MAIMON WORKING PAPER: No 18 SEPTEMBER 2025

MEASUREMENT IN HTA: PART 2

FROM WILLFUL NEGLECT TO RELATIVISM

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

ABSTRACT

This paper (Part II) explains how and why health technology assessment (HTA) has endured for five decades despite resting on numerical artifacts rather than measures. Part I established that HTA's core constructs fail the axioms of representational measurement theory (RMT). By the 1970s, the relevant measurement standards were explicit: Stevens had drawn the line between ranks and measures; Suppes, Luce, and Tukey had codified the axioms that warrant additivity; Foundations of Measurement had consolidated these results; Rasch had shown how to transform ordered responses into interval scales under strict unidimensionality; and Wright had identified Rasch as uniquely consistent with RMT. Proceeding with utilities and QALYs was therefore not epistemic innocence but willful neglect.

Part II traces HTA's subsequent descent into what is best described as relativistic measurement practice. The analysis begins with the TTO, showing that it cannot yield unidimensional observations and therefore cannot ground measurement. It then demonstrates that multiattribute utility algorithms such as those attached to the EQ-5D manufacture a single number from disparate domains without evidence of a common continuum, violating unidimensionality, invariance, additivity, and any coherent notion of a unit. The QALY is shown to be impossible in principle: multiplying a pseudo-number by time does not create a new quantity. The reference-case model is revealed as institutionalized non-science; an elaborate simulation that embeds non-measures, layers untestable assumptions, aggregates heterogeneous costs into an uninterpretable numerator, and yields outputs that are unfalsifiable and therefore non-evaluable.

The paper then asks why, given these failures, the QALY/reference-case complex has persisted and expanded. The answer is sociological rather than scientific. HTA functions as a memeplex sustained by consensus, authority, and curricular omission. Within this community, "evidence" is what the group agrees to treat as evidence; success is secured by comparability rituals, thresholds, and replication of assumptions rather than replication of phenomena. The analysis draws on the Strong Programme in the sociology of scientific knowledge to explain how such systems survive regardless of truth-value, and on Frankfurt's distinction between lying and bullshit to argue that much HTA discourse is marked by cultivated indifference to truth: numbers are used because they perform administrative work, not because they measure. Bullshit is institutionalized.

The remedy is not repair but replacement. HTA can produce scientific claims only on two admissible foundations: linear ratio scales for manifest attributes (time, costs, resource units) and

Rasch logit ratio scales for latent constructs developed under strict unidimensionality and invariance. These two forms exhaust legitimate options for quantifying outcomes in a way that preserves empirical structure and supports falsifiable, replicable claims. Everything else—multiattribute indices, utilities, QALYs, and reference-case outputs—belongs to numerical storytelling. Rebuilding HTA on these twin foundations would restore dimensional coherence and open value claims to genuine testing; continuing with the current memeplex perpetuates a world outside science, maintained by persuasion and procedure rather than measurement and truth.

INTRODUCTION

In Part I of these two papers, it was demonstrated that health technology assessment (HTA) was never built on legitimate measurement foundations [1]. Despite presenting itself as the scientific guardian of healthcare resource allocation, the methods it adopted, time trade off (TTO), utilities, QALYs, and the reference case model claims, were never measures in the first place. They were numbers masquerading as measures, ordinal rankings of preference paraded as though they sustained the arithmetic of interval or ratio scales. When multiplied by time, a true ratio measure, these pseudo-numbers yielded not a coherent quantity but an incoherent hybrid, violating the principle of dimensional homogeneity at its most basic level. From the outset, failure was inevitable. The only question worth asking is how such an obvious failure not only occurred but persisted for fifty years.

Part I concluded that this genesis could not be explained as epistemic ignorance. By the time HTA was consolidating itself in the 1970s, the standards of measurement were already explicit and widely known. Stevens' 1946 typology had long distinguished between ordinal rankings and interval or ratio measures, warning that misuse leads only to "nonsense" [2]. Suppes had clarified the axioms of extensive measurement [3]. Luce and Tukey had demonstrated the cancellation requirements of conjoint measurement [4]. The publication of *Foundations of Measurement* in 1971 codified these advances in definitive form [5]. Rasch had shown, by 1960, how ordinal observations could be transformed into interval measures under strict conditions, and Wright in 1977 explicitly linked Rasch to the axioms of representational measurement, identifying it as the unique model consistent with them [6 7]. The science was settled. By the 1970s, there was no excuse for ignorance. To proceed with utilities and QALYs was to set aside measurement deliberately, not accidentally. It was willful neglect.

Part II begins from this recognition and asks the harder question: why has HTA persisted in this neglect? Why, despite fifty years of critique from measurement theory and repeated demonstrations of incoherence, has the field insulated itself from science? Why has the QALY/reference case complex survived, defended, and institutionalized itself as orthodoxy? The answer is not scientific but sociological. It lies not in data but in the survival strategies of a memeplex.

The trajectory of HTA from the 1970s onward illustrates a descent from willful neglect into what can only be described as relativistic measurement madness. Once the QALY was institutionalized the field became less about discovering valid knowledge and more about sustaining its own coherence through consensus, authority, and curricular omission. What held HTA together was not empirical warrant but rhetorical force, bureaucratic usefulness, and the insulation of its

practices from critique. Its leaders defended the indefensible not by argument but by silencing, ignoring, or excluding awkward questions.

The purpose of Part II to deconstruct this relativistic universe by returning to the single principle that invalidates it: unidimensionality. If this requirement is enforced, the entire edifice of HTA falls. Time trade-off preferences collapse as ordinal indices. Multiattribute instruments such as the EQ-5D collapse as composites with no common continuum. Utility algorithms collapse as arbitrary conventions. The QALY collapses as the impossible multiplication of a pseudo-number with time. The reference case model collapses as a simulation that embeds non-measures and produces results that cannot be falsified. Each of these failures will be examined and made explicit.

