representation, while numerical predictions have been presented as scientific evidence. The result is a curriculum that reproduces foundational measurement errors across successive generations of researchers and practitioners.

The reconstruction of HTA requires curriculum replacement, not curriculum supplementation**.** Adding lectures on measurement theory to an otherwise unchanged program would preserve the

fundamental contradiction. A curriculum cannot simultaneously insist that lawful measurement precede arithmetic while endorsing the multiplication of unvalidated utility scores by survival duration.

A practical alternative already exists. Maimon Research has developed a nine-unit program for the reconstruction of HTA, providing a structured transition from legacy methods to an empirical science founded on representational measurement [28] . The program begins with the definition of attributes, empirical relational structures, measurement scales, admissible transformations, dimensional homogeneity and commensurability. The program demonstrates that the entire methodological baggage of ISPOR and legacy HTA can be discarded by returning to two established foundations of quantitative measurement: linear ratio measures for manifest attributes and Rasch logit interval measures for latent attributes. Both begin with explicitly defined, unidimensional attributes rather than multidimensional health-state descriptions.

This returns HTA to normal science, where quantitative claims are grounded in lawful measures and subjected to empirical testing and falsification. There is no longer any need for absurd debates over health states supposedly worse than death, negative utilities or alternative valuation anchors. These problems arise from the artificial construction of utility scores through regression-estimated decrements from unity. Rejecting that construction eliminates the problem rather than attempting to repair it. The required measurement approaches already exist and have done for 80 years.

The program proceeds to the construction of measurable therapy-impact claims. Treatment outcomes must be expressed in the units of defined attributes rather than hypothetical composite indices. Survival gains are measured in time, resource consequences in specified resource quantities, and patient-reported outcomes through instruments with demonstrated measurement properties. Claims must be empirically testable, reproducible and potentially falsifiable. Simulation models may have a supporting role, but their inputs must first satisfy the measurement requirements appropriate to their intended use.

This reconstruction also changes the purpose of economic evaluation. Instead of asking whether a simulated cost-per-QALY ratio falls below a conventional threshold, the emphasis shifts to demonstrating measurable treatment effects and identifying the resources required to achieve them. Monetary accounting remains relevant, but it cannot substitute for measurement of the underlying resource quantities or therapy outcomes.

The significance of the nine-unit program is that the transition can begin now**.** The foundations of representational measurement are established, and the program provides an organized route for faculty development, graduate education and research reconstruction. It is available through structured distance learning and can be adapted for institutional teaching and faculty training. There is no need to await another generation of methodological debate.

ISPOR's responsibility cannot be discharged through another Good Practices report. Its professional standards and educational activities have contributed to the institutional acceptance of legacy HTA. The appropriate response is to acknowledge the measurement failures and cease presenting unestablished numerical constructs as scientific measures. Reconstruction, however, does not depend upon ISPOR's approval. Universities, research institutions and professional educators can initiate the transition independently.

The reconstruction of HTA is not an exercise in improving numerical storytelling. It is the replacement of arithmetic without measurement by an empirical science founded on lawful measures. The nine-unit program demonstrates that this transition is possible now. Continuing to teach legacy HTA as established quantitative science, without demonstrating its measurement foundations, perpetuates the very failure that reconstruction is intended to overcome.

.CONCLUSION

The ISPOR legacy HTA framework for health-status assessment offers an apparently attractive approach to quantitative analysis for those who have little understanding of, or interest in, the foundations of scientific measurement. There is no requirement to begin with a defined, unidimensional attribute, establish its measurement properties, distinguish between interval and ratio scales, or demonstrate compliance with the axioms of representational measurement. The rules governing admissible arithmetic are effectively set aside. What matters is the ability to generate numerical scores and manipulate them within an accepted analytical framework.

This creates a methodological environment in which the production of numbers becomes an end in itself. Health-state descriptions are assigned utility scores, duration ratios are reinterpreted as preference measures, utilities are multiplied by survival time, and the resulting QALYs are incorporated into cost-effectiveness calculations. At no stage is there a systematic requirement to demonstrate that the numbers possess the measurement properties necessary to justify these operations. Mathematical convenience replaces empirical justification, while professional convention substitutes for the axioms of measurement.

The attraction of this framework is precisely that it imposes few measurement constraints upon the analyst. Numerical scores can be combined, adjusted, aggregated and projected over a lifetime without first establishing whether the underlying attributes support such arithmetic. Statistical sophistication and increasingly elaborate simulation models then provide an appearance of scientific precision. It is easy to teach with few demands on the instructor.

The consequence is a system in which manipulating numbers takes precedence over measuring attributes**.** ISPOR's continued endorsement of this approach perpetuates the distinction between the appearance of quantitative science and its empirical foundations. The central failure of legacy HTA is not an inability to perform arithmetic. It is the failure to establish whether that arithmetic has any legitimate measurement meaning. Until representational measurement determines what calculations are admissible, health-status valuation will remain an exercise in numerical storytelling rather than scientifically defensible quantitative assessment. This is the global ISPOR message selling legacy HTA.

