

MAIMON WORKING PAPER No 5 APRIL 2025**NUMERICAL STORYTELLING: WHY THE QALY WAS NEVER POSSIBLE**

Paul C Langley, Adjunct Professor, Graduate Faculty, College of Pharmacy, University of Minnesota, Minneapolis MN and School of Pharmacy, University of Wyoming, Laramie WY

ABSTRACT

Once the axioms of fundamental measurement are introduced as a reference point for value claims in health technology assessment (HTA), the extent of numerical storytelling becomes apparent. This unfortunate situation can be attributed to the proposal by Klarman et al in 1968 that life years should be adjusted by the quality of those life years lived. This raises significant epistemological issues in the theory of measurement because although time is a unidimensional linear ratio measure, it is mathematically impossible to construct a time discount factor that has the required constant absolute difference properties to adjust the time measure. This follows from the axioms of fundamental measurement where the required discount proportion cannot support multiplication. This constraint was not recognized with the result that, over the next 60 years, significant effort was put into trying to create a preference score or discount factor that could be applied to create QALYs. While this was claimed the process and the characteristics of the discount factor failed to meet required measurement properties. The purpose of this paper is to detail how this transpired. The result is that HTA has instruments creating scores and reference case models with QALYs which are just numerical storytelling. Rather than attempting to cover this up, the solution, as these multiattribute health status scores have no meaning in measurement terms, is to focus on value claims which meet Rasch standards for constant relative difference for disease and target patient specific outcomes. The baggage of HTA scores to include the QALY and reference case models has to be abandoned. The QALY was never possible.

Keywords: numerical storytelling, measurement failure, epistemic disaster, Rasch standards

INTRODUCTION

When Klarman *et al* in 1968 proposed what seemed at the time an elegant solution to the integration of life expectancy and morbidity: *One way to handle this is to adjust life-years according to the quality of life during those years. For example, if the quality of life is regarded as half that of perfect health, then a year of life would be counted as one-half year*, there was, clearly, no expectation that it would lead to the effective end of health technology assessment (HTA) ¹. It would take 60 years. The irony is that if there had been any understanding of the axioms of fundamental measurement, this entire HTA global odyssey endorsing numerical storytelling would have never got off the ground.

The measurement mistake is trivial. Time is a unidimensional ratio measure with constant absolute differences. To create a discounted version called a QALY the requirement was multiplication by another measure with the same properties. Given the Klarman et al proposal, that discounting time would have to be proportional; a multiplicand in the range 0 to 1, the mistake was that the

multiplicand had to have an absolute difference property. By the axioms of fundamental measurement that was impossible. Yet HTA grew, from a hard core of disciples to a global phenomenon, populated by analysts and others who had not the first idea of the axioms of measurement. To compound this epistemic failure, the proportion had to be multiattribute: a description of health states that had universal application. The measurement failure was compounded with the various attempts to construct this holy grail of time discounting even further removed from the reality of measurement.

The purpose of this brief note is to make clear the magnitude of what can only be described, in measurement terms, as an avoidable epistemic disaster. A disaster unparalleled in the history of science and the social sciences. A meme, not a paradigm, that persists to this day with its practitioners unaware of their unwavering commitment to numerical storytelling.

PSEUDOSCIENCE AND NUMERICAL STORYTELLING

While it might be unusual to describe a subject area as not only pseudoscience, but one that is devoted to numerical storytelling, it is entirely appropriate for HTA. Not only is HTA pseudoscience because of its rejection of the standards of normal science and rejection of any commitment to the evolution of objective knowledge, but it is numerical storytelling because of its flagrant disregard of the standards of fundamental measurement in its pursuit of ersatz preference measures to discount time and create mathematically impossible QALYs.

Numerical storytelling describes a dog's breakfast of measurement error. In HTA it refers to the construction of plausible, model-driven narratives that rely on the manipulation of numbers to simulate evidence rather than derive it ². It gives the appearance of precision and objectivity, but the numbers involved, typically derived from ordinal preference scores and assumptions lack measurement validity. These stories, notably the reference case model, are not generated from observed data governed by the axioms of fundamental measurement, but from projections driven by belief, convention, and the demand for policy-relevant answers. The results are treated as evidence, even though they are nothing more than artifacts of internal consistency, divorced from empirical reality.

