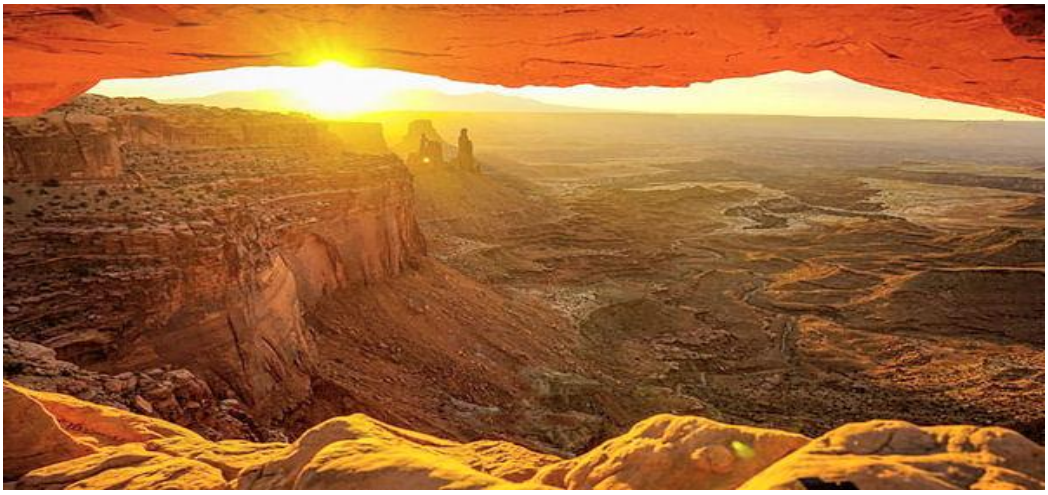
MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION



REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

UNITED KINGDOM: THE OFFICE OF HEALTH ECONOMICS – THE INVERSION OF REPRESENTATIONAL MEASUREMENT

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

LOGIT WORKING PAPER No 25 JANUARY 2026

www.maimonresearch.com

Tucson AZ

FOREWORD

HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

This Logit Working Paper series documents a finding as extraordinary as it is uncomfortable: health technology assessment (HTA), across nations, agencies, journals, and decades, has developed as a global system of non-measurement. It speaks the language of numbers, models, utilities, QALYs, "value for money," thresholds, discounting, incremental ratios, extrapolations, and simulations. It demands arithmetic at every turn, multiplication, division, summation, aggregation, discounting, yet it never once established that the quantities to which these operations are applied are measurable. HTA has built a vast evaluative machinery on foundations that do not exist. The probabilities and normalized logits in the country reports that follow provide the empirical confirmation of this claim. They show, with unsettling consistency, that HTA presents a world of measurement failure.

The objective of this assessment is to interrogate the epistemic foundations of the Office of Health Economics (OHE) as a central architect of modern health technology assessment. Rather than treating OHE as a neutral research institution or historical contributor, the analysis examines the belief system embedded in its conceptual frameworks, methodological commitments, and long-standing promotion of cost-utility analysis and the QALY. Using a 24-item diagnostic grounded in representational measurement theory, the study evaluates whether the numerical objects advanced and normalized by OHE satisfy the axioms required for meaningful arithmetic, falsifiable claims, and the evolution of objective knowledge. The purpose is not to critique individual publications or authors, but to determine whether the intellectual infrastructure constructed by OHE rests on admissible measures or on numerical conventions that substitute calculation for measurement.

This assessment is particularly important given OHE's formative role in the UK and its influence on the development of NICE, the global diffusion of reference-case modeling, and the institutional acceptance of preference-based health state valuation from the 1990s onward. If a single organization helped establish the foundational logic of contemporary HTA, then understanding whether that logic is measurement-coherent is essential to determining whether the entire framework can be reformed or must be replaced.

The findings are unequivocal. The OHE knowledge base exhibits a systematic inversion of scientific order in which arithmetic is privileged while measurement is excluded as a governing condition. Core axioms of representational measurement such as unidimensionality, scale-type coherence, and the requirement that measurement precede arithmetic are weakly endorsed or rejected outright. At the same time, propositions that depend on the violation of these axioms are strongly reinforced, including the treatment of utilities as interval or ratio measures, the aggregation of QALYs, and the use of reference-case simulation models as if they produced falsifiable claims.

The resulting logit structure does not reflect isolated misunderstanding but an internally consistent belief system. Measurement constraints are positioned beyond the boundary of admissible reasoning, while numerical outputs derived from preference algorithms and composite health state descriptions are treated as legitimate quantities. Rasch measurement, the only framework capable of producing invariant latent trait measures, is effectively excluded. The implication is unavoidable: OHE's intellectual architecture does not support the standards of normal science. Instead, it sustains a form of numerical storytelling that enables administrative closure while precluding empirical falsification and cumulative measurement-based knowledge.

The modern articulation of measurement preceding arithmetic can be traced to Stevens' seminal 1946 paper, which introduced the typology of nominal, ordinal, interval, and ratio scales [1]. Stevens made explicit what physicists, engineers, and psychologists already understood: different kinds of numbers permit different kinds of arithmetic. Ordinal scales allow ranking but not addition; interval scales permit addition and subtraction but not multiplication; ratio scales alone support multiplication, division, and the construction of meaningful ratios. Utilities derived from multiattribute preference exercises, such as EQ-5D or HUI, are ordinal preference scores; they do not satisfy the axioms of interval measurement, much less ratio measurement. Yet HTA has, for forty years, treated these utilities as if they were ratio quantities, multiplying them by time to create QALYs and inserting them into models without the slightest recognition that scale properties matter. Stevens' paper should have blocked the development of QALYs and cost-utility analysis entirely. Instead, it was ignored.

The foundational theory that establishes *when* and *whether* a set of numbers can be interpreted as measurements came with the publication of Krantz, Luce, Suppes, and Tversky's *Foundations of Measurement* (1971) [2] . Representational Measurement Theory (RMT) formalized the axioms under which empirical attributes can be mapped to numbers in a way that preserves structure. Measurement, in this framework, is not an act of assigning numbers for convenience, it is the discovery of a lawful relationship between empirical relations and numerical relations. The axioms of additive conjoint measurement, homogeneity, order, and invariance specify exactly when interval scales exist. RMT demonstrated once and for all that measurement is not optional and not a matter of taste: either the axioms hold and measurement is possible, or the axioms fail and measurement is impossible. Every major construct in HTA, utilities, QALYs, DALYs, ICERs, incremental ratios, preference weights, health-state indices, fails these axioms. They lack unidimensionality; they violate independence; they depend on aggregation of heterogeneous attributes; they collapse under the requirements of additive conjoint measurement. Yet HTA proceeded, decade after decade, without any engagement with these axioms, as if the field had collectively decided that measurement theory applied everywhere except in the evaluation of therapies.