The bottom line is that we have known since the 1970s that there are only two measures that can such HTA claims for therapy response. These are the linear ratio scale for manifest constructs and the Rasch logit ratio scale for latent constructs. The former represents constant absolute differences and the latter constant relative differences, calibrated in logits. There are options. The axioms of RMT allied to the Rasch model are all that is required. These exhaust the options available and do not require a descent into relativistic measurement madness.

THE TIME TRADE-OFF: THE FIRST STEP INTO RELATIVISTIC MATHEMATICAL MADNESS

Despite its continued place in textbooks and HTA training programs, the time trade-off (TTO) is the clearest signal that health technology assessment was destined from the start for measurement failure. The reason is simple: from the perspective of measurement theory, the TTO does not yield unidimensional observations. Responses to TTO tasks generate what may be misleadingly described as ordinal "utilities," but these are nothing more than rankings over multidimensional health state descriptions. Such constructs are not recognized in representational measurement theory, because they cannot be located on a single continuum.

Torrance was the central architect of this misstep**.** His early papers, including Torrance, Thomas, and Sackett's *A utility maximization model for evaluation of health care programs* (1972), his *social preferences for health states* (1976), and his later defense in *Utility approach to measuring health-related quality of life* (1987), laid the foundation for the QALY project [8] [9] [10] . Torrance presented the TTO as a pragmatic solution to the problem of valuing health states: simple to administer, intuitively anchored between "dead" (0) and "full health" (1), and easily aggregated across individuals to create societal weights.

In his 1987 article in the *Journal of Chronic Diseases*, Torrance described the TTO as "the most appropriate method currently available" for eliciting utilities, precisely because it involved explicit trade-offs between multidimensional quality and length of life". This was offered as justification for its use as the basis of QALY construction. Yet nowhere in Torrance's work is there the faintest recognition that these values were not measures. Nowhere does he acknowledge Stevens' warning that ordinal numbers cannot sustain arithmetic. Nowhere does he engage with Suppes' axioms of extensive measurement, Luce and Tukey's demonstration of cancellation, or the codification of the representational program in *Foundations of Measurement* (1971). And nowhere does he mention Rasch's demonstration of how ordinal data might legitimately be transformed into interval

measures. There is nothing in measurement theory that can defend Torrance's extraordinary claim that "the utility approach can be used to measure a single cardinal value, usually between 0 and 1, that reflects the health-related quality of life of the individual at a particular point in time." This statement is not only indefensible in the language of science; it is incoherent when judged against the axioms of RMT. Torrance presents the "utility approach" as if it were a legitimate bridge from preference under uncertainty to the measurement of health states, but it is nothing of the sort.

First, there is no such thing as a "single cardinal value" derived from multiattribute health state descriptions. By definition, such descriptions are composites of different dimensions—mobility, pain, anxiety, self-care—none of which lie on a unidimensional continuum. RMT makes clear that interval and ratio properties are possible only when variation is captured along a single attribute. To collapse multiple, qualitatively distinct domains into a single "value" is to abandon unidimensionality and thereby abandon measurement. The claim that the result is "cardinal" is rhetorical bluster; no test of additivity, cancellation, solvability, or Archimedean consistency has ever been applied or could be passed. It is worth noting that Stevens in his seminal 1946 paper rejected the word 'cardinal' as it blurred the distinction between interval and ratio measures.;

Second, the invocation of "modern utility theory" as a foundation is a sleight of hand. Von Neumann–Morgenstern expected utility theory is not a theory of measurement but a framework for rational choice under uncertainty. The "utilities" of vNM are defined only up to positive affine transformation. They represent orderings of lotteries, not quantities that admit arithmetic combination with physical magnitudes such as time. To claim that this provides a measurement of health-related quality of life is to commit a category mistake of the highest order. What Torrance labels a "cardinal value" is in fact nothing more than an ordinal ranking, dressed up with decimals and forced into the 0–1 interval by convention.

Third, the assertion that this single number "reflects the health-related quality of life of the individual" is an act of pure invention. No demonstration of invariance exists. No empirical test has established that two individuals at the same "utility" are comparable in their subjective state. The number has no unit, no zero, and no continuity. It is not a measure but a preference index, and even as an index it is unstable, sensitive to framing, elicitation method, and cultural context. To call it a measure is to deny the very distinction that Stevens warned of in 1946, when he cautioned that "failure to observe this principle leads to nonsense."

The consequence of Torrance's claim was to legitimize the QALY as a construct, but it was a legitimacy bought at the price of science. Measurement theory was cast aside in favor of a rhetorical appeal to decision theory. What should have been recognized as ordinal preference rankings were passed off as "cardinal values," and their multiplication with time was presented as though it yielded a meaningful quantity. In reality, this was pseudo-measurement—numbers generated by algorithms, endowed with authority by repetition and institutional uptake, but utterly without measurement warrant.

Torrance's formulation should be seen for what it is: a flight from science into numerical storytelling. It is not ignorance, for by 1987The silence of the 1980s and 1990s the axioms of measurement and the availability of Rasch modeling were well established. It is not a mistake, for the conceptual gap between decision theory and measurement was already explicit. It is, rather,

the willing substitution of rhetoric for rigor, a statement that sounds authoritative but collapses under the most elementary scrutiny. To cite "modern utility theory" as a foundation for health state measurement is to erect a monument to confusion: the construction of pseudo-numbers masquerading as measures. That this quote continues to be recycled in HTA literature is not evidence of its scientific validity but of the collective failure of the discipline to grasp the elementary conditions of measurement.

That Torrance never once engaged with measurement theorists damning. His defense of the TTO rested not on science but on pragmatism. It was "good enough" to generate numbers that looked like measures. It produced decimals between 0 and 1, which could be multiplied by time and slotted neatly into cost-effectiveness ratios. It gave policy makers the single index they craved. But from the standpoint of measurement, the enterprise was fraudulent. The TTO does not and cannot produce unidimensional measures. It generates ordinal rankings of multidimensional profiles and then treats them as though they were interval or ratio data.