Pseudoscience, in this context, is the systematic pretense of scientific legitimacy in the absence of adherence to the standards of normal science. HTA qualifies as pseudoscience because it abandons the requirements of testability, reproducibility, and measurement validity ³. The question of demarcation is entirely absent. It embraces constructs, such as the EQ-5D-3L, that violate basic axioms of arithmetic and dimensional analysis, yet continues to propagate them as if they were meaningful quantities. Its claims are not empirically falsifiable but are embedded in reference case models designed to produce acceptability, not truth. In this way, HTA sustains itself through professional ritual rather than scientific rigor. It puts to one side the idea of progress in science or what Popper would describe as the evolution of objective knowledge ⁴

THE AXIOMS OF MEASUREMENT

Measurement, in its scientific sense, is the assignment of numbers to objects or events according to rules that preserve the properties of the thing being measured. In his seminal 1946 paper, Stevens

classified measurement scales into four types: nominal, ordinal, interval, and ratio. Each scale permits specific mathematical operations; only interval and ratio scales support arithmetic manipulations such as addition and multiplication⁵. Crucially, each scale type adheres to a specific set of axioms that constrain what can be validly inferred from numerical representations.

Nominal scales assign labels without implying order (e.g., blood type). Ordinal scales introduce rank order (e.g., pain ratings), but do not support assumptions about the magnitude of differences between ranks. Interval scales preserve equal distances between units (e.g., temperature in Celsius), allowing addition and subtraction. But only ratio scales, which have both equal intervals and a true zero, support multiplication and division. Time, weight, and length are classic ratio scales; preferences are not.

The failure to distinguish between constant absolute and constant relative differences lies at the heart of the measurement failure of HTA. Fundamental measurement requires that numerical operations be limited to quantities with clearly defined and invariant properties. Classical statistical analysis and arithmetic operations such as addition and multiplication demand either linear ratio scales where units exhibit constant absolute differences or logit-based scales with constant relative differences. In the former the existence of a true zero permits proportional comparison. For objectively observed phenomena, only interval and ratio measures with constant absolute differences are admissible. For latent constructs, such as symptoms, need-fulfillment, or functional capacity, measurement must adhere to a different standard: the Rasch model, which yields a logit-based ratio scale defined by constant relative differences⁶. These two measures, the linear interval scales and Rasch logit ratio scales, are the only scientifically defensible foundations for empirical claims and statistical analysis.

Even if one attempted to transform ordinal preference scores to produce a scale of constant relative differences, such as through Rasch measurement (log-odds or logit scales), the resulting metric would still be incompatible with the ratio structure of time. Constant relative differences (expressed in logits) and constant absolute differences (expressed on linear scales) cannot be validly combined in arithmetic operations. They belong to different mathematical systems. The entire premise of the QALY assumes this incompatibility away, to include the reference case simulation modelling.

UNIDIMENSIONALITY AND BUNDLING

The failure to recognize the necessity of compatible measures is only the beginning. Those advocating for the construction of an impossible proportion also failed to grasp a core axiom of fundamental measurement: arithmetic operations involving ratio measures embody unidimensionality. This principle is non-negotiable. Yet, the preference scores used to generate QALYs are inherently multidimensional, collapsing disparate health domains into a single index. This creates an immediate and insurmountable obstacle. When bundled measures are used where "health status" is an amalgam of multiple attributes, the resulting scale lacks any defensible measurement property. There is no theoretical basis or empirical method by which such a bundled construct can yield a scale with the required properties of additivity or multiplicability. In fundamental measurement, dimensional consistency is a precondition for meaningful operations. Without it, equivalence is lost, and the numbers are nothing more than symbolic artifacts, useful

only for storytelling, not for science. To describe an instrument as multiattribute is to exclude from meaningful arithmetic operations.

If bundling is rejected then we have to reject a critical part of the HTA's instrument base: the various families of multiattribute instruments. While one might sympathize with these attempts to create generic health status scores the various EQ-5D, ST-36 and HUI instruments, the fact is that they fail the standards for fundamental measurement and hypothesis testing that has been in place since the scientific revolution of the 17th century⁷. Hypothesis testing involving value claims only has epistemic meaning when the underlying measure satisfies all four conditions: unidimensionality, linearity, ratio properties, and invariance. Otherwise, what may appear as a hypothesis test is merely statistical manipulation on numbers without measurement meaning; it is nothing more than numerical storytelling.