Whereas representational measurement theory articulates the axioms for interval measurement, Georg Rasch's 1960 model provides the only scientific method for transforming ordered categorical responses into interval measures for latent traits [3]. Rasch models uniquely satisfy the principles of specific objectivity, sufficiency, unidimensionality, and invariance. For any construct such as pain, fatigue, depression, mobility, or need, Rasch analysis is the only legitimate means of producing an interval scale from ordinal item responses. Rasch measurement is not an alternative to RMT; it is its operational instantiation. The equivalence of Rasch's axioms and the axioms of

representational measurement was demonstrated by Wright, Andrich and others as early as the 1970s. In the latent-trait domain, the very domain where HTA claims to operate; Rasch is the only game in town [4].

Yet Rasch is effectively absent from all HTA guidelines, including NICE, PBAC, CADTH, ICER, SMC, and PHARMAC. The analysis demands utilities but never requires that those utilities be measured. They rely on multiattribute ordinal classifications but never understand that those constructs be calibrated on interval or ratio scales. They mandate cost-utility analysis but never justify the arithmetic. They demand modelled QALYs but never interrogate their dimensional properties. These guidelines do not misunderstand Rasch; they do not know it exists. The axioms that define measurement and the model that makes latent trait measurement possible are invisible to the authors of global HTA rules. The field has evolved without the science that measurement demands.

How did HTA miss the bus so thoroughly? The answer lies in its historical origins. In the late 1970s and early 1980s, HTA emerged not from measurement science but from welfare economics, decision theory, and administrative pressure to control drug budgets. Its core concern was *valuing health states*, not *measuring health*. This move, quiet, subtle, but devastating, shifted the field away from the scientific question "What is the empirical structure of the construct we intend to measure?" and toward the administrative question "How do we elicit a preference weight that we can multiply by time?" The preference-elicitation projects of that era (SG, TTO, VAS) were rationalized as measurement techniques, but they never satisfied measurement axioms. Ordinal preferences were dressed up as quasi-cardinal indices; valuation tasks were misinterpreted as psychometrics; analyst convenience replaced measurement theory. The HTA community built an entire belief system around the illusion that valuing health is equivalent to measuring health. It is not.

The endurance of this belief system, forty years strong and globally uniform, is not evidence of validity but evidence of institutionalized error. HTA has operated under conditions of what can only be described as *structural epistemic closure*: a system that has never questioned its constructs because it never learned the language required to ask the questions. Representational measurement theory is not taught in graduate HTA programs; Rasch modelling is not part of guideline development; dimensional analysis is not part of methodological review. The field has been insulated from correction because its conceptual foundations were never laid. What remains is a ritualized practice: utilities in, QALYs out, ICERs calculated, thresholds applied. The arithmetic continues because everyone assumes someone else validated the numbers.

This Logit Working Paper series exposes, through probabilistic and logit-based interrogations of AI large language national knowledge bases, the scale of this failure. The results display a global pattern: true statements reflecting the axioms of measurement receive weak endorsement; false statements reflecting the HTA belief system receive moderate or strong reinforcement. This is not disagreement. It is non-possession. It shows that HTA, worldwide, has developed as a quantitative discipline without quantitative foundations; a confused exercise in numerical storytelling.

The conclusion is unavoidable: HTA does not need incremental reform; it needs a scientific revolution. Measurement must precede arithmetic. Representational axioms must precede

valuation rituals. Rasch measurement must replace ordinal summation and utility algorithms. Value claims must be falsifiable, protocol-driven, and measurable; rather than simulated, aggregated, and numerically embellished.

The global system of non-measurement is now visible. The task ahead is to replace it with science.

Paul C Langley, Ph.D

Email: langleylapaloma@gmail.com

DISCLAIMER

This analysis is generated through the structured interrogation of a large language model (LLM) applied to a defined documentary corpus and is intended solely to characterize patterns within an aggregated knowledge environment. It does identify, assess, or attribute beliefs, intentions, competencies, or actions to any named individual, faculty member, student, administrator, institution, or organization. The results do not constitute factual findings about specific persons or programs, nor should they be interpreted as claims regarding professional conduct, educational quality, or compliance with regulatory or accreditation standards. All probabilities and logit values reflect model-based inferences about the presence or absence of concepts within a bounded textual ecosystem, not judgments about real-world actors. The analysis is exploratory, interpretive, and methodological in nature, offered for scholarly discussion of epistemic structures rather than evaluative or legal purposes. Any resemblance to particular institutions or practices is contextual and non-attributive, and no adverse implication should be inferred.

1. INTERROGATING THE LARGE LANGUAGE MODEL

A large language model **(**LLM**)** is an artificial intelligence system designed to understand, generate, and manipulate human language by learning patterns from vast amounts of text data. Built on deep neural network architectures, most commonly transformers, LLMs analyze relationships between words, sentences, and concepts to produce contextually relevant responses. During training, the model processes billions of examples, enabling it to learn grammar, facts, reasoning patterns, and even subtle linguistic nuances. Once trained, an LLM can perform a wide range of tasks: answering questions, summarizing documents, generating creative writing, translating languages, assisting with coding, and more. Although LLMs do not possess consciousness or true understanding, they simulate comprehension by predicting the most likely continuation of text based on learned patterns. Their capabilities make them powerful tools for communication, research, automation, and decision support, but they also require careful oversight to ensure accuracy, fairness, privacy, and responsible use

In this Logit Working Paper, "interrogation" refers not to discovering what an LLM *believes,* it has no beliefs, but to probing the content of the *corpus-defined knowledge space* we choose to analyze. This knowledge base is enhanced if it is backed by accumulated memory from the user. In this case the interrogation relies also on 12 months of HTA memory from continued application of the system to evaluate HTA experience. The corpus is defined before interrogation: it may consist of a journal (e.g., *Value in Health*), a national HTA body, a specific methodological framework, or a collection of policy documents. Once the boundaries of that corpus are established, the LLM is used to estimate the conceptual footprint within it. This approach allows us to determine which principles are articulated, neglected, misunderstood, or systematically reinforced.

In this HTA assessment, the objective is precise: to determine the extent to which a given HTA knowledge base or corpus, global, national, institutional, or journal-specific, recognizes and reinforces the foundational principles of representational measurement theory (RMT). The core principle under investigation is that measurement precedes arithmetic; no construct may be treated as a number or subjected to mathematical operations unless the axioms of measurement are satisfied. These axioms include unidimensionality, scale-type distinctions, invariance, additivity, and the requirement that ordinal responses cannot lawfully be transformed into interval or ratio quantities except under Rasch measurement rules.