This was the first step in what can only be described as the descent into relativistic mathematical madness. The TTO was elevated into a standard technique, despite the fact that it fails the most basic criterion of measurement: unidimensionality. From there, the logic unfolded inexorably. Multiattribute instruments such as the EQ-5D-3L were constructed using TTO data as their foundation. Regression models were fitted to these ordinal responses, despite the fact that regression demands interval or ratio dependent variables. Algorithms were concocted to generate "utility weights" for arbitrary health states, even though no unidimensional continuum could possibly underwrite them. The resulting utilities, lacking any measurement warrant, were multiplied by time to yield the impossible QALY. The absurdity reached its zenith with the reference case model, in which entire policy frameworks were built on numbers that never rose above ordinal status.

Torrance must therefore be recognized as a major contributor to the failure of HTA to match the RMT standards. His endorsement of the TTO normalized the idea that multidimensional ordinal preferences could be treated as measures, paving the way for decades of incoherent practice as witnessed in major texts [11] . By refusing to engage with measurement theory, whether through ignorance or willful neglect, he helped set HTA on a course from which it has never recovered.

The TTO is thus more than a technical curiosity. It is the original sin of HTA: a procedure that, by its very design, guaranteed the impossibility of measurement. To call it "the most appropriate method currently available" was not an act of science but an act of expedience. It signaled the triumph of policy convenience over measurement truth, and it remains the starting point of the QALY's descent into mathematical nonsense.

THE UTILITY ALGORITHM: THE SECOND STEP INTO RELATIVISTIC MATHEMATICAL MADNESS

The so-called multiattribute utility is not a measure. It is a number manufactured by a sequence of steps that fail every axiom of representational measurement and violate the most basic requirement of science: unidimensionality. Yet this manufactured number has been passed off for

three decades as if it were a metric of health, forming the cornerstone of HTA's cost-effectiveness industry. To examine the procedure in detail is to see its incoherence exposed.

The creation of so-called utilities or preference scores through the EQ-5D algorithm is a textbook case of measurement failure disguised as arithmetic [12]. [13]A respondent first describes their health state by selecting a level on each of five dimensions, mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. That five-tuple is not a measure of anything; it is a categorical, multiattribute description. The next move is purely algorithmic: each chosen level is assigned a tariff weight that was estimated earlier from time trade-off (TTO) valuation studies using regression. Those level-weights are then combined, typically as a weighted sum with ad-hoc adjustments to fit the data, with the result is subtracted from unity, "perfect health," to yield a single utility number. Mild profiles land between 0 and 1; severe profiles can be driven below zero (e.g., to –0.59 in the UK 3L tariff) and labelled "worse than dead." The entire performance has the look and feel of measurement with inputs, coefficients, output with decimals, but none of it is substance. There was no mention of how such as procedure would yield a unidimensional; measure with RMT requirements.

Start with the raw material. TTO responses are ordinal. A respondent's willingness to trade years of life to avoid a state, ranks that state relative to another; it does not endow the states with equal intervals or a true zero. Treating those ordinal rankings as if they were interval data suitable for linear regression violates the most basic requirement of econometrics: the dependent variable in an OLS model must admit meaningful differences. Replacing OLS with ordered models does not rescue the enterprise, because such models still deliver a latent ordering, not a demonstrated interval or ratio scale, and they do nothing to justify additivity across distinct attributes. The cancellation axioms that would warrant additivity in conjoint measurement are neither tested nor satisfied; they are assumed away.

Now consider the tariff itself. The weights are created by convention, model specification choices, dummy variables, "N3" penalties, country-specific calibrations, not by any demonstration that, say, moving from pain level 1 to 2 is commensurate with moving from mobility level 1 to 2, or that the same "distance" holds anywhere on the putative scale. That is, equal numerical steps are not shown to correspond to equal differences in any single attribute. The resulting score therefore fails unidimensionality by construction: it collapses five distinct domains into one number without evidence that they lie on a common continuum. Without unidimensionality there is no continuity; without continuity interval and ratio talk is empty.

Invariance is absent as well. A genuine measure must not depend on who is being measured or which instrument instance is used. EQ-5D utilities vary with the tariff set (UK, US, Japan, etc.), the modelling choices used to derive that tariff, and even the vintage of the tariff as fashions shift. The same descriptive profile can yield materially different "utilities" across locales and eras. That is not a property of a measure; it is the hallmark of an index tuned to local convention.

Additivity is simply asserted. The algorithm presumes that decrements can be stacked across domains and that cross-domain trade-offs preserve structure. Yet there is no empirical warrant that a one-level worsening in mobility can be "added" to a one-level worsening in anxiety to produce a meaningful composite quantity. Without demonstrated additivity (via cancellation tests), the

weighted sum is just arithmetic on labels. Subtracting that sum from 1 further compounds the fiction by pretending there is a true, absolute zero anchored at "dead" and a true unit anchored at "one year in full health." The appearance of a ratio frame is rhetorical; there is no proof that "0.8" is twice "0.4" in any empirically preserved sense.

Negative utilities expose the artifice. They are not empirical discoveries about health states; they are mathematical artefacts of a linear model with an intercept and penalties that push the weighted sum beyond one. The tariff creates them; observation does not. That a profile computes to –0.59 says only that the chosen regression and coding allow the arithmetic to cross an arbitrary threshold. It tells us nothing about a continuous health attribute, because no such attribute has been demonstrated.

Nor does the algorithm define a unit. What is the unit of an EQ-5D utility? It is not hours, meters, dollars, or logits. There is no operational procedure that shows what one unit difference means or how to calibrate instruments to reproduce it. Without a unit, there is no scale; without a scale, there is no measurement. The output is therefore a pseudo-number: decimals, bounded, and portable, but with no admissible interpretation under the axioms of representation and uniqueness.

Proponents sometimes reply that the utilities are "cardinal enough" for policy or that averaging across respondents confers legitimacy. Neither move works. Averaging ordinal or pseudo-interval scores does not alchemize them into measures; it merely produces a mean of what was never a scale. Claiming "cardinality for policy" is a confession that the standard of truth has been replaced by expedience. Scientific measurement is not made legitimate by intent or usefulness; it is warranted by preserving empirical structure. Here, structure is neither identified nor preserved.

