MAIMON WORKING PAPER No 22 November 2025

THE VENERATION OF THE QALY

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

ABSTRACT

The Quality-Adjusted Life Year (QALY) emerged in the 1970s from efforts by Torrance, Weinstein, and Zeckhauser to create a single construct that combined length and quality of life. Although widely adopted as the central metric for health technology assessment (HTA) and cost-effectiveness analysis, the QALY has no basis in measurement theory and cannot support the arithmetic operations on which its use depends. Its foundation in ordinal, multidimensional preference scores violate the axioms of representational measurement. The underlying utilities are not measures of any defined variable but composite indices derived from value judgments about dissimilar attributes such as mobility, pain, and anxiety. By attempting to multiply these ordinal scores by time, a linear ratio measure, HTA creates a mathematical hybrid that is internally inconsistent and without empirical meaning. The result is not a quantification of health benefit but a numerical fiction: an artifact of misplaced arithmetic masquerading as science.

The introduction of fundamental measurement theory through the work of Stevens, Krantz, Rasch, and Wright exposes this error decisively. Arithmetic operations, including addition, multiplication, and discounting, require a unidimensional scale where equal numerical differences represent equal empirical differences. The EQ-5D-3L and similar instruments are inherently multidimensional, ensuring that no such scale exists. Moreover, when constructs are subjective or latent, the Rasch model provides the only lawful path from observation to measurement. By jointly estimating item difficulty and person ability on a common logit scale, the Rasch framework creates an interval measure that can, by transformation, yield a ratio structure within that latent domain. However, these logits belong to a different mathematical system from linear ratio measures like time and cannot be combined with them.

The conclusion is unavoidable: utilities are not measures, and QALYs are impossible. Once this is understood, the foundations of current HTA practice collapse. The discipline must abandon numerical storytelling based on non-measures and rebuild its evaluative frameworks on lawful standards of measurement, ensuring that every value claim, manifest or latent, is credible, evaluable, and replicable.

1.INTRODUCTION: FAITH IN NUMBERS

The Quality-Adjusted Life Year, or more commonly, the QALY, has become the central icon of health technology assessment. It is the most venerated artefact in the field: an instrument that promises to unite mortality and morbidity in a single scale, to convert human experience into a universal metric of value, and to permit comparative judgments across therapies, diseases, and populations. For more than 40 years it has guided decisions in agencies such as NICE, PBAC, ICER, and CADTH, and its authority has become axiomatic. It no longer needs justification; its

use alone has come to constitute evidence of validity. Yet this is the paradox that demands exposure. The QALY is not a measure. It was never a measure. It is a symbolic construct that gained prestige through repetition and bureaucratic inertia. The problem is not its convenience but its metaphysical pretensions: the illusion that preference scores derived from the time trade-off or standard gamble can bear the weight of arithmetic.

The purpose of this essay is to expose that illusion. The veneration of the QALY has displaced science with ritual. The field that claims to quantify value has become one of the least quantitative in the human sciences. In short, the title "The Veneration of the QALY" conveys a misplaced, almost devotional faith in a measure that cannot meet even the most basic standards of arithmetic or scientific coherence. The purpose here is not to mock belief but to return to measurement; to insist that if numbers are to govern access to care, they must be lawful representations of the empirical world. That simple demand undermines the QALY entirely. Indeed, if one set out to create a believable yet manifestly false construct to support health care decisions and resource allocation, the QALY is the ideal vehicle.

2. THE QALY: DISCOVERED OR INVENTED

If we begin from the position that mathematics is discovered, not made up, then the role of invention is immediately constrained. Discovery gives us the lawful structures: ordered sets, additive structures, ratio systems, and the conditions under which numbers can stand for relations in the world. Invention then has a narrower task: to construct instruments, scales, and applied metrics that preserve those discovered structures when we move from the abstract to the observable [1]. That translation only succeeds if the transformation from the empirical system to the numerical system satisfies the axioms of representational measurement, including the Rasch rules for transforming subjective observations to interval measures

This means we cannot defend a metric simply by saying it is a useful convention. Once we claim that a number measures an attribute, we are bound by the axioms that make measurement possible. The central point is that invention does not license violations of discovery. You cannot take an ordinal ordering of preferences and, by fiat, promote it to an interval or ratio scale. Unless the empirical relations support additivity, solvability, and the relevant cancellation conditions, the numerical assignment is not unique in the required sense, and so it is not a measure. It is only a labeling exercise.