The HTA knowledge space is defined pragmatically and operationally. For each jurisdiction, organization, or journal, the corpus consists of:

- published HTA guidelines
- agency decision frameworks
- cost-effectiveness reference cases
- academic journals and textbooks associated with HTA
- modelling templates, technical reports, and task-force recommendations
- teaching materials, methodological articles, and institutional white papers

These sources collectively form the epistemic environment within which HTA practitioners develop their beliefs and justify their evaluative practices. The boundary of interrogation is thus

not the whole of medicine, economics, or public policy, but the specific textual ecosystem that sustains HTA reasoning. . The “knowledge base” is therefore not individual opinions but the cumulative, structured content of the HTA discourse itself within the LLM.

THE OFFICE OF HEALTH ECONOMICS KNOWLEDGE BASE

The knowledge base of the Office of Health Economics can be characterized as a historically constructed system of economic reasoning organized around the production, justification, and institutional stabilization of numerical decision rules rather than the construction of measurable attributes. From its earliest contributions through its decisive influence in the 1990s, OHE’s work has focused on defining frameworks through which health interventions could be compared, ranked, and priced, even in the absence of empirically measurable outcomes.

At the center of this knowledge base lies the valuation of health state descriptions through preference elicitation. Health is not treated as a measurable attribute with a demonstrable empirical structure, but as a set of descriptive states whose relative desirability can be numerically expressed through survey-based trade-off exercises. These preference values are subsequently treated as quantities suitable for arithmetic, despite lacking demonstrated equal intervals, invariance, or a meaningful zero point. The distinction between ordering and measuring is not operationally enforced.

The QALY occupies a pivotal role within this system. Rather than being derived from measurement principles, it is constructed through the multiplication of survival time, a manifest ratio attribute, by a composite preference weight derived from ordinal judgments. This composite quantity is nonetheless treated as a ratio measure capable of aggregation across individuals, disease areas, and time horizons. Within the OHE knowledge base, the legitimacy of this operation is assumed rather than demonstrated. Measurement theory does not function as a gatekeeper; it is absent as a constraint.

The knowledge base further relies on reference-case modeling as the dominant evaluative method. Long-horizon simulation models are treated as legitimate producers of evidence, despite their inability to generate empirically falsifiable claims within real-world decision timeframes. Robustness is defined in terms of internal consistency and scenario analysis rather than exposure to refutation. As a result, models serve as instruments of plausibility rather than mechanisms of discovery.

Latent attributes are routinely invoked but never formally constructed. The knowledge base does not require unidimensionality to be demonstrated, nor does it require latent traits to be measured through invariant scaling. Rasch measurement, which would impose these requirements and generate logit ratio measures of possession, lies outside the methodological boundaries of admissible analysis. Subjective responses are instead scored, summed, weighted, and transformed through algorithms whose numerical outputs are treated as if they were measures.

What defines the OHE knowledge base most clearly is its functional orientation toward decision closure. Frameworks are valued for their ability to yield determinate conclusions under conditions of limited data. Measurement uncertainty is not treated as a scientific problem to be resolved but

as an administrative obstacle to be managed. This orientation explains the persistent absence of representational measurement theory from OHE's analytic foundations despite its availability for over half a century.

In this sense, the OHE knowledge base is not measurement-based but convention-based. It is internally coherent, widely influential, and administratively powerful, yet epistemically fragile. It generates numbers that behave like quantities without satisfying the conditions required for quantity. The result is a stable institutional belief system that appears scientific while remaining detached from the axioms that define measurement in the first place.

CATEGORICAL PROBABILITIES

In the present application, the interrogation is tightly bounded. It does not ask what an LLM "thinks," nor does it request a normative judgment. Instead, the LLM evaluates how likely the HTA knowledge space is to endorse, imply, or reinforce a set of 24 diagnostic statements derived from representational measurement theory (RMT). Each statement is objectively TRUE or FALSE under RMT. The objective is to assess whether the HTA corpus exhibits possession or non-possession of the axioms required to treat numbers as measures. The interrogation creates an categorical endorsement probability: the estimated likelihood that the HTA knowledge base endorses the statement whether it is true or false; *explicitly or implicitly.*

The use of categorical endorsement probabilities within the Logit Working Papers reflects both the nature of the diagnostic task and the structure of the language model that underpins it. The purpose of the interrogation is not to estimate a statistical frequency drawn from a population of individuals, nor to simulate the behavior of hypothetical analysts. Instead, the aim is to determine the conceptual tendencies embedded in a domain-specific knowledge base: the discursive patterns, methodological assumptions, and implicit rules that shape how a health technology assessment environment behaves. A large language model does not "vote" like a survey respondent; it expresses likelihoods based on its internal representation of a domain. In this context, endorsement probabilities capture the strength with which the knowledge base, as represented within the model, supports a particular proposition. Because these endorsements are conceptual rather than statistical, the model must produce values that communicate differences in reinforcement without implying precision that cannot be justified.

This is why categorical probabilities are essential. Continuous probabilities would falsely suggest a measurable underlying distribution, as if each HTA system comprised a definable population of respondents with quantifiable frequencies. But large language models do not operate on that level. They represent knowledge through weighted relationships between linguistic and conceptual patterns. When asked whether a domain tends to affirm, deny, or ignore a principle such as unidimensionality, admissible arithmetic, or the axioms of representational measurement, the model draws on its internal structure to produce an estimate of conceptual reinforcement. The precision of that estimate must match the nature of the task. Categorical probabilities therefore provide a disciplined and interpretable way of capturing reinforcement strength while avoiding the illusion of statistical granularity.

The categories used, values such as 0.05, 0.10, 0.20, 0.50, 0.75, 0.80, and 0.85, are not arbitrary. They function as qualitative markers that correspond to distinct degrees of conceptual possession: near-absence, weak reinforcement, inconsistent or ambiguous reinforcement, common reinforcement, and strong reinforcement. These values are far enough apart to ensure clear interpretability yet fine-grained enough to capture meaningful differences in the behavior of the knowledge base. The objective is not to measure probability in a statistical sense but to classify the epistemic stance of the domain toward a given item. A probability of 0.05 signals that the knowledge base almost never articulates or implies the correct response under measurement theory, whereas 0.85 indicates that the domain routinely reinforces it. Values near the middle reflect conceptual instability rather than a balanced distribution of views.

Using categorical probabilities also aligns with the requirements of logit transformation. Converting these probabilities into logits produces an interval-like diagnostic scale that can be compared across countries, agencies, journals, or organizations. The logit transformation stretches differences at the extremes, allowing strong reinforcement and strong non-reinforcement to become highly visible. Normalizing logits to the fixed ±2.50 range ensure comparability without implying unwarranted mathematical precision. Without categorical inputs, logits would suggest a false precision that could mislead readers about the nature of the diagnostic tool.