Finally, the entire construction is non-homogeneous dimensionally. The input is a five-dimensional categorical description; the output is a unit-less index masquerading as a quantity. Even if one granted, for the sake of argument, that the algorithm produced a legitimate dimensionless proportion, it would still be incommensurable with duration unless and until its own measurement properties were established as ratio. They are not. Thus, when this pseudo-number is later multiplied by time to form the QALY, the product is undefined in measurement theory. One does not rescue a non-measure by combining it with a measure; one only propagates the error.

In sum, every safeguard that separates measurement from numerology is breached. The EQ-5D utility algorithm violates unidimensionality, ignores cancellation, lacks invariance, invents a zero, fabricates negative values, and provides no unit. It is not a measure of health; it is an index built for rhetorical portability. To treat it as a metric or worse, to build reimbursement policy upon it, is not science but numerical storytelling with decimals.

THE QALY: THE THIRD STEP INTO RELATIVISTIC MATHEMATICAL MADNESS

There was no doubt an air of triumph when the QALY was unveiled, as though the problem of rationing scarce health resources had finally met its mathematical solution. Here, at last, was a formula: the quantity of time lived, discounted by a "utility" score that purported to represent the quality of that life. The result was the quality-adjusted life year, a currency of health outcomes that seemed to place all diseases, all interventions, all lives, on the same ledger. Policy makers, health

economists, and even clinicians could be forgiven for embracing the promise. But this was nothing more than a will-o'-the-wisp, a conjuring trick that led directly to the pseudo-science of the reference case simulation model and its attendant paraphernalia of cost-outcome ratios, probabilistic sensitivity analysis, and imaginary claims for cost-effectiveness. The satisfaction was misplaced because the mathematics was incoherent from the start.

The problem rests on a fundamental category error. Time is, on Stevens' typology, the paradigm of a ratio measure. It has a true zero, equal intervals, and the full range of permissible arithmetic operations. It is one of the most robust and uncontroversial measures in science. But the "utility" with which it was to be multiplied was never a measure at all. It was a pseudo multiattribute ordinal index, derived from an algorithm that had no meaning in measurement theory. Treating these ordinal numbers as if they were interval or ratio measures is exactly the "nonsense" Stevens had warned against in 1946. The arithmetic collapse is obvious once dimensional homogeneity is recalled: a ratio measure can be modified only by another measure that is either unitless, properly scaled, and invariant, or that shares its dimension in a way that preserves the structure of the quantity. Utilities fail on both counts.

What then is a "quality-adjusted life year"? It is not a unit of time, because the utility multiplier is not dimensionless. It is not a unit of health, because no such dimension has been established. It is a hybrid with no empirical warrant, a number with decimals but no meaning. By insisting that a preference-derived index could be laid on top of time to yield a new, compound unit, HTA analysts engaged in numerical invention, not measurement. The satisfaction that accompanied the first QALY tables was not the satisfaction of science meeting reality but of politics acquiring a rhetorical instrument dressed in quantitative garb.

The absurdity compounds when this chimera is embedded in the reference case model. Once utilities and QALYs are admitted as "data," they can be multiplied, summed, and discounted further in elaborate simulations. Cost-per-QALY ratios emerge; probabilistic sensitivity analyses spin out thousands of iterations; thresholds are invoked as if they marked the boundaries of rational policy. Yet every one of these products rests on the original sleight of hand: the illegitimate multiplication of a true ratio measure by an ordinal pseudo-number. No amount of simulation can repair that error. Instead, it only magnifies it, wrapping it in ever thicker layers of false statistical ritual. The result is an edifice of apparent rigor built on a foundation of sand.

If dimensional homogeneity is the non-negotiable standard of science, then the QALY was doomed at birth. Time can be added, compared, discounted, because it is a ratio quantity. Utilities cannot be treated as if they share those properties. Their combination yields not a measure but an illusion, a flight of fancy sustained by convention and repetition rather than empirical warrant. What began as a false step hardened into orthodoxy, and from that error the entire machinery of HTA, the reference case model, cost-outcome ratios, and probabilistic analysis, was born. It is not measurement; it is storytelling with numbers, an exercise in mathematical madness masquerading as science.

For forty years the QALY has not only survived but flourished. A simple PubMed search shows more than 27,000 references, a staggering literature built around what is, at root, a non-measure. The persistence of this construct poses an uncomfortable question: how could such a profound

error, identified clearly by the standards of measurement theory, become the most widely accepted device in health economics? The answer cannot be found in science. It lies instead in sociology, politics, and the psychology of professional consensus. It is impossible to believe that so many people were so easily fooled.

THE REFERENCE CASE:THE FOURTH STEP IN RELATIVISTIC MATHEMATICAL MADNESS

If the QALY represented the conceptual impossibility at the heart of health technology assessment (HTA), the reference case model was its institutional entrenchment. By embedding QALYs into a standardized simulation framework, agencies such as NICE, PBAC, and ICER created an elaborate machinery for generating cost-effectiveness ratios. This machinery has been presented as rigorous, transparent, and indispensable. In reality, it is absurd. It takes as input numbers that are not measures, processes them through assumptions that cannot be tested, and produces outputs that carry no truth-value. Judged against the standards of normal science, the reference case model is a construct outside science.

Science advances by generating claims that can be evaluated as true or false. At its core is falsifiability, the principle articulated by Karl Popper: a claim is scientific only if it could, in principle, be proven wrong. Alongside falsifiability, normal science requires credibility, meaning that methods are transparent and systematic; evaluability, meaning that claims can be assessed against observable evidence; and replicability, meaning that results can be reproduced by independent investigators. Together, these standards safeguard against conjecture, speculation, and self-confirming exercises.

In HTA, these standards are indispensable. If a therapy is said to improve survival or reduce hospitalization, the claim can be tested, replicated, and potentially falsified. If a therapy is said to improve patient-reported quality of life, Rasch-based instruments can provide interval-level measures open to the same scientific discipline. But if a therapy is said to yield “0.3 additional QALYs” at a cost of $50,000 per QALY, no such test is possible. The number is not an empirical claim but an artifact of assumptions.