Applied to HTA, this is fatal for the QALY and all its QALY-like variants. The utilities that are meant to represent "quality" are not generated on a structure that supports the transformations required for interval or ratio meaning. When these are multiplied by time, the resulting number is not protected by uniqueness; different admissible representations of the underlying preferences would not preserve the arithmetic. That is the hallmark of a failed translation from discovery to application. In short, any so-called invention in HTA that does not preserve the discovered axioms of measurement is not a lower grade of measure; it is a non-measure, however widely used.

Given the importance pf patient reported outcomes (PROs) in HTA the transformation of subjective observations to interval measures is only possible under strict mathematical

conditions, and the Rasch model is the unique framework that meets them [2]. When we observe subjective responses, statements of agreement, intensity of feeling, or self-reported status, the Rasch rules have to be applied.

We begin with ordinal data: ordered categories without equal intervals. To achieve measurement, we must transform these into a scale where the differences between adjacent points are constant and the origin and unit are invariant. The Rasch model accomplishes this by situating both persons and items on a common latent continuum, expressed in logits, which reflect the relative probabilities of a given response pattern. A transformation that satisfies the axioms of representational measurement ensures the resulting interval scale preserves the discovered mathematical structure. Every other approach to scoring subjective data, such as summing raw responses or assigning arbitrary weights, fails these axioms and cannot claim the status of measurement. Rasch modeling is thus not a statistical option but the necessary bridge that lawfully transforms subjective observations into interval measures, ensuring that the translation from discovery to application retains mathematical legitimacy. Accepting this means that essentially all HTA PRO instruments fail the standards for lawful arithmetic.

3. ORIGINS OF VENERATION

The QALY had to be invented, even if it contradicted the required axioms of representational measurement. The QALY's genealogy is well known. It arose in the 1970s from attempts by Torrance, Weinstein, Zeckhauser and colleagues to combine length and quality of life into a single construct [3] [4] [5] [6]. The logic seemed impeccable: assign a utility to each health state between 0 (death) and 1 (perfect health), multiply it by time, and obtain a ratio measure of benefit. This construction assumed, without evidence, that the utilities were themselves ratio-scaled. The assumption violated every axiom of representational measurement theory. Preferences elicited through trade-offs or lotteries are ordinal; they express order but not magnitude. The operations of multiplication and addition have no meaning on such scales. The QALY, therefore, is the product of two incommensurable quantities: an ordinal index and an interval of time. Its appearance of precision masks a category error.

Stevens' 1946 paper on scale types had already made clear that lawful arithmetic requires unidimensionality, invariance and additivity [7]. Krantz, Luce, Suppes, and Tversky formalized those axioms in *Foundations of Measurement* (1971), showing that measurement is a mapping from an empirical relational system to a numerical one that preserves structure [1]. Importantly, this was supported by Wright in 1977 who showed that for latent traits, the Rasch rules for fitting data to the Rasch logit model articulated these axioms [8] [9].

Nothing in the time trade-off or standard gamble procedures satisfies those requirements. Yet the QALY survived because it was useful. It allowed economists to populate models, to fill the barren matrices of cost-effectiveness analysis with numbers that looked scientific. When NICE institutionalized the reference case, the QALY was enthroned not as a provisional construct but as revealed truth.

3. INSTITUTIONAL CONSECRATION

Once adopted by national agencies, the QALY ceased to be an analytic tool and became a moral standard. Its use defined the boundaries of legitimate discourse. To question it was to question the foundations of HTA itself. The academic ecosystem adjusted accordingly: journals, conferences, and doctoral programs taught students to accept the QALY as measurement. Its composite or multiattribute ordinal foundations were forgotten. The ritual of multiplication by time replaced the discipline of empirical validation. In this process of institutionalization, the QALY became what sociologists of science would call a boundary object, ambiguous enough to unite diverse constituencies, rigid enough to prevent dissent.

At this point, it may be useful to imagine the inner reasoning of the faithful. It is not malicious; it is defensive. Faced with criticism that the QALY violates fundamental measurement, its adherents respond with the weary pragmatism of the clergy who have heard every heresy before. "Yes," they say, "perhaps it is imperfect, but it works. It has been used for decades. It guides resource allocation. It allows comparability." Usage becomes the new epistemology.