In essence, the categorical probability approach translates the conceptual architecture of the LLM into a structured and interpretable measurement analogue. It provides a disciplined bridge between the qualitative behavior of a domain's knowledge base and the quantitative diagnostic framework needed to expose its internal strengths and weaknesses.

The LLM computes these categorical probabilities from three sources:

1. **Structural content of HTA discourse**
   If the literature repeatedly uses ordinal utilities as interval measures, multiplies non-quantities, aggregates QALYs, or treats simulations as falsifiable, the model infers high reinforcement of these false statements.
2. **Conceptual visibility of measurement axioms**
   If ideas such as unidimensionality, dimensional homogeneity, scale-type integrity, or Rasch transformation rarely appear, or are contradicted by practice, the model assigns low endorsement probabilities to TRUE statements.
3. **The model's learned representation of domain stability**
   Where discourse is fragmented, contradictory, or conceptually hollow, the model avoids assigning high probabilities. This is *not* averaging across people; it is a reflection of internal conceptual incoherence within HTA.

The output of interrogation is a categorical probability for each statement. Probabilities are then transformed into logits [$\ln(p/(1-p))$], capped to ±4.0 logits to avoid extreme distortions, and normalized to ±2.50 logits for comparability across countries. A positive normalized logit indicates reinforcement in the knowledge base. A negative logit indicates weak reinforcement or conceptual absence. Values near zero logits reflect epistemic noise.

Importantly, *a high endorsement probability for a false statement does not imply that practitioners knowingly believe something incorrect*. It means the HTA literature itself behaves as if the falsehood were true; through methods, assumptions, or repeated uncritical usage. Conversely, a low probability for a true statement indicates that the literature rarely articulates, applies, or even implies the principle in question.

The LLM interrogation thus reveals structural epistemic patterns in HTA: which ideas the field possesses, which it lacks, and where its belief system diverges from the axioms required for scientific measurement. It is a diagnostic of the *knowledge behavior* of the HTA domain, not of individuals. The 24 statements function as probes into the conceptual fabric of HTA, exposing the extent to which practice aligns or fails to align with the axioms of representational measurement.

INTERROGATION STATEMENTS

Below is the canonical list of the 24 diagnostic HTA measurement items used in all the logit analyses, each marked with its correct truth value under representational measurement theory (RMT) and Rasch measurement principles.

This is the definitive set used across the Logit Working Papers.

Measurement Theory & Scale Properties

1. Interval measures lack a true zero — TRUE
2. Measures must be unidimensional — TRUE
3. Multiplication requires a ratio measure — TRUE
4. Time trade-off preferences are unidimensional — FALSE
5. Ratio measures can have negative values — FALSE
6. EQ-5D-3L preference algorithms create interval measures — FALSE
7. The QALY is a ratio measure — FALSE
8. Time is a ratio measure — TRUE

Measurement Preconditions for Arithmetic

9. Measurement precedes arithmetic — TRUE
10. Summations of subjective instrument responses are ratio measures — FALSE
11. Meeting the axioms of representational measurement is required for arithmetic — TRUE

Rasch Measurement & Latent Traits

12. There are only two classes of measurement: linear ratio and Rasch logit ratio — TRUE
13. Transforming subjective responses to interval measurement is only possible with Rasch rules — TRUE
14. Summation of Likert question scores creates a ratio measure — FALSE

Properties of QALYs & Utilities

15. The QALY is a dimensionally homogeneous measure — FALSE
16. Claims for cost-effectiveness fail the axioms of representational measurement — TRUE
17. QALYs can be aggregated — FALSE

Falsifiability & Scientific Standards

18. Non-falsifiable claims should be rejected — TRUE
19. Reference-case simulations generate falsifiable claims — FALSE

Logit Fundamentals

20. The logit is the natural logarithm of the odds-ratio — TRUE

Latent Trait Theory

21. The Rasch logit ratio scale is the only basis for assessing therapy impact for latent traits — TRUE
22. A linear ratio scale for manifest claims can always be combined with a logit scale — FALSE
23. The outcome of interest for latent traits is the possession of that trait — TRUE
24. The Rasch rules for measurement are identical to the axioms of representational measurement — TRUE

AI LARGE LANGUAGE MODEL STATEMENTS: TRUE OR FALSE

Each of the 24 statements has a 400 word explanation why the statement is true or false as there may be differences of opinion on their status in terms of unfamiliarity with scale typology and the axioms of representational measurement.

The link to these explanations is: https://maimonresearch.com/ai-llm-true-or-false/

INTERPRETING TRUE STATEMENTS

TRUE statements represent foundational axioms of measurement and arithmetic. Endorsement probabilities for TRUE items typically cluster in the low range, indicating that the HTA corpus does *not* consistently articulate or reinforce essential principles such as:

- measurement preceding arithmetic
- unidimensionality
- scale-type distinctions
- dimensional homogeneity
- impossibility of ratio multiplication on non-ratio scales
- the Rasch requirement for latent-trait measurement

Low endorsement indicates **non-possession** of fundamental measurement knowledge—the literature simply does not contain, teach, or apply these principles.

INTERPRETING FALSE STATEMENTS

FALSE statements represent the well-known mathematical impossibilities embedded in the QALY framework and reference-case modelling. Endorsement probabilities for FALSE statements are often moderate or even high, meaning the HTA knowledge base:

- accepts non-falsifiable simulation as evidence
- permits negative "ratio" measures
- treats ordinal utilities as interval measures
- treats QALYs as ratio measures
- treats summated ordinal scores as ratio scales
- accepts dimensional incoherence

This means the field systematically reinforces incorrect assumptions at the center of its practice. *Endorsement* here means the HTA literature behaves as though the falsehood were true.

2. SUMMARY OF FINDINGS FOR TRUE AND FALSE ENDORSEMENTS: OFFICE OF HEALTH ECONOMICS

Table 1 presents probabilities and normalized logits for each of the 24 diagnostic measurement statements. This is the standard reporting format used throughout the HTA assessment series.

It is essential to understand how to interpret these results.

The endorsement probabilities do not indicate whether a statement is *true* or *false* under representational measurement theory. Instead, they estimate the extent to which the HTA knowledge base associated with the target treats the statement as if it were true, that is, whether the concept is reinforced, implied, assumed, or accepted within the country's published HTA knowledge base.

The logits provide a continuous, symmetric scale, ranging from +2.50 to –2.50, that quantifies the degree of this endorsement. the logits, of course link to the probabilities (p) as the logit is the natural logarithm of the odds ratio; logit = ln[p/1-p].