The reference case model begins with utilities derived from preference elicitation or tariffs. These utilities are multiplied by time to produce QALYs. Costs, expressed in currency, are then divided by these QALYs to yield incremental cost-effectiveness ratios (ICERs). On the surface this looks elegant. By standardizing the method—requiring particular tariffs, discount rates, time horizons, and modeling techniques—the reference case promises consistency and comparability across submissions. Two drugs for different diseases can, it seems, be ranked by their cost per QALY. But this comparability is an illusion. The QALY is not a measure. Its denominator collapses under every axiom of representational measurement theory. The ICER, therefore, is undefined. What appears as a ratio of costs to outcomes is in fact a ratio of costs to non-measures, a number without dimensional coherence.

The outputs of the reference case model are therefore imaginary claims. They cannot be falsified because they do not represent empirical structures. Change the tariff, the discount rate, or the extrapolation horizon, and the ICER changes. Yet no dataset could ever prove the ICER wrong,

because it is not a claim about the world but about the assumptions built into the model. This is the antithesis of science. Scientific claims are provisional precisely because they can be rejected by evidence. The reference case model immunizes itself against such rejection. Its claims are conditional, assumption-dependent, and non-evaluable. They are not hypotheses but numerical stories.

The deeper problem lies in the status of the assumptions themselves. Every element of the reference case, transition probabilities in Markov models, survival curve extrapolations, discounting rules, decrements for health states, rests not on measurement but on induction. The belief that a set of observations can be stretched forward into the future is a fragile one even under the best of circumstances. In the reference case, it is doubly fragile, because the base data, other than clinical inputs, are not measures. Utilities are ordinal preferences masquerading as interval numbers. Extrapolating from them compounds error on error. Induction here is not disciplined by measurement, theory, or falsification; it is simply convention. Analysts assume survival continues according to a fitted curve, costs accumulate according to a chosen horizon, and quality-of-life decrements persist indefinitely. No empirical test can arbitrate between one set of assumptions and another. The projections are untethered from reality, justified only by rhetorical appeal to consistency.

This abuse of induction is compounded by the layering of conventions. Discount rates are imposed not because of any scientific law but because policy makers prefer to weight the present more heavily than the future. Tariffs differ across jurisdictions, producing different ICERs for the same therapy depending on geography. Sensitivity analyses are conducted not on empirical quantities but on the arbitrary assumptions themselves, generating ranges of numbers that look like robustness checks but are only variations on fictions. Probabilistic sensitivity analysis, far from disciplining uncertainty, only formalizes ignorance, assigning probability distributions to assumptions that cannot be verified. Each layer adds an aura of sophistication while leaving the foundations untouched.

The cost side of the equation fares no better. The numerator of the ICER is no more coherent than its denominator. Aggregate costs in the reference case are a dog's breakfast of assumptions about resource use, time spent in different states, and unit costs. Drug acquisition, hospitalization, physician visits, laboratory tests, and ancillary care are all folded together, each based on assumptions about frequency and duration. These aggregated costs have no independent meaning because they cannot be disentangled into verifiable units of therapy impact. A true claim about a therapy's effect must be made in terms of resource units, doses dispensed, days in hospital, outpatient visits avoided. Costs can then be attached separately by any health system for its own accounting. The ICER erases this distinction, presenting a blended numerator that is neither standardized nor scientifically evaluable. Thus both the numerator and denominator of the ratio collapse. One is an aggregate of assumptions, the other a non-measure, and the ratio between them is meaningless.

Defenders of the reference case often point to transparency and replicability. Anyone, they argue, can reproduce the results by applying the same assumptions. But replication of assumptions is not replication of evidence. Two analysts may agree on an ICER not because it represents reality but because they have agreed on the same fictions. What is replicated is the arithmetic of the model,

not the empirical phenomena of health or cost. The defense confuses internal consistency with scientific warrant.

The institutionalization of this model has had profound consequences. Health systems base reimbursement decisions on ICERs, thresholds, and probabilistic sensitivity analyses. Committees deliberate as though these numbers were empirical measures. Thresholds are debated as though they were grounded in dimensional reality. In truth, the exercise is closer to astrology than to science: a formalized system of symbols whose authority rests not on validity but on institutional inertia. Once HTA adopted the QALY, the reference case model was inevitable. A single metric required a standardized framework; otherwise, its arbitrariness would be too obvious. By embedding utilities and QALYs in simulation models, agencies could maintain the appearance of rigor. Markov chains, half-cycle corrections, survival extrapolations, and Monte Carlo simulations all created a façade of technical precision. But the precision is only in the arithmetic. It cannot rescue the absence of measurement.

The inevitability of absurdity is thus clear. Starting from ordinal utilities, ignoring unidimensionality, violating dimensional homogeneity, and compounding these errors with inductive projections and cost aggregates, the reference case model could never yield measures. It produces numbers without truth-value, outputs that cannot be falsified and therefore cannot be scientific. By the time the model was codified, the standards of measurement and science were explicit. Rasch had provided the method for legitimate subjective measurement. Representational measurement theory had codified the axioms. Popper had clarified falsification as the demarcation of science. HTA ignored them all [14].

The collapse is therefore total. From measurement, the reference case fails because its denominator is not a measure and its numerator is an incoherent aggregate. From science, it fails because its outputs are unfalsifiable and its assumptions immune to test. The combination is fatal. HTA's central constructs do not belong in the domain of science. They belong in the domain of numerical storytelling, narratives with decimals, thresholds, and confidence intervals that persuade by form while offering no content. For fifty years HTA has produced claims that cannot be true or false, cannot be replicated in the scientific sense, and cannot guide knowledge. The decisions taken on their basis are decisions made without science, sustained only by the authority of institutions committed to defending the indefensible.

RELATIVISM AND THE SURVIVAL OF HTA

By the 1970s, the intellectual and methodological tools were in place to prevent the construction of pseudo-measures. Yet HTA went ahead. It adopted utilities, constructed the QALY, and embedded these in the reference case model. Each of these was impossible from the standpoint of measurement and incoherent from the standpoint of science. The pressing question, then, is not only why this happened but why it has endured. How could such a patent failure survive for fifty years? Was it simply ignorance of measurement standards? Was it neglect of available knowledge? Or was it something more insidious: the construction of a self-sustaining belief system that thrives by relativism rather than science?