With the rejection of multiattribute generic instruments, HTA has to address the standards for disease specific instruments. As the overwhelming majority of these are attempting to capture latent constructs there is still another hurdle: virtually all will fail the standards for Rasch modelling. This is the only technique available for creating measurement structures for assessing therapy response. Whether by design or not, Rasch techniques for creating constant relative difference logit ratio scales have been ignored in HTA. The fact remains: Rasch predates by almost 20 years the Klarman suggestion for a preference score to discount time. It is one thing to present a proposal that had no chance of ever meeting measurement standards; it is another to fail to endorse an approach, Rasch measurement for constant relative differences, that would have provided a robust and validated logit framework for the measurement of structures in latent traits for disease areas and target patient groups. A failure which, unfortunately, persists to the present day.

THE TIME TRADE-OFF DEBACLE

The Time Trade-Off (TTO) technique stands as one of the most striking violations of the axioms of fundamental measurement in health technology assessment (HTA). Promoted as a means to anchor health states on a hypothetical interval scale from 0 (death) to 1 (perfect health), the TTO presumes that a trade-off between time and quality can reveal the relative value of living in different health states. It is a fiction. What TTO actually reveals is the epistemic negligence of those who built and sustained it, blissfully ignorant of the most basic requirement of arithmetic operations: dimensional homogeneity.

The foundational flaw is simple yet devastating. The TTO attempts to elicit a 'preference score' by asking respondents how many years of life in a less-than-optimal health state they would trade for fewer years in full health. This response is interpreted as indicating how desirable that health state is. From this, a ratio-like score is generated, which is then used to 'adjust' actual time in the QALY framework. But for such a multiplication to be valid, both operands must meet the requirements of fundamental measurement. Time is a ratio scale, with a true zero and equal intervals. The preference score derived from the TTO is neither ratio, nor interval, nor unidimensional. It is impossible to apply the preference score.

The key failure is the absence of unidimensionality. In fundamental measurement, a valid scale must measure a single construct. In the TTO, the underlying attribute supposedly being measured

is “preference for health states”, a concept already vague and ontologically unstable. Worse still, the health states themselves are defined in terms of multiple, incommensurable dimensions: mobility, pain, self-care, mental health, social function, and so on. Asking someone to collapse these into a single trade-off is like asking a respondent to weigh kilograms against decibels. There is no common dimension, no invariant rule, no justifiable arithmetic. Yet the TTO forces precisely such an incoherent judgment, treating the outcome as if it represented a true scalar quantity. It does not.

The arithmetic pretense is absurd. If a respondent indicates that 10 years in a compromised health state is equivalent in preference to 6 years in full health, the TTO algorithm transforms this into a ‘value’ of 0.6 for that state. This number is then interpreted as a ratio; often used in subsequent modeling as if it bore the properties of a linear scale. But this is nothing more than numerical storytelling. The respondent was never operating on a unidimensional ratio scale of ‘health preference.’ Rather, they were confronting a multidimensional and irreducible judgment, synthesizing an internal narrative shaped by life experience, framing effects, mood, and cognitive bias. No psychometric evaluation is conducted. No test for dimensionality is applied. No axioms are invoked, because if they were, the entire technique would collapse under its own epistemic ignorance.

Let us be clear: multiplication in science is governed by strict rules. It is only permissible when applied to quantities measured on scales that support constant absolute differences, such as a linear ratio scale. Multiplication is not defined for quantities measured on logit scales, which express log-odds and operate under a different mathematical structure. Nor is it valid for ordinal scales, which provide only rank order without fixed intervals. Multiplying a time value, a valid ratio measure, by a preference score derived from TTO, which is at best ordinal and almost certainly multidimensional, is an operation with no scientific justification. This is not a harmless simplification; it is a fundamental mathematical error. It renders every QALY derived from such preference scores incoherent. It is as if physics had attempted to calculate acceleration by multiplying speed by color, an operation without meaning, logic, or interpretability.

Moreover, there is no process of calibration or standardization within TTO that could even hint at satisfying unidimensionality. Rasch measurement, the gold standard for constructing logit or constant relative difference measures from ordinal responses, is never applied. The result is not only non-linear, it is non-measurement. The TTO produces numbers, but they are symbolic. They represent a belief system, not a measurement framework.