At conferences and editorial meetings, one often hears a voice, sometimes ironic, sometimes earnest, that could be condensed into the following communiqué, circulating perhaps among the imagined HTA Illuminati:

> *Dear colleagues in measurement reform, we appreciate your zeal, but you misunderstand the foundations of our practice. The QALY is not merely a tool; it is an institution. It unites policy, research, and morality under one symbol. To abandon it would be to unmake our discipline. You speak of axioms, of additivity, of invariance; we speak of decision thresholds and willingness to pay. The world runs on budgets, not on Rasch logits. If the QALY has guided forty years of policy, must that not attest to its validity? The mere fact of use is proof of worth. Measurement theory is elegant, but we have patients to serve and submissions to review. You cannot ask us to start again. Our models have momentum; our journals depend on them; our reputations rest upon their continuance. The QALY may be ordinal, but so, perhaps, is faith.*

This fictional letter captures the essential defense of the orthodoxy: the appeal to history and utility rather than truth. The argument is circular, it is valid because we use it, and we use it because it is valid; yet it has sustained an entire global industry. What matters is not coherence but continuity. The language of evidence masks an anxiety of loss: the fear that abandoning the QALY would expose four decades of economic modeling as numerology.

Unfortunately, although there have been long-standing criticisms of the QALY and its exclusion from federally funded decision making in the United States, the debate has rarely confronted the underlying issue of lawful measurement. Once the contributions of Stevens, Krantz, Rasch, and Wright are introduced, the measurement cat is well and truly out of the HTA bag. For the first time, there is explicit recognition that any claim to quantify health must be compatible with the

axioms of measurement and, in the case of latent constructs, with Rasch requirements for conjoint, invariant measurement. This shifts the ground entirely. It is no longer a matter of taste, convenience, or policy pragmatism; it is a matter of mathematical admissibility. In that setting, when a defender of QALYs or reference-case modeling appeals to the supposed practical benefits of numerical storytelling, that defense can now be challenged directly on scientific grounds. The field cannot go on pretending that an ordinal, multidimensional, preference-based index can be multiplied by time and presented as if it were a measure. There is, increasingly, an appreciation that the emperor has no clothes.

4. THE GUARANTEE OF MEASUREMENT FAILURE

With its commitment to valuing health state descriptions through preference exercises, HTA built measurement failure into its foundations. The logic is straightforward. Arithmetic operations in health technology assessment, whether for comparison, aggregation, or discounting, presuppose a unidimensional scale where differences have the same meaning across the range. The EQ-5D-3L utility algorithm, however, guarantees multidimensionality because it fuses qualitatively distinct attributes, mobility, self-care, usual activities, pain, and anxiety/depression, into a single index. Once that aggregation occurs, there is no longer a single underlying variable to be measured, so there is no lawful basis for treating the resulting utility as if it were an interval or ratio measure. From an axiomatic perspective, the utility score lacks meaning because it cannot satisfy the conditions for additivity or invariance, and different admissible representations of the underlying preferences would not preserve the numerical relationships. In effect, HTA tried to move directly from ordered preferences over multidimensional health descriptions to arithmetic, bypassing the prior question of whether measurement was even possible. That is why subsequent operations, multiplying by time to form QALYs, discounting over future periods, or comparing across therapies, are mathematically indefensible. They act on numbers that do not map a coherent empirical system. The failure is not in the implementation of the EQ-5D-3L but in the decision to treat multiattribute health descriptions as if they could ever yield a single, legitimate measure.

The deeper problem is that HTA never even reaches Stevens; most practitioners have never heard of him, let alone of representational measurement. The field is therefore not misapplying an established framework but operating in ignorance of the preconditions for lawful arithmetic. When a multiattribute utility such as the EQ-5D-3L is presented as if it were a measure, this is not a matter of interpretation but a category error. There is no recognition that a scale must be unidimensional before anyone can speak meaningfully about intervals, ratios, or admissible transformations. Instead, HTA moves directly from convenience data to cost-per-QALY outputs, bypassing the prior question of whether the numbers correspond to a coherent empirical structure. This is why the standard defense, it is widely used, is scientifically empty. Widespread adoption does not cure a breach of axioms; it merely institutionalizes it. These utilities are not weak or imperfect measures; they were never measures at all, and every QALY constructed from them necessarily inherits that illegitimacy.

Arithmetic operations depend on unidimensionality because without a single underlying dimension, the relations among the quantities being combined have no consistent meaning. In measurement theory, numbers acquire legitimacy only when they represent magnitudes along

one continuum; whether it is length, temperature, pressure, or ability. Unidimensionality guarantees that every numerical difference or ratio corresponds to an identical empirical difference or ratio in the attribute being measured. Addition, subtraction, multiplication, and division only make sense when those operations mirror relationships that exist in the empirical system itself.

If a measure combines more than one dimension, there is no single empirical variable to which the arithmetic refers. Adding or averaging across dimensions such as pain and mobility assumes a common scale of comparison that does not exist. The numerical operations then become symbolic manipulations detached from any real structure. This is why, in both RMT axioms and the equivalent Rasch rules, unidimensionality is the first condition tested: only when item responses or observations reflect variation along one latent variable can we claim that a unit change has the same meaning across the scale.