The answer requires a historical progression. In the earliest days of HTA, there is a case to be made for epistemic ignorance. The late 1960s and early 1970s were a moment when decision theory utilities were being popularized, and economists believed they had found a neat way to import preference numbers into policy. The time trade-off and the standard gamble were presented with confidence, as though their outputs could be treated as utilities in the cardinal sense. The attraction lay in the fact that the method generated numbers, decimals that looked like ratios, anchored between zero and one. There was no hesitation in using them as multipliers with time. In retrospect, this looks naïve, but it is plausible that practitioners genuinely did not recognize the measurement issues. They believed they were constructing measures when in fact they were only generating rankings. This was ignorance, though hardly excusable, since Stevens' warning was already three decades old.

By the late 1970s and 1980s, however, the innocence of ignorance was no longer available. Representational measurement theory had been codified in the first volume of *Foundations of Measurement*. Rasch modeling had moved beyond specialist psychometrics into wider circulation. Wright had published his 1977 paper linking Rasch explicitly to the axioms of RMT, making clear that Rasch was the unique model that yielded legitimate measurement. At this point, to proceed with multiattribute indices and utilities was to embrace non-science. The requirements of RMT were accepted. To continue to build health technology assessment on preference indices treated as measures was to ignore what was known.

But neglect soon hardened into something more. The 1980s and 1990s saw the QALY entrenched as the lingua franca of HTA. Agencies were created, journals launched, professional societies established, all with the QALY at their center. The reference case model codified its use. The illusion of rigor was established by demanding probabilistic sensitivity analyses, thresholds, and consistency of assumptions. What had begun as ignorance, and passed through neglect, now became defense of the indefensible. By this stage, practitioners knew that utilities were unstable, that different methods gave different results, that QALYs varied with tariffs and cultures. They knew the arbitrariness of discounting rules and the fragility of extrapolated models. They did not defend these as measures; they defended them as "good enough" for policy.

This shift from aspiring to science to defending fictions marks the embrace of relativism. In its relativist form, HTA no longer seeks to discover truth but to construct consensus. The claim is not that utilities are valid measures, but that they are useful. The QALY is not defended as true but as indispensable. The reference case model is not justified by correspondence to reality but by the comparability it enforces. What counts is not measurement but acceptance. Truth is redefined as consensus; evidence is redefined as whatever the community agrees to treat as evidence.

Here the sociology of scientific knowledge, and in particular the Strong Programme, provides an illuminating lens [15]. The Strong Programme argued that the success of scientific ideas should be explained in the same sociological terms regardless of whether they are true or false. Knowledge is constructed within communities; evidence is not discovered but invented within paradigms. The validity of a program rests not on its correspondence to reality but on its ability to mobilize support. On this view, the persistence of HTA can be explained sociologically rather than epistemically. It survives not because it is true but because it has built a community that treats its fictions as evidence.

The QALY is the archetypal construct of such a community. It offers a simple, communicable number that policymakers can understand. It serves as a common currency across diseases, allowing comparability where otherwise there would be none. It lends itself to thresholds and ratios that give the impression of objectivity. These rhetorical and institutional advantages make it a successful meme in Dawkins' sense: a replicator that thrives not because it is true but because it reproduces effectively [16]. Around it has grown a memeplex: the reference case model, probabilistic sensitivity analyses, incremental cost-effectiveness ratios, thresholds, and the professional societies that sustain them. This memeplex is self-reinforcing, reproducing itself through guidelines, curricula, and career incentives.

Memeplexes survive by defending themselves against critique. They exclude awkward questions, stigmatize dissent, and indoctrinate the next generation. HTA exemplifies all of these. Critiques of the QALY are dismissed as philosophical quibbles, irrelevant to practical policy. Measurement theory is absent from training curricula; students are taught procedures, not foundations. ISPOR and related networks indoctrinate new entrants, presenting QALYs and ICERs as the unquestioned standards. Authority figures reinforce the orthodoxy by their positions in agencies and journals. Careers advance through compliance, not questioning. In this way, the memeplex ensures its own survival.

Wittgenstein's proposition that "at the end of reasons comes persuasion" is apt: rational argument cannot proceed indefinitely; at some point, persuasion (or conversion to a framework) replaces reason [17]. The defenders of HTA do not attempt to prove that utilities are measures or that QALYs are coherent. They persuade instead by repetition, by rhetorical neatness, by the authority of institutions. The numbers look precise, the models look sophisticated, the thresholds look authoritative. Policymakers are persuaded, not because the constructs are true, but because they are presented with rhetorical force. Science has been displaced by persuasion, evidence by rhetoric.

In this relativist frame, the survival of HTA is not puzzling at all. It survives precisely because it is not tied to reality. If evidence is what a community says it is, then utilities and QALYs can be evidence by definition. If truth is consensus, then the agreement of NICE suffices. If science is rhetoric, then the persuasive power of ICER tables and probabilistic sensitivity profiles is enough. In this world, there is no failure to correct, because failure is defined out of existence. The charade is maintained because its maintenance is the very condition of the memeplex survival.

The question then becomes whether this relativist defense can stand indefinitely. Other pseudoscientific memeplexes have survived for decades or centuries: Ptolemaic astronomy with its epicycles, phlogiston theory with its non-existent substance, psychoanalysis with its unfalsifiable constructs. Each persisted because it mobilized a community, enforced consensus, and excluded critique. Each eventually collapsed when new frameworks displaced them. HTA may be unique in that it has been institutionalized by governments and agencies, embedding its fictions in policy. But the logic is the same: it persists not because it corresponds to reality but because it persuades a community to accept it.