It is worth emphasizing that the failure of TTO is not merely technical; it is foundational. The technique presumes a stable, transitive preference structure among respondents. It presumes that respondents understand what it means to sacrifice years of life for quality. It presumes, without justification, that preferences are linear, reversible, and comparable across individuals and time. All of this is asserted, never demonstrated. There is no validity. There is no reliability. There is no dimensional coherence. What is left is a ritualized exercise in assumption, the production of numbers that facilitate model-building, not evidence generation.

And yet, from this invalid structure flows the entire edifice of QALY-based cost-effectiveness analysis. The TTO preference score, a psychometric chimera, becomes the foundation for

simulated health economic evaluations. Prices are set. Coverage decisions are made. Access is granted or denied. All because a handful of respondents imagined a scenario in which they might trade years of life for hypothetical health gains, scenarios that, even if they could be coherently framed, say nothing about the empirical consequences of therapy in real populations.

It is not just that the TTO is methodologically weak. It is that it is built on an impossible proposition: that human judgment under multidimensional uncertainty can yield a unidimensional, ratio-scale value. The proposition was always absurd. That it has been accepted and institutionalized speaks not to its scientific strength but to the professional incentives of those who have refused to interrogate its assumptions.

This is not measurement. It is an article of faith dressed in the garb of science. The HTA should have rejected it decades ago. The refusal to do so has created a decades-long detour into pseudoscience, one that cannot be corrected by incremental improvements or new checklists, but only by returning to first principles: unidimensionality, invariance, and fundamental measurement. Without those, the TTO is not a method. It is a mistake.

THE EQ-5D-3L DECONSTRUCTED

As far as can be determined, the development of the EQ-5D-3L gave no consideration to the constraints imposed by fundamental measurement; a neglect that similarly characterizes other multiattribute instruments⁸. It would strain credulity to suggest that the developers understood the requirement for unidimensionality under the axioms of measurement theory and simply chose to ignore it. Yet, in practical terms, that is exactly what occurred. The consequence is unavoidable: any application of these instruments is mathematically incoherent from the outset. Their use in cost-effectiveness modeling, regression analysis, or as the basis for value claims introduces errors that cannot be corrected *post hoc*, because the inputs themselves are not legitimate measures.

The valuation results for the EQ-5D-3L health states are based on applying the TTO technique to the 243 health state descriptions defined by combinations of five dimensions, mobility, self-care, usual activities, pain/discomfort, and anxiety/depression, each with three levels of severity (no problems, some problems, extreme problems). In typical valuation studies (e.g., the UK MVH study), samples from the general population are asked to assign values to selected health states using the TTO method. These responses are then modeled using regression techniques (often ordinary least squares with dummy variables) to estimate utility weights for 10 of the 15 health state parameters; the 5 no problem states are the reference group and are assigned zero. These weights form the basis of the EQ-5D-3L preference algorithm; an algorithm that assigns a single index score to reported health states. The score is capped at unity for perfect health.

If the TTO technique is a monument to the abandonment of measurement standards, the EQ-5D-3L is the architectural extension; a prefabricated framework designed to exploit its outputs while compounding their errors. It is marketed as a simple, standardized, generic instrument to generate quality-of-life scores across disease areas. In practice, it is a pseudometric construction that violates every known requirement for meaningful measurement. Worse, it exports these violations into statistical modeling, where its outputs are treated as dependent variables in regression analysis, as if they bore the attributes of ratio or even interval scales. They do not. The EQ-5D-3L score,

derived from a preference algorithm built on TTO data, is a fundamentally incoherent multiattribute construct. It cannot support multiplication, regression, or even interpretation.

The transformation from descriptive health states to preference scores presumes an underlying unidimensional utility construct. However, the five domains are not manifestations of a single latent trait. They are ontologically distinct, with no common unit of comparison and no empirical justification for bundling. There is no reason, beyond bureaucratic convenience, to suppose that anxiety can be arithmetically traded against mobility, or pain against usual activities. To assign weights to each level of each attribute and combine them into a single score is to commit the cardinal sin of measurement: the construction of a composite from incommensurable components. The resulting index is not unidimensional, not linear, and not invariant across populations or individuals.