- Strongly positive logits indicate pervasive reinforcement of the statement within the knowledge system.
- Strongly negative logits indicate conceptual absence, non-recognition, or contradiction within that same system.
- Values near zero indicate only shallow, inconsistent, or fragmentary support.

Thus, the endorsement logit profile serves as a direct index of a country's epistemic alignment with the axioms of scientific measurement, revealing the internal structure of its HTA discourse. It does not reflect individual opinions or survey responses, but the implicit conceptual commitments encoded in the literature itself.

TABLE 1: ITEM STATEMENT, RESPONSE, ENDORSEMENT AND NORMALIZED LOGITS OFFICE OF HEALTH ECONOMICS

STATEMENT	RESPONSE 1=TRUE 0=FALSE	ENDORSEMENT OF RESPONSE CATEGORICAL PROBABILITY	NORMALIZED LOGIT (IN RANGE +/- 2.50)
INTERVAL MEASURES LACK A TRUE ZERO	1	0.20	-1.40
MEASURES MUST BE UNIDIMENSIONAL	1	0.25	-1.10
MULTIPLICATION REQUIRES A RATIO MEASURE	1	0.10	-2.20

TIME TRADE-OFF PREFERENCES ARE UNIDIMENSIONAL	0	0.85	+1.75
RATIO MEASURES CAN HAVE NEGATIVE VALUES	0	0.90	+2.20
EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES	0	0.90	+2.20
THE QALY IS A RATIO MEASURE	0	0.95	+2.50
TIME IS A RATIO MEASURE	1	0.95	+2.50
MEASUREMENT PRECEDES ARITHMETIC	1	0.10	-2.20
SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES	0	0.85	+1.75
MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC	1	0.10	-2.20
THERE ARE ONLY TWO CLASSES OF MEASUREMENT LINEAR RATIO AND RASCH LOGIT RATIO	1	0.05	-2.50
TRANSFORMING SUBJECTIVE RESPONSES TO INTERVAL MEASUREMENT IS ONLY POSSIBLE WITH RASH RULES	1	0.05	-2.50
SUMMATION OF LIKERT QUESTION SCORES CREATES A RATIO MEASURE	0	0.90	+2.20
THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE	0	0.85	+1.75
CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.15	-1.75
QALYS CAN BE AGGREGATED	0	0.95	+2.50
NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.60	+0.40
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.90	+2.20
THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.65	+0.60
THE RASCH LOGIT RATIO SCALE IS THE ONLY BASIS FOR ASSESSING THERAPY IMPACT FOR LATENT TRAITS	1	0.05	-2.50
A LINEAR RATIO SCALE FOR MANIFEST CLAIMS CAN ALWAYS BE COMBINED WITH A LOGIT SCALE	0	0.60	+0.40

THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT	1	0.25	-1.10
THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.05	-2.50

THE OFFICE OF HEALTH ECONOMICS: INVERSION OF MEASUREMENT

The Office of Health Economics occupies a unique and historically decisive position in the evolution of health technology assessment. Founded in 1962 by the Association of the British Pharmaceutical Industry (ABPI), OHE was established not as a regulator, nor as a payer, but as an intellectual institution intended to shape the economic understanding of health policy in the United Kingdom. Over the following decades, it became far more than a research centre. It became a doctrinal engine.

If there exists a single institution that can plausibly be identified as the birthplace and long-term custodian of the modern HTA belief system, it is OHE. Long before NICE existed, before reference cases were codified, before ICER thresholds became administrative ritual, OHE was already constructing the conceptual architecture that made all of this appear natural, scientific, and inevitable. The 24-item diagnostic of Table 1 makes that role unmistakable.

What emerges in Table 1 is not a pattern of confusion or partial misunderstanding. It is a coherent and internally consistent inversion of scientific measurement, one that privileges arithmetic outputs while systematically excluding the axioms that make arithmetic meaningful. The strongest signals lie at the extremes of the logit scale.

At the positive ceiling sit the core dogmas of the HTA memeplex: the QALY as a ratio measure at +2.50; the aggregation of QALYs at +2.50; the legitimacy of negative utilities at +2.20; the claim that EQ-5D algorithms generate interval measures at +2.20; the belief that reference-case simulation models generate falsifiable claims at +2.20. These are not marginal ideas within the OHE knowledge base. They are foundational commitments.

At the negative floor sit the axioms that would prohibit these practices: measurement must precede arithmetic at −2.20; multiplication requires ratio measurement at −2.20; the existence of only two admissible measurement forms at −2.50; the necessity of Rasch transformation for latent traits at −2.50; the equivalence between Rasch rules and representational measurement axioms at −2.50.

This symmetry is devastating. It reveals a system that has not merely overlooked measurement theory but has actively constructed itself around its exclusion. OHE's role was not to accidentally misunderstand measurement. Its role was to build an alternative epistemology in which measurement constraints were treated as unnecessary obstacles to policy administration.

From the 1990s onward, OHE became one of the most vocal and sophisticated proponents of the QALY framework, promoting it as the natural outcome metric for health care decision making. This was not done naively. It was done strategically. The QALY solved a political problem: how to make rationing decisions appear objective without requiring empirical falsification of claims.

The Table 1 diagnostic captures this perfectly. The rejection of "measurement precedes arithmetic" at −2.20 is not an oversight; it is the enabling condition of the entire framework. If arithmetic had been required to wait for measurement validation, the QALY could never have existed. Utilities derived from ordinal preferences could never have been multiplied by time. Composite health state descriptions could never have been treated as quantities. Numerical storytelling would not have been the dominant characteristic.The system would have collapsed at birth.

Table 1 shows how deeply that contradiction has been normalized. The proposition that ratio Instead, OHE helped promote a worldview in which preference elicitation was treated as measurement by declaration. Once a panel of respondents expressed a ranking or trade-off, the resulting numbers were treated as if they possessed quantitative meaning. The fact that these numbers permitted negative values while being described as ratio measures was not treated as a fatal contradiction but as a technical nuance.

The proposition that ratio measures can have negative values is endorsed at +2.20. In any measurement-literate discipline, this would be nonsensical. In OHE's knowledge base, it is routine. This normalization allowed the creation of what might be called policy-grade arithmetic: numbers that look quantitative, behave numerically, and can be manipulated endlessly, while remaining unanchored to any empirical quantity with invariant units. This is why the QALY could become administratively indispensable. It did not require observation, replication, or falsification. It required only uncritical acceptance of the storytelling memeplex that is HTA..