From ignorance to neglect to relativist defense, the trajectory of HTA reveals a deeper pathology. At its beginning, there may have been ignorance of measurement standards. Soon there was neglect of known axioms and models. But once institutionalized, HTA became something else: a

memeplex that defends itself against reality. Its survival is not a mystery once relativism is recognized. Evidence in HTA is not discovered in the world but invented within the community. Its success rests not on its ability to generate new knowledge but on its ability to mobilize support. Its defenses are formidable: awkward questions are excluded, measurement is absent from curricula, networks indoctrinate the next generation. At the end of reason comes persuasion, and HTA has relied on persuasion, rhetoric, and authority to maintain its position.

This diagnosis is more troubling than ignorance or neglect. Ignorance can be corrected; neglect can be remedied. But a memeplex built on relativism resists correction. Its survival depends precisely on denying that correction is needed. To admit failure would be to collapse decades of work, discredit agencies, and expose careers. Too much is at stake. Better to defend the indefensible than to concede the truth.

In this light, HTA stands revealed as a world outside science. It borrows the symbols of science, numbers, ratios, thresholds, but it denies the standards of science. It is not falsifiable, not measurable, not empirical. It is sustained by consensus, authority, ignorance, and persuasion. Its outputs are not discoveries but inventions, fictions treated as evidence because a community agrees to treat them so. It is, in Dawkins' sense, a successful memeplex with global reach: a cultural parasite that reproduces itself regardless of truth, infecting institutions and policy, sustained by authority and rhetoric.

The survival of HTA can thus be traced from epistemic ignorance to willful neglect to relativism. What began as naivety hardened into negligence and now persists as a self-reinforcing memeplex. This is why nonsense has survived. It is not because utilities are measures or QALYs coherent; it is because the community of HTA has agreed to treat them as such. The Strong Program explains the logic: success depends not on generating knowledge but on mobilizing support. Wittgenstein reminds us that reason ends in persuasion. Dawkins shows us that memes survive not by truth but by replication. Together, these insights explain why HTA, though impossible as science, has survived as practice.

Judged by the standards of RMT and normal science, the HTA memeplex is clearly in the non-science or pseudo-science camp by denying the possibility of the falsification of claims. But there is a more disquieting possibility: is it bullshit? While the term may appear jarring in an academic context, its use is legitimate and accepted within the philosophy of science, most notably in Harry Frankfurt's seminal 1985 essay *On Bullshit*. Frankfurt distinguished bullshit from both truth and lying [18]. A liar, he argued, is tethered to the truth because in order to lie, one must know the truth yet attempt to conceal it. The bullshitter, by contrast, is indifferent to the truth altogether. Bullshit is characterized not by the intention to deceive about facts but by the indifference to whether a statement is true or false. Its purpose is persuasive, rhetorical, or strategic, unconcerned with empirical warrant.

If we apply this distinction to HTA, the question becomes whether the defenders of utilities, QALYs, and the reference case model are engaged in lying, knowingly presenting false constructs as measures, or whether their enterprise is better understood as bullshit. Certainly, one could argue that fraud is involved: the standards of measurement were explicit by the 1970s, Rasch had demonstrated how ordinal responses could be legitimately transformed, and yet HTA pressed

ahead in defiance of these standards. In that sense, the claim that utilities are cardinal measures, or that ICERs have truth-value, is knowingly false. But Frankfurt's analysis suggests another reading. For many within HTA, the truth-status of utilities and QALYs is simply irrelevant. What matters is their utility as tools of policy, their rhetorical function in providing decision-makers with numbers that appear precise, comparable, and actionable.

This indifference to truth fits the Frankfurtian category of bullshit. When an HTA committee debates whether an intervention is cost-effective at $50,000 per QALY, the central concern is not whether the QALY is a measure, nor whether the ICER is scientifically evaluable, but whether the number performs the function of legitimizing rationing decisions. The QALY and reference case outputs are bullshit not because they are deliberate lies but because their truth-status is beside the point to their users. They persist not through empirical warrant but through institutional convenience, curricular inertia, and rhetorical force.

Whether this distinction between fraud and bullshit is meaningful for HTA depends on how one evaluates culpability. Fraud suggests intentional deception; bullshit suggests cultivated indifference. Both are corrosive to science, but the latter may be more insidious, because it replaces the pursuit of truth with the pursuit of usefulness, collapsing the boundary between evidence and rhetoric. By this standard, HTA is not merely pseudo-science. It is bullshit institutionalized.

CONCLUSIONS

HTA has survived for fifty years by cultivating the illusion of scientific authority while denying the very standards that define science. From its inception it was built on error: the decision to value health states through preference exercises and treat the resulting ordinal rankings as measures guaranteed that its constructs could never meet the axioms of RMT. Utilities were never measures. They lacked unidimensionality, true zeros, interval spacing, and invariance. To multiply them by time in the guise of QALYs was not an approximation but a categorical mistake. The reference case model compounded the error by embedding these pseudo-measures in elaborate simulations, producing ratios that are neither credible nor falsifiable. These devices do not generate knowledge; they generate numbers with the appearance of precision, numbers that cannot be tested against reality.

The analysis presented in Part I has shown that HTA was never built on measurement but on its absence. The foundations of representational measurement theory were in place by the early 1970s. Stevens had drawn the boundary between numbers and measures; Suppes, Luce, and Tukey had codified the axioms that guarantee additivity; Rasch had provided the model that transforms ordinal responses into interval measures; and Wright had made explicit Rasch's unique status as the only model consistent with measurement theory. Against this background, the elevation of utilities to the status of measures and their multiplication with time to create QALYs was indefensible. It was a categorical error, an abandonment of unidimensionality, dimensional homogeneity, and invariance. The subsequent embedding of these pseudo-measures in reference case models only compounded the incoherence.

From the beginning, HTA has defended its methods not on the grounds of measurement legitimacy but on their supposed policy utility. The argument has been simple: health systems must ration

scarce resources; therefore, any device that allows outcomes to be expressed in a single metric is better than none. The QALY and the reference case model, defenders insist, are not perfect, but they are practical. They produce "approximate information" sufficient for policy. This is the rhetoric of relativism. It abandons the idea that numbers must be structure-preserving mappings of empirical relations. It abandons falsification as a scientific standard. It substitutes institutional convenience for epistemic warrant.