This index is derived through a decrement-from-unity algorithm. The preference algorithm assumes a base value of 1 for perfect health and subtracts from this baseline based on the levels of impairment in each dimension. These decrements are not measured; they are imputed from pooled TTO responses using ordinary least squares (OLS) regression on dummy variables. The resulting algorithm is entirely artificial, an engineering of preference fiction. Coefficients are manipulated to 'fit' a dataset of ordinal judgments generated through a method (TTO) that cannot, by any standard, yield interval data. The decrements often create negative values. The score might be anchored at 1.0 but there is no lower bound, although efforts were made to tweak the algorithm to be as close to zero as possible.

The moment we allow a scale to take on negative values without a clear, interpretable zero point, we are no longer dealing with a measurement scale in any scientific sense. In the EQ-5D-3L, the zero point is not even anchored in physiology or patient experience, it is arbitrarily defined by modeling convenience. The notion of 'death' becomes a symbolic anchor, not a scientifically meaningful point of comparison. As a result, both the magnitude and direction of changes on this scale are uninterpretable.

The concept of decrement-from-unity is fundamentally flawed. The use of unity as an upper bound presumes a bounded and normalized scale, yet this is incompatible with the requirements of valid measurement. To support multiplication, such as discounting time in the construction of QALYs, the preference scale must be a ratio scale. This demands a true zero, constant absolute differences (i.e., a linear structure), and no upper bound. The EQ-5D-3L preference scale, like any other instrument constructed from ordinal preference data, violates all three conditions. It lacks a true zero, fails to demonstrate equal intervals, and imposes an artificial upper limit at 1.0. As such, it cannot support the arithmetic operations required to produce a valid QALY.

Moreover, the discount factor itself is not just mathematically incoherent, it is impossible in principle. The inputs to the EQ-5D-3L algorithm, or any other preference-based instrument, are subjective ordinal responses, intended to represent a latent health-related construct. The only legitimate path to transform such ordinal responses into a measure suitable for statistical analysis is through the Rasch logit model, which satisfies the axioms of fundamental measurement by producing unidimensional, interval-scaled logits. However, logits support only additive operations; they cannot be multiplied. This means that any attempt to discount time, a valid ratio

scale, by a logit-transformed preference score is impossible. We have reached, as detailed below, a conceptual dead end: time cannot be discounted by any derived value from these instruments. The QALY fails not only mathematically but epistemically.

To understand the collapse of the QALY, it is essential to distinguish between the properties of ratio and logit scales. A ratio scale, such as time, possesses a true zero, equal intervals, and no upper bound. It supports all arithmetic operations—addition, subtraction, multiplication, and division—because the units are constant and comparisons are meaningful across the scale. In contrast, a logit scale, as produced by Rasch modeling, is an interval scale on a log-odds metric, supporting only additive operations. While it allows for the meaningful comparison of differences in ability or need across individuals or items, it cannot be used in multiplication or ratio formation. To multiply time by a logit violates the mathematical structure of both scales. Since no valid preference-based measure supports multiplication, the QALY—defined as time multiplied by a utility—is a construct built on a category mistake. It attempts to perform an operation that no valid measurement framework permits.

It is difficult to overstate the epistemic irresponsibility of embedding these scores into health technology assessment frameworks. No engineer, physicist, or biostatistician would tolerate the use of such a scale for modeling or prediction. The claim that this is “close enough” for public policy rests on the presumption that real measurement is too difficult or unnecessary. This is the logic of the administrator, not the scientist. It is the logic of narrative, not of evidence.

The EQ-5D-3L is not a measurement instrument. It is a composite index derived from ordinal responses, forced into the shape of a scale by algorithmic prestidigitation. It fails unidimensionality, fails interval properties, and fails all tests of invariance and interpretability. When its scores are used in regression models, cost-effectiveness analyses, or simulation frameworks, the outputs are invalid, numerical artifacts with the semblance of science but none of its substance.

RASCH MEASUREMENT: A SCIENCE OF THE SUBJECTIVE

In the wake of the epistemic incomprehension represented by the EQ-5D-3L and its algorithmic exploitation of TTO-based preference scores, one might ask whether there exists any legitimate path forward for measuring subjective health outcomes. The answer is yes; but only if we abandon the fantasy of preferences with time discounting and embrace the principles of fundamental measurement. This means recognizing that claims about patient experience must begin with the axiomatic structure of measurement theory, not the convenience of incoherent measurement models. This path is defined by Rasch measurement theory, which alone offers a mathematically coherent method for transforming ordinal responses into invariant, interval-level measures; so long as the response items reflect a unidimensional latent trait.