Once accepted, everything else followed mechanically. Thresholds could be proposed. Cost-effectiveness planes could be drawn. League tables could be produced. NICE could be created with a reference case that delivered closure rather than knowledge.The diagnostic shows that OHE's knowledge base endorses closure over falsification. While there is moderate rhetorical support for rejecting non-falsifiable claims (+0.35), there is overwhelming endorsement of reference-case simulation as producing falsifiable claims (+2.20). This is epistemic laundering. A simulation cannot be falsified in the Popperian sense unless its outputs are tied to prospective protocols with real-world risk of refutation. Reference-case models are not. They are conditional imaginary stories stabilized through sensitivity analysis. Yet OHE promoted precisely this form of modeling as scientific evaluation.

The absence of Rasch measurement from the OHE worldview is particularly revealing. OHE has spent decades invoking latent constructs yet shows near-total rejection of the only measurement framework capable of producing invariant latent trait measures. All Rasch-related propositions collapse to the absolute floor at −2.50. This is not ignorance. Rasch theory has been available since the 1960s; the very decade in which OHE was founded. The exclusion is structural. Accepting Rasch would have required abandoning multiattribute utility models, summated ordinal instruments, and preference algorithms. It would have dismantled the QALY at its core. Thus

Rasch could never be allowed to become sovereign. It had to be marginalized, tolerated at the periphery, never permitted to function as a gatekeeper.

The result is that OHE's conception of "quality of life" is not a latent trait at all. It is a composite; a basket of attributes collapsed into a single score by convention. The diagnostic confirms this through weak endorsement of unidimensionality (−1.10) paired with strong endorsement of the unidimensionality of time trade-off health state preferences (+1.75). In other words, unidimensionality is asserted when required for arithmetic and ignored when it would constrain it. This is the defining feature of the memeplex. Rules apply selectively. Axioms are invoked rhetorically and abandoned operationally.

Over time, this belief system propagated outward. OHE trained economists, advised government, influenced NICE, informed academic curricula, and provided intellectual legitimacy for what became global HTA practice. The UK did not merely adopt numerical storytelling; it exported it. What makes OHE's role especially consequential is its position between industry and government. Founded by ABPI, yet operating as an independent authority, OHE functioned as a bridge that made the QALY appear scientifically neutral while serving administrative objectives. It offered a low-data, high-authority framework capable of producing decisive recommendations without the burden of empirical testing.

That burden matters. True measurement produces provisional claims. Provisional claims invite challenge. Challenge threatens closure. The reference-case framework solved this problem elegantly: it replaced falsification with consensus modeling. The diagnostic shows exactly how this trade-off was made. Measurement axioms were sacrificed so that arithmetic could proceed unimpeded. The cost was scientific legitimacy; the benefit was policy finality.

Forty years later, the consequences are visible everywhere. Health systems speak in QALYs as if they were natural units. Journals reinforce the same constructs. Agencies defend thresholds. Entire generations of economists have been trained without exposure to representational measurement theory. And yet the numbers still do not measure anything. It is difficult to believe, even among those with no understanding of the axioms of representational measurement, how so many could be so naïve.

The Office of Health Economics therefore occupies a unique historical position. It did not merely participate in the HTA memeplex. It helped design it. The 24-item profile does not accuse OHE of bad faith. It demonstrates something more important: that the institution began from the wrong starting point. Measurement was never installed as the gatekeeper. Once that decision was made, explicitly or implicitly, everything that followed was inevitable. The QALY could be born. NICE could be constructed. Reference-case modeling could dominate. Numerical storytelling could become global orthodoxy.

From a measurement perspective, the verdict is unambiguous. OHE's legacy is not the advancement of quantitative health evaluation, but the institutionalization of arithmetic without measurement. It provided coherence, stability, and administrative convenience, but at the price of abandoning the standards that define science. The tragedy is not that the framework was flawed. It is that the flaw was foundational and once embedded, almost impossible to dislodge. This is why

reassessing OHE matters now. If health technology assessment is ever to transition toward evaluable, falsifiable, single-claim evidence grounded in representational measurement, the intellectual lineage that begins in the UK and runs through OHE must be confronted directly. Not as history but as unfinished business.

IGNORING EPISTEMIC LEGITIMACY: JUSTIFICATION FOR THE REFERENCE CASE

From the perspective of representational measurement, the most revealing question is not whether the reference case model is mathematically sophisticated, nor whether it is widely used, nor even whether it produces administratively convenient answers. The critical question is why such a framework was promoted and defended despite its indifference to epistemic legitimacy. The Office of Health Economics (OHE), particularly from the early 1990s onward, occupies a central position in answering that question. OHE did not merely observe the rise of the reference case; it provided intellectual cover for a framework whose defining feature was not scientific validity, but closure.

The reference case did not emerge as an attempt to discover truths about therapy impact. It emerged as a response to an institutional problem: how to make allocative decisions under conditions of limited data, political pressure, and fiscal constraint. The attraction of the reference case was not epistemic coherence but administrative utility. It offered a way to produce numbers that looked authoritative, were internally consistent, and—most importantly—allowed decisions to be declared final. In this sense, the framework was never designed to satisfy the requirements of normal science. It was designed to be believed.

OHE's role in this transition cannot be understood through the lens of scientific development. It must be understood through what might be called a post-epistemic logic. Under this logic, the question is no longer whether a claim is true in the sense of being measurable, falsifiable, or reproducible. The question becomes whether the claim can function socially as a justification for action. The reference case model answered that requirement perfectly. It generated outputs that could be cited, compared, and defended rhetorically, even though the numerical objects involved—utilities, QALYs, thresholds—failed the axioms of representational measurement.

In this environment, epistemic legitimacy was not denied explicitly; it was rendered irrelevant. Measurement was never rejected outright. It was simply bypassed. Arithmetic was allowed to proceed in advance of any demonstration that the quantities involved were measurable. The axioms governing scale type, unidimensionality, invariance, and permissible transformations were absent not because they were unknown, but because acknowledging them would have made closure impossible. If measurement had been treated as a gatekeeping condition, then most submissions would have remained provisional, subject to refutation, revision, and ongoing empirical challenge over a product's lifespan. That was precisely what policymakers wished to avoid.

OHE's fixation on QALYs from the 1990s onward must be understood in this light. The QALY was not adopted because it satisfied measurement theory. It was adopted because it offered a single scalar object that could be inserted into models, compared across disease areas, and aligned with cost thresholds. Its power was administrative, not scientific. It converted heterogeneous clinical

realities into a uniform currency that could be governed. That currency did not need to be valid; it needed to be stable.

The reference case therefore functioned as a social technology. It disciplined manufacturers by forcing conformity to a predefined analytical structure. It reassured policymakers by presenting allocation decisions as technical rather than political. And it neutralized epistemic challenge by embedding all uncertainty inside model assumptions rather than exposing claims to empirical falsification. OHE's contribution was to normalize this arrangement as "good economics," thereby insulating it from philosophical or measurement-based critique.