Yet for more than forty years this edifice has not only survived but thrived. The QALY has generated over 27,000 PubMed citations, has been institutionalized by NICE and has become the unquestioned lingua franca of HTA. How can a construct that fails every axiom of measurement theory have endured so successfully? Two explanations present themselves.

The first is collective amnesia. The standards of measurement theory were available, but the HTA community ignored or forgot them. Economists and policymakers alike were seduced by the appeal of a single metric and, in their rush to relevance, simply put aside the requirements of science. If this is the explanation, HTA suffers from a profound disciplinary failure: the loss of intellectual memory that every other empirical science preserved.

The second is fraud. On this reading, economists and agencies recognized that utilities and QALYs could never meet the standards of measurement, but adopted them anyway because they were politically expedient. The QALY's pseudo-numbers could be manipulated into cost-effectiveness ratios, thresholds, and rankings that gave the appearance of rigor. It was a conscious deception, perpetuated through guidelines, journals, and curricula, in order to deliver administratively convenient answers.

But perhaps the most compelling diagnosis is Frankfurt's category of bullshit. The hallmark of bullshit is indifference to truth. The bullshitter does not lie, which requires a commitment to the false; rather, he is unconcerned whether what he says is true or false. What matters is the impression created and the function served. This captures HTA with disturbing precision. The architects of the QALY were not primarily concerned with whether it was a measure; they were concerned with whether it could be used. The QALY looks like science, generates ratios, fills textbooks, and provides a language for rationing. Whether it is true or false, a measure or a number, is irrelevant.

This explains the QALY's survival: not as a triumph of science but as the institutionalization of bullshit. HTA has insulated itself from criticism by embedding the QALY/reference case complex in a memeplex sustained by agencies, journals, societies, and curricula. It persists not because it is right, but because it is useful.

The bottom line is that we have known since the 1970s that there are only two admissible forms of measurement on which HTA can base credible claims for therapy response: the linear ratio scale for manifest constructs and the Rasch logit ratio scale for latent constructs. The linear ratio scale applies where attributes vary along a unidimensional continuum with a true zero, permitting all arithmetic operations and proportional comparisons. Time, costs, and resource units fall into this category; their standing is uncontroversial and their structure self-evident. The Rasch logit ratio scale, by contrast, is essential where constructs are latent and must be inferred from ordered

categorical responses. Here the Rasch model provides the only rigorous transformation from ordinal data to interval measurement, securing unidimensionality, invariance, and interval spacing, and yielding constant relative differences in logits. Together, these two approaches exhaust the legitimate options: they cover manifest and latent attributes alike, and they alone preserve empirical structure in a way that supports falsifiable, replicable scientific claims. Everything else is excluded.

Multiattribute indices, utilities, QALYs, and their associated algorithms collapse at the threshold of unidimensionality and dimensional homogeneity. They are not approximations of measures but categorical mistakes that cannot be rehabilitated. To continue to privilege them in guidelines, journals, and decision frameworks is not merely a technical lapse but an abandonment of science itself. If HTA is to regain credibility, it must discard the pseudo-measures and rebuild on these twin foundations. Without them, HTA remains numerical storytelling, not science.

ACKNOWLEDGEMENT

Portions of this education program were drafted and edited with assistance from ChatGPT (version 5; OpenAI). The author reviewed, verified, and refined all AI-assisted text and assumes full responsibility for the accuracy, integrity, and originality of the final content

REFERENCES

[1] Langley P. **M**easurement in HTA: Part 1 From epistemic ignorance to willful neglect. Maimon Working Paper No. 17 September 2025 www.maimonresearch.com

[2] Stevens S. On the Theory of Scales of Measurement. *Science*. 1946;103(2684):677-80

[3] Suppes P. A Set of Independent Axioms for Extensive Quantities. *Portugaliae Mathematica* 1951; 10: 163–172

[4] Luce R, Tukey J. Simultaneous Conjoint Measurement: A New Type of Fundamental Measurement. J *Math Psychol.* 1964*;* 1(1): 1–27

[5] Krantz D, Luce R, Suppes P, Tversky A**.** *Foundations of Measurement*, Volumes I–III. New York: Academic Press, 1971 (Vol. I); 1989 (Vol. II); 1990 (Vol. III).

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

[7] Wright B. "Solving Measurement Problems with the Rasch Model." *Journal of Educational Measurement* 14, no. 2 (1977): 97–116

[8] Torrance G, Thomas J, Sackett D. A Utility Maximization Model for Evaluation of Health Care Programs. *Health Services Research.* 1972; 7(2): 118–133

[9] Torrance G. Social Preferences for Health States: An Empirical Evaluation of Three Measurement Techniques. *Socio-Economic Planning Sciences.* 1976; 10(3): 129–136

[10] Torrance G. Utility Approach to Measuring Health-Related Quality of Life." J *Chronic Diseases.* 1987; 40(6): 593–600

[11] 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

[12] EuroQol Group**.** EuroQol—A New Facility for the Measurement of Health-Related Quality of Life. *Health Policy* 1990;16(3): 199–208.

[13] Dolan Paul. Modeling Valuations for EuroQol Health States. *Medical Care* 1997; 11: 1095-1108.

[14] Popper K. *Objective Knowledge: An Evolutionary Approach*. Revised edition. Oxford: Clarendon Press, 1979

[15] Bloor D. *Knowledge and Social Imagery.* London: Routledge & Kegan Paul, 1976.

[16] Dawkins , R. *The Selfish Gene*. 2nd ed. Oxford: Oxford University Press, 1989.

[17] Wittgenstein L. *On Certainty.* Edited by G. E. M. Anscombe and G. H. von Wright. Translated by Denis Paul and G. E. M. Anscombe. Oxford: Blackwell, 1969. §612 [See also Wootton D. The Invention of Science: A New History of the Scientific Revolution. New York: HarperCollins. 2015. P. 44]
.

[18] Frankfurt H. *On Bullshit.* Princeton, NJ: Princeton University Press, 2005 [Frankfurt H. On Bullshit. *Raritan Quarterly Review*. 1985; 6(2): 81–100].