The Rasch model, first articulated by Georg Rasch in the 1960s, does not pretend to “value” multidimensional health states. It does not rely on preferences or arbitrary utility assignments. Instead, it offers a rigorous method for mapping the interaction between item difficulty and respondent ability on a common, unidimensional logit scale for a manifestation of a latent construct. When applied to a well-constructed item set designed to reflect a single trait, such as

need-fulfillment, symptom severity, or functional capacity, the Rasch model produces measures with logit ratio properties. These measures are meaningful. They support arithmetic operations. They can be replicated across samples and, most critically, they respect the logic of measurement.

It is essential to emphasize that the Rasch model is explicitly designed to construct a measure, expressed in logits, that meets the requirements of unidimensionality, linearity, invariance, and a true zero. For Rasch, the data must fit the model, not the reverse. Items in a disease-specific instrument are tested for fit to the Rasch model, and any item that violates the model's assumptions, such as multidimensionality, local dependence, or disordered thresholds, is discarded. This represents a decisive conceptual break from traditional statistical modeling. The Rasch approach is confirmatory and predictive: it presumes a specific measurement structure and tests whether the observed data conform to it. In contrast, traditional models, such as classical test theory or factor analysis, are exploratory and descriptive. They seek to account for as much variance in the data as possible and retain all items that contribute to this explained variance, regardless of whether the resulting score supports fundamental measurement properties. For example, factor analysis may extract latent components based on shared variance across items, but the resulting scores are weighted composites that lack unidimensionality unless the factor structure is strictly constrained. Moreover, these scores are typically ordinal, with unequal intervals and no true zero.

The accountability of traditional models to all observed data, rather than to a strict measurement criterion, ensures that the final instrument cannot be unidimensional, linear, ratio, or invariant. It may be useful for descriptive purposes, but it cannot produce a valid measure suitable for rigorous scientific testing or statistical analysis. At its core, the Rasch model operates by establishing a probabilistic structure around the response process. The likelihood of a respondent affirming an item is modeled as a logistic function of the difference between the person's location on the latent trait (ability) and the item's location (difficulty). When the data fit the Rasch model, we can be confident that the instrument is unidimensional and additive: the ordering of persons remains invariant across item subsets, and the ordering of items remains invariant across different samples of persons. This property, known as specific objectivity, is the linchpin of scientific measurement. Without it, comparisons are meaningless.

The implications are profound. Rasch transforms subjective observations, ordinal responses from patients, into interval-level measures that support legitimate comparisons within and between populations. This is not wishful thinking. It is a mathematical transformation grounded in axioms, confirmed through fit statistics, and capable of detecting when those axioms are violated. If the data do not fit the model, the problem lies not with the Rasch framework, but with the instrument: the items may not reflect a single construct, or they may be poorly targeted. In this way, Rasch theory becomes not only a measurement model but a diagnostic tool, identifying where and why instruments fail.

Contrast this with the EQ-5D-3L and its ilk. These instruments assume from the outset that ordinal responses can be summed, transformed by regression weights, and treated as utility scores. No tests of dimensionality are conducted. No examination of item invariance is made. The "value" is preordained, imposed by an external algorithm, and justified only by convention. Rasch rejects this entirely. It requires that the structure of measurement emerge from the data itself, within strict probabilistic bounds; a requirement proposed and accepted over 60 years ago.

Moreover, Rasch modeling aligns with the scientific requirement that constructs be independently testable and reproducible. Once an instrument has demonstrated good fit to the Rasch model, it can be used in multiple contexts without reweighting or re-norming. The logit scale it produces is fixed: 1 logit difference always represents the same relative difference in the latent trait, regardless of the population being assessed. This is what makes measurement science and what makes Rasch the only defensible approach to constructing subjective health outcome measures that capture constant relative differences.