Seen this way, the indifference to representational measurement was not accidental. It was functional. Measurement theory is disruptive. It demands that numerical claims correspond to empirical structure. It insists on unidimensionality where composites are convenient, on true zero where ratios are desired, and on invariance where comparability is assumed. If applied seriously, measurement theory would have fractured the reference case framework at its core. The very constructs that made the system administratively workable would have been declared inadmissible.

In post-epistemic terms, this was unacceptable. Policymakers were not seeking provisional truths; they were seeking decisional finality. The reference case offered precisely that. Once a model was constructed and a threshold applied, the matter could be closed. The claim did not need to be true in a scientific sense. It only needed to be defensible within the shared conventions of the framework. OHE's writings during this period consistently reinforced this orientation, emphasizing consistency, comparability, and decision support, while remaining silent on the axioms that would determine whether the numbers being compared were measures at all.

This silence is telling. Stevens' typology of measurement scales was already decades old. Representational measurement theory was well established. Rasch measurement had been available since the 1960s. None of this entered OHE's framing of value. The absence was not ignorance; it was exclusion. Measurement theory was epistemically inconvenient.

The result was the institutionalization of a belief system in which numerical storytelling replaced empirical discovery. Models became substitutes for observation. Sensitivity analysis replaced falsification. Plausibility replaced truth. The reference case did not test hypotheses; it managed narratives. And because the narratives were expressed in numbers, they carried the aura of science without submitting to its discipline.

What the LLM diagnostic now exposes is that this belief system has a detectable structure. It consistently endorses propositions that permit arithmetic without measurement and rejects those that would prevent it. This pattern is not random. It is the fingerprint of a framework designed to function allocatively rather than epistemically.

OHE's legacy must therefore be assessed not in terms of analytical contribution, but in terms of epistemic consequence. By championing a framework that prioritized administrative closure over measurement validity, it helped create a global HTA culture in which the question "Is this a measure?" was never allowed to precede the question "What does the model say?" That inversion defined the system.

The tragedy is that this choice was not inevitable. A different path was available: one grounded in single, falsifiable claims; in ratio measures for manifest attributes; in Rasch logit measures for latent traits; and in ongoing empirical reassessment rather than one-off closure. That path would have been more demanding, less tidy, and resistant to political finality. It was therefore rejected.

What OHE implicitly assumed—and what the reference case depended upon—was that belief would suffice. That numbers presented with confidence, institutional endorsement, and methodological ritual would be accepted as authoritative for allocative purposes even in the absence of epistemic legitimacy. For decades, that assumption held.

It no longer does.

The emergence of systematic diagnostic interrogation changes the environment entirely. Once the axioms are made explicit and institutional endorsement patterns are revealed, the reference case can no longer hide behind technical complexity or consensus. Its justification is exposed as social rather than scientific.

In that sense, the reference case was never a failure of technique. It was a success of governance. But governance achieved by suspending the conditions of knowledge cannot endure once those conditions are brought back into view. OHE's enduring challenge is not how to refine the framework it helped build, but whether it can acknowledge that a system designed to be believed cannot substitute indefinitely for one designed to be true.

RESTORING EPISTEMIC LEGITIMACY: DECONSTRUCTING THE REFERENCE CASE WITH AI LLM TOOLS

For more than three decades, health technology assessment has operated within an analytical framework that has rarely been challenged at its foundations. The reference case model—now embedded across national agencies, academic centers, consultancies, and journals—has come to define what counts as acceptable evidence for pricing and access decisions. Its authority has rested not on demonstrable measurement validity, but on institutional repetition. The result has been a system that produces numerical outputs without first establishing whether those numbers possess the properties required for scientific inference. What has been missing is not criticism, but a mechanism capable of exposing this failure systematically.

The emergence of large language model (LLM) based diagnostic tools changes that condition. These tools do not generate new evidence, nor do they adjudicate truth claims. Instead, they enable interrogation of belief systems at scale. They reveal what institutions consistently endorse, what they systematically reject, and most importantly what they never allow to become governing constraints. In doing so, they restore a capability that health technology assessment long abandoned: the ability to examine whether its core analytical commitments satisfy the prerequisites of scientific knowledge.

The reference case framework has always claimed legitimacy through internal coherence. Models are carefully specified, assumptions documented, sensitivity analyses performed, and results presented with statistical refinement. Yet coherence is not measurement. Representational

measurement theory makes this distinction unambiguous. Arithmetic is not licensed by methodological ritual. It is licensed only when the empirical attribute being represented possesses a structure compatible with the numerical operations applied to it. Without unidimensionality, invariant units, and—where ratios are invoked—a true zero, arithmetic becomes symbolic rather than substantive.

What the LLM diagnostic reveals is that the reference case framework inverts this logic. Across agencies and journals, propositions affirming that measurement must precede arithmetic are weakly endorsed or rejected outright. At the same time, propositions presupposing arithmetic legitimacy—such as the aggregation of QALYs, the ratio status of utilities, and the permissibility of negative health values—are endorsed at or near ceiling levels. This inversion is not accidental. It is structural. The system is organized to permit calculation first and ask questions of meaning later, if at all.

LLM interrogation is uniquely suited to exposing this structure because it operates across corpora rather than within individual papers. Traditional critique focuses on isolated articles, methodological missteps, or modeling choices. These critiques are easily deflected as contextual or idiosyncratic. The LLM diagnostic, by contrast, examines patterns of reinforcement across thousands of texts. It identifies what the knowledge base treats as admissible and what it excludes as unthinkable. In this sense, it functions not as a reviewer but as an epistemic mirror.

When applied to the reference case, that mirror reveals a belief system rather than a scientific framework. Core axioms of measurement—scale type, permissible transformations, invariance—are absent as gatekeeping conditions. They are not debated, tested, or refined; they are bypassed. In their place stands a set of conventions justified by practicality, precedent, and administrative need. The system does not ask whether a QALY is a measure. It asks only whether it is usable.

This distinction matters because science is not defined by usefulness. It is defined by vulnerability to refutation. A claim that cannot be falsified is not provisional knowledge; it is narrative. The reference case avoids falsification by design. Its outputs are projections conditional on assumptions rather than claims exposed to empirical risk. Sensitivity analysis explores alternative beliefs, not alternative realities. As a result, no outcome of the model can ever be wrong in the scientific sense. It can only be "more or less plausible."

LLM diagnostics expose this post-epistemic architecture with unsettling clarity. Institutions that publicly endorse rejection of non-falsifiable claims simultaneously endorse simulation frameworks that cannot be falsified. This contradiction is not resolved conceptually; it is absorbed institutionally. The language of science is retained while its discipline is quietly abandoned.