ACKNOWLEDGEMENT.

I acknowledge that I have used OpenAI technologies, including the large language model, to assist in the development of this work. All final decisions, interpretations, and responsibilities for the content rest solely with me

ABOUT THE AUTHOR

Dr Paul Langley is a health economist. Originally from the UK, Dr Langley completed his graduate studies in Canada and has taught in the UK, Canada, Australia and the United States. After more than 20 years teaching at university in Australia, Dr Langley moved to take up academic positions in the United States. He now lives in Tucson, Arizona, to enjoy the Catalina Mountains, the Sonoran Desert, the absence of mosquitoes and the swimming pool.

REFERENCES

[1] Maimon Research. ISPOR. https://maimonresearch.com/ispor/

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

[3] Stevens S. (1946). On the theory of scales of measurement. *Science*, **103**(2684), 677–680.

[4] Luce RD, Suppes P. Representational measurement theory. In: Pashler H, Wixted JT, editors. *Stevens' Handbook of Experimental Psychology*. 3rd ed. Vol. 4: Methodology in Experimental Psychology. New York: John Wiley & Sons; 2002. pp. 1–41

[5] Helmholtz, H. von (1887**).** Zahlen und Messen, erkenntnistheoretisch betrachtet [Numbering and measuring from an epistemological viewpoint]. In *Philosophische Aufsätze, Eduard Zeller gewidmet*. Leipzig: Fues's Verlag. English translation: *Counting and Measuring* (1930). Princeton, NJ: D. Van Nostrand.

[6] Hölder, O. (1901**).** Die Axiome der Quantität und die Lehre vom Mass [The axioms of quantity and the theory of measurement]. *Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-Physische Klasse*, **53**, 1–64.

[7] Campbell, N. R. (1920**).** *Physics: The Elements*. Cambridge: Cambridge University Press. Republished as *Foundations of Science: The Philosophy of Theory and Experiment* (1957). New York: Dover Publications

[8] Campbell, N. R. (1928). *An Account of the Principles of Measurement and Calculation*. London: Longmans, Green and Co

[9] Suppes, P., & Zinnes, J. L. (1963**).** Basic measurement theory. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of Mathematical Psychology* (Vol. 1, pp. 3–76). New York: John Wiley & Sons

[10] Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, **1**(1), 1–27

[11] Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of Measurement, Volume I: Additive and Polynomial Representations*. New York: Academic Press

[12] Suppes, P., Krantz, D. H., Luce, R. D., & Tversky, A. (1989). *Foundations of Measurement, Volume II: Geometrical, Threshold, and Probabilistic Representations*. San Diego, CA: Academic Press.

[13] Luce, R. D., Krantz, D. H., Suppes, P., & Tversky, A. (1990)**.** *Foundations of Measurement, Volume III: Representation, Axiomatization, and Invariance*. San Diego, CA: Academic Press.

[14] Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research. Expanded edition (1980), Chicago: University of Chicago Press.

[15] Langley P. Legacy HTA ISPOR – The endorsement of measurement inversion in practice pattern standards. Logit Working Paper No. 2000. https://maimonresearch.com/logit-working-paper-no-2000-august-2026/

[16] Dawkins, R. (1976). *The Selfish Gene*. Oxford: Oxford University Press. Chapter 11, "Memes: The New Replicators

[17] Sperber, D. (1996)**.** *Explaining Culture: A Naturalistic Approach*. Oxford: Blackwell Publishers.

[18] Boudry, M., & Braeckman, J. (2011)**.** Immunizing strategies and epistemic defense mechanisms. *Philosophia*, **39**(1), 145–161

[19] Wolowacz, S. E., Briggs, A., Belozeroff, V., et al. (2016). Estimating health-state utility for economic models in clinical studies: An ISPOR Good Research Practices Task Force Report. *Value in Health*, **19**(6), 704–719

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

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

[22] Torrance, G. W., Thomas, W. H., & Sackett, D. L. (1972). A utility maximization model for evaluation of health care programs. *Health Services Research*, **7**(2), 118–133

[23] Torrance, G. W. (1976). Toward a utility theory foundation for health status index models. *Health Services Research*, **11**(4), 349–369

[24] Sampson, C., Parkin, D., & Devlin, N. (2024). Is anchoring at 'dead' a theoretical requirement for health state valuation? *Health Economics*, **33**(9), 1929–1935

[25] Wailoo AJ, Hernandez-Alava M, Manca A et al. (2017). Mapping to Estimate Health-State Utility from Non–Preference-Based Outcome Measures: An ISPOR Good Practices for Outcomes Research Task Force Report. *Value in Health*. 20(1):18–27

[26] Institute for Clinical and Economic Review**.** *ICER Reference Case for Economic Evaluations.* Boston, MA: ICER; 2026.

[27] Institute for Clinical and Economic Review**.** *2023–2026 Value Assessment Framework.* Boston, MA: ICER; updated November 2025

[28] Maimon Research. *Reconstructing HTA Programs: Measurement before Arithmetic*. https://maimonresearch.com/hta-reconstruction-program-and-fees/