Rasch also avoids the absurdities of negative scoring. There is no arbitrary or socially imposed zero point in a Rasch scale. Instead, the logit scale is constructed around a mathematically defined origin: the point at which person ability and item difficulty are equal. This zero is not subjective, it reflects an inflection in the probability curve where the likelihood of a positive response is 50%. It is anchored in the structure of the data and the model itself, not in conjecture or preference. As such, the Rasch logit scale is a ratio scale with constant relative differences, supporting all arithmetic operations permissible under the axioms of measurement. Unlike preference-based systems that introduce ill-defined decrements from unity or allow negative values to represent so-called “states worse than death,” the Rasch framework offers a coherent, interpretable continuum. Higher logit scores consistently reflect greater possession of the latent trait being measured, and each unit difference has a fixed meaning across the scale. There is no fictional arithmetic, no utility imputation, and no conceptual sleight of hand, only a unidimensional, invariant, and empirically grounded measurement structure. This is what science demands: a scale whose origin, intervals, and interpretation are determined by the construct itself, not by social convention or algorithmic convenience.

This makes Rasch uniquely suited to the development of patient-reported outcome measures that aim to support evaluable claims from latent constructs. A Rasch-based instrument, grounded in a defined latent trait and constructed with rigorous item design and testing, can provide real-time, actionable, and reproducible metrics of change. It does not rely on indirect utility values or imagined trade-offs. It does not need to be “converted” into QALYs or fed into imaginary reference case models. It stands on its own as an empirical measure of a manifest phenomenon: how a patient reports the experience of their condition, or of a need being fulfilled, and their possession of it over time.

CONCLUSION: THE BAGGAGE OF NUMERICAL STORYTELLING

Once the favorite instruments of HTA are deconstructed, there is considerable numerical storytelling baggage that must be discarded. While this is not surprising after 60 years of unchallenged orthodoxy, the axioms of fundamental measurement demand the wholesale rejection of preference-based techniques such as the TTO and, by extension, all multiattribute instruments. This necessitates the abandonment of the QALY and its associated reference case models, as well as the cost-per-QALY thresholds used to rationalize pricing and access decisions. Abolishing generic instruments does not protect disease-specific instruments that fail Rasch standards for unidimensionality, invariance, and constant relative differences. Their survival is no less a violation of the principles of science.

What remains, until the emergence of Rasch-based disease-specific instruments to capture latent construct traits, are value claims for therapy impact that meet linear measurement standards for constant absolute differences. These include objective clinical scores, biomarkers, value claims for compliance, therapy switching, and resource utilization, all framed as unidimensional, linear, and evaluable measures. They can be tracked and reported in defined time frames and permit meaningful comparisons.

Cleansing the HTA Augean stable will meet fierce resistance, not merely technical or bureaucratic, but institutional and ideological. Entire careers, journals, and consultancy practices have been built on the maintenance of these flawed constructs. What is proposed here is not an adjustment but a reckoning. Yet there is no alternative. If HTA is to survive, not as a professional ritual but as a scientific discipline, then culling must occur. The survival of HTA demands a return to first principles, and that begins with the rejection of everything that measurement theory cannot tolerate.

ACKNOWLEDGEMENT

The author gratefully acknowledges the contribution of ChatGPT, an AI developed by OpenAI, as a collaborative assistant in the preparation of this work. Through structured dialogue, technical support, and critical editorial feedback, ChatGPT provided assistance in drafting, refining, and organizing sections of the text. Responsibility for the final content, interpretation, and conclusions, however, rests solely with the author.

REFERENCES

¹ Klarman H, Francis J, Rosenthal, G. *Cost-effectiveness analysis applied to the treatment of chronic renal disease. Medical Care.* 1968; 6(1): 48–54.

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

³ Pigliucci M. *Nonsense on Stilts: How to tell science from bunk.* Chicago: University of Chicago Press, 2010

⁴ Popper K. *Objective Knowledge: An Evolutionary Approach* (Rev Ed). New York: Oxford University Press, 1972

⁵ Stevens S. On the Theory of Scales of Measurement. *Science.* 1946;103:677-80

⁶ Bond T, Yan Z, Heene M. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (4th Ed). New York: Routledge, 2021

⁷ Wootton D. *The Invention of Science: A new history of the scientific revolution.* New York: Harper Collins, 2015

⁸ Devlin, N. J., Brooks, R. EQ-5D and the EuroQol Group: Past, present and future. *Applied Health Econ Health Policy.* 2017; 15(2), 127–137.