Arithmetic is therefore not a privilege of convenience but a logical consequence of measurement structure. Once unidimensionality is lost, the mapping from empirical to numerical relations breaks down; equal differences in numbers no longer signify equal differences in the construct. What remains is calculation without correspondence; numbers that look quantitative but have no lawful connection to what they purport to measure.

5. DISCOUNTING TIME

You cannot discount time by multiplying with a multidimensional scale. When such a composite score is multiplied by time, the result, “quality-adjusted life years” has no mathematical meaning, because there is no single underlying construct being measured. It is like multiplying “height plus apples” by time: the operation is undefined. Only unidimensional, interval or ratio measures allow arithmetic manipulation. Without unidimensionality, the numbers are simply labels or rankings, not quantities that can be multiplied or averaged in any lawful scientific sense.

You can, of course, multiply by a dimensionless number when the underlying scale has valid interval or ratio properties, because the operation preserves meaning. A ratio scale possesses a true zero and equal intervals across its range, allowing multiplication or division to express proportional change. For example, multiplying a time measure by 2.0 doubles duration; multiplying a dose by 0.5 halves it. The dimensionless multiplier functions as a scalar, not a measure in itself. This is only lawful when the target quantity is unidimensional and the zero point represents total absence of the property being measured. Problems arise when the “number” used is not dimensionless in this sense but rather an ordinal or composite score, as with utility values. Utilities lack a true zero, are bounded between 0 and 1, and represent heterogeneous attributes, so they cannot serve as lawful multipliers. Multiplying such values by time produces a hybrid quantity without coherent dimensional meaning. Lawful multiplication thus depends on both the dimensional purity of the base measure and the unbounded, ratio-scale nature of the multiplier; a condition the QALY’s utility component manifestly fails to meet.

6. MANIPULATING QALYS

Manipulating QALYs is impossible because the QALY does not meet the requirements for lawful arithmetic operations. This should be obvious, yet it remains one of the most persistent misconceptions in health technology assessment. Arithmetic can only be performed on data measured on interval or ratio scales; scales where differences and proportions have consistent meaning. The QALY, however, is constructed from utility scores derived from time trade-off or standard gamble tasks, which generate only ordinal data. These scores merely rank health states; they do not quantify the magnitude of difference between them. Consequently, adding, subtracting, multiplying, or averaging these numbers has no mathematical or empirical meaning. When an ordinal utility score is multiplied by time, the result is not a measurable quantity but a numerical fiction; a product of non-comparable dimensions. This is akin to multiplying a person's shoe size by their age and claiming it represents "foot-years." The operation produces a number but not a measure. The illusion of precision masks the absence of measurement structure. Until health technology assessment recognizes that the QALY rests on ordinal data masquerading as interval measures, any manipulation, discounting, averaging, summation, will remain mathematically indefensible and scientifically incoherent.

7. THE COST OF FALSE BELIEF

The persistence of the QALY has consequences beyond academic debate. It shapes access to care, pricing, and innovation. By treating an ordinal index as a ratio measure, HTA agencies create thresholds that appear objective but are in fact arbitrary. A drug priced at $150,000 per QALY may be rejected not because it fails to deliver measurable benefit but because the measure itself is fictive. This is pseudoscience institutionalized: decisions justified by metrics that cannot be replicated or falsified. The absence of falsifiability, as Popper observed, marks the boundary between science and myth.

The cost is epistemological as well as moral. When analysts build simulation models on QALY foundations, they propagate uncertainty exponentially. Each parameter is drawn from studies that themselves rest on non-interval data. The resulting cost-effectiveness ratios are numerical storytelling; quantitative theater performed for regulatory audiences. The problem is not modeling per se but its pretense to measurement. As long as the dependent variable is illegitimate, no refinement of the independent variables can redeem the model. Garbage, even when simulated in ten million iterations, remains garbage.

The damage extends to education. Entire graduate programs teach students to manipulate QALYs without once addressing the axioms of measurement. They learn to run probabilistic sensitivity analyses but not to ask whether the underlying numbers can be added. In this sense, the QALY has functioned as an instrument of epistemic control. It has defined the curriculum of ignorance. The young are trained to worship at the altar of the model rather than to question its metaphysics.

There are, of course, significant legal implications if a decision is believed to have been based, even in part, on constructs such as the QALY or on derivative reference-case cost-effectiveness models. The application of cost-per-QALY thresholds compounds the problem because it gives

the appearance of objectivity while resting on numerically meaningless foundations. Once it is established that the underlying utilities are ordinal, multidimensional, and fail the axioms required for arithmetic, any decision justified by those figures becomes vulnerable to legal challenge. A rejected formulary submission or an imposed price cap could be contested on the grounds that the evidence used was not merely flawed but scientifically indefensible; unsupported by valid measurement.

It would take only one high-profile lawsuit to expose this weakness. A manufacturer denied reimbursement or a patient group excluded from access could argue that the agency's methodology rests on a non-measure, making the decision irrational or procedurally unlawful. In the United States, this could invoke administrative law principles of arbitrary and capricious action; in other jurisdictions, similar standards of reasonableness and due process apply. Once a court hears expert testimony explaining that QALYs and their associated models cannot support arithmetic operations and therefore cannot produce valid comparisons, the credibility of the entire framework could unravel. Even a single judgment acknowledging that cost-per-QALY reasoning lacks scientific legitimacy would have cascading consequences for HTA practice, compelling agencies and payers to abandon these pseudo-quantitative methods in favor of protocols grounded in lawful measurement.

8. CONCLUSION: THE RETURN TO SCIENCE

To end veneration is not to end valuation claims. Health systems require evidence-based decisions, but those decisions must rest on measurement. The alternative to the QALY is not chaos but clarity. Claims must be divided into two classes: manifest and latent. Manifest claims concern observable quantities; these are direct measures on ratio scales. Latent claims concern subjective experience, pain, fatigue, need fulfillment, satisfaction. These require Rasch measurement to transform ordinal responses into interval logits. In both cases, the criterion is the same: unidimensionality, linearity, and invariance.

Every value claim must be accompanied by a protocol that specifies its empirical basis, population, timeframe, and analytic method. Reproducibility replaces assumption; measurement replaces modeling. The role of agencies such as NICE and ICER should be not to compute cost per QALY but to evaluate whether proposed claims meet these standards of fundamental measurement. Only then can HTA become a science rather than a belief system.

The reform will not be easy. Institutions built on the QALY will resist, citing the authority of tradition and the inertia of policy. They will invoke the argument of the HTA Illuminati: "We cannot start again." But science is precisely the art of starting again. When a construct fails, it is abandoned, not worshiped. The shift from veneration to verification will require courage, but the alternative is continued intellectual stagnation.

The QALY's endurance is testimony not to its validity but to the power of repetition. It has survived because it served administrative needs and provided the illusion of comparability. Yet its arithmetic is unlawful, its assumptions unfounded, and its consequences profound. To continue to use it is to perpetuate pseudoscience. The time has come to acknowledge what measurement theory has always required: numbers are not measures unless they map the

structure of reality. The QALY does not. It is an idol carved from convenience, gilded by usage, and enthroned by habit.

If HTA is to regain scientific legitimacy, it must dismantle the shrine it built to the QALY. The path forward is not incremental adjustment but epistemic reformation; a return to the foundations of measurement, to the recognition that arithmetic demand's structure, and that only through structure can evidence acquire meaning. The QALY was never a measure; it was a belief system. Science begins when belief yields to measurement.

ACKNOWLEDGEMENT

The author acknowledges the use of ChatGPT (version 5, OpenAI) in drafting and editing portions of this paper. All AI-assisted text was reviewed, verified, and substantively revised by the author, who assumes full responsibility for the final content and interpretation.

REFERENCES

[1] Krantz, David H., R. Duncan Luce, Patrick Suppes, and Amos Tversky. 1971. *Foundations of Measurement, Volume I: Additive and Polynomial Representations*. New York: Academic Press

[2] Bond T, Zhiqiang Yan, Heene M *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. 4th ed. New York: Routledge, 2021

[3] Zeckhauser R, Shepard. D. Where Now for Saving Lives? *Law and Contemporary Problems* 1976; 0 (4): 5–45

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

[5] Weinstein M, Stason W. Foundations of Cost-Effectiveness Analysis for Health and Medical Practices. *New England Journal of Medicine*. 1976; 296 (13): 716–721

[6] Torrance G, Feeny D. Utilities and Quality-Adjusted Life Years. *International Journal of Technology Assessment in Health Care.* 1989; 5 (4): 559–575.

[7] Stevens S. On the Theory of Scales of Measurement. *Science.* 1946; 103 (2684): 677–680

[8] Rasch, Georg. 1980 [1960]. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: University of Chicago Press

[9] Wright B. Solving Measurement Problems with the Rasch Model. *J Educational Measurement.* 1977; 14 (2): 97–116