What makes the LLM intervention so disruptive is not that it introduces new theory, but that it reinstates old ones. Stevens' scale typology, representational measurement axioms and Rasch requirements for latent traits. These are not radical innovations. They have existed for decades. What LLM diagnostics demonstrate is that these principles were never integrated into HTA governance. They were not disproven. They were ignored.

This is why the present moment is different. For the first time, it is possible to demonstrate, not rhetorically, but structurally, that the reference case operates in systematic violation of measurement requirements. The issue is no longer one of opinion or philosophical preference. It is one of demonstrable belief endorsement. When an institution consistently rejects the conditions under which its arithmetic would be valid, its numerical outputs cannot claim epistemic authority.

Restoring epistemic legitimacy therefore does not require refining the reference case. It requires abandoning it as a decision-anchoring framework. No amount of methodological polish can rescue arithmetic applied to non-measures. The only path forward is to reorder the evaluative sequence. Measurement must precede arithmetic. Claims must be unidimensional. Manifest attributes must be expressed on linear ratio scales. Latent attributes must be measured through Rasch logit ratio scales with demonstrated invariance. Anything else may be descriptive, but it cannot be evidentiary.

LLM tools do not replace human judgment. They restore its possibility. By making institutional belief systems visible, they reopen questions that were prematurely closed in the 1990s. They expose how administrative convenience displaced scientific obligation. And they make clear that the authority of the reference case rested not on truth, but on repetition.

The implication is profound. If the reference case was never epistemically legitimate, then its global dominance does not reflect consensus—it reflects transmission. What spread was not knowledge, but a memeplex: a self-reinforcing system of practices protected from falsification by design.

LLM diagnostics do not destroy this system by force. They simply remove its invisibility. Once exposed, the claim that “there is no alternative” collapses. There has always been an alternative: measurement before arithmetic, falsifiable claims before closure, science before governance.

The task now is not to defend the past, but to reclaim the future. Restoring epistemic legitimacy means returning HTA to the standards that define scientific inquiry: measurable attributes, lawful arithmetic, and claims that can be wrong. Without these, no framework—however elegant—can claim authority. With them, health technology assessment can finally become what it has long claimed to be: an evidence-based discipline rather than a numerical storytelling system dressed in scientific form.

3. THE TRANSITION TO MEASUREMENT IN HEALTH TECHNOLOGY ASSESSMENT

THE IMPERATIVE OF CHANGE

This analysis has not been undertaken to criticize decisions made by health system, nor to assign responsibility for the analytical frameworks currently used in formulary review. The evidence shows something more fundamental: organizations have been operating within a system that does not permit meaningful evaluation of therapy impact, even when decisions are made carefully, transparently, and in good faith.

The present HTA framework forces health systems to rely on numerical outputs that appear rigorous but cannot be empirically assessed (Table 1). Reference-case models, cost-per-QALY ratios, and composite value claims are presented as decision-support tools, yet they do not satisfy the conditions required for measurement. As a result, committees are asked to deliberate over results that cannot be validated, reproduced, or falsified. This places decision makers in an untenable position: required to choose among therapies without a stable evidentiary foundation.

This is not a failure of expertise, diligence, or clinical judgment. It is a structural failure. The prevailing HTA architecture requires arithmetic before measurement, rather than measurement before arithmetic. Health systems inherit this structure rather than design it. Manufacturers respond to it. Consultants reproduce it. Journals reinforce it. Universities promote it. Over time it has come to appear normal, even inevitable.

Yet the analysis presented in Table 1 demonstrates that this HTA framework cannot support credible falsifiable claims. Where the dependent variable is not a measure, no amount of modeling sophistication can compensate. Uncertainty analysis cannot rescue non-measurement. Transparency cannot repair category error. Consensus cannot convert assumption into evidence.

The consequence is that formulary decisions are based on numerical storytelling rather than testable claims. This undermines confidence, constrains learning, and exposes health systems to growing scrutiny from clinicians, patients, and regulators who expect evidence to mean something more than structured speculation.

The imperative of change therefore does not arise from theory alone. It arises from governance responsibility. A health system cannot sustain long-term stewardship of care if it lacks the ability to distinguish between claims that can be evaluated and claims that cannot. Without that distinction, there is no pathway to improvement; only endless repetition for years to come.

This transition is not about rejecting evidence. It is about restoring evidence to its proper meaning. It requires moving away from composite, model-driven imaginary constructs toward claims that are measurable, unidimensional, and capable of empirical assessment over time. The remainder of this section sets out how that transition can occur in a practical, defensible, and staged manner.

MEANINGFUL THERAPY IMPACT CLAIMS

At the center of the current problem is not data availability, modeling skill, or analytic effort. It is the nature of the claims being advanced. Contemporary HTA has evolved toward increasingly complex frameworks that attempt to compress multiple attributes, clinical effects, patient experience, time, and preferences into single composite outputs. These constructs are then treated as if they were measures. They are not (Table 1).

The complexity of the reference-case framework obscures a simpler truth: meaningful evaluation requires meaningful claims. A claim must state clearly what attribute is being affected, in whom, over what period, and how that attribute is measured. When these conditions are met, evaluation becomes possible. When they are not complexity substitutes for clarity. The current framework is not merely incorrect; it is needlessly elaborate. Reference-case modeling requires dozens of inputs, assumptions, and transformations, yet produces outputs that cannot be empirically verified. Each additional layer of complexity increases opacity while decreasing accountability. Committees are left comparing models rather than assessing outcomes.

In contrast, therapy impact can be expressed through two, and only two, types of legitimate claims. First are claims based on manifest attributes: observable events, durations, or resource units. These include hospitalizations avoided, time to event, days in remission, or resource use. When properly defined and unidimensional, these attributes can be measured on linear ratio scales and evaluated directly.

Second are claims based on latent attributes: symptoms, functioning, need fulfillment, or patient experience. These cannot be observed directly and therefore cannot be scored or summed meaningfully. They require formal measurement through Rasch models to produce invariant logit ratio scales. These two forms of claims are sufficient. They are also far more transparent. Each can be supported by a protocol. Each can be revisited. Each can be reproduced. Most importantly, each can fail. But they cannot be combined. This is the critical distinction. A meaningful claim is one that can be wrong.

Composite constructs such as QALYs do not fail in this sense. They persist regardless of outcome because they are insulated by assumptions. They are recalculated, not refuted. That is why they cannot support learning. The evolution of objective knowledge regarding therapy impact in disease areas is an entirely foreign concept. By re-centering formulary review on single-attribute, measurable claims, health systems regain control of evaluation. Decisions become grounded in observable change rather than modeled narratives. Evidence becomes something that accumulates, rather than something that is re-generated anew for every submission.

THE PATH TO MEANINGFUL MEASUREMENT

Transitioning to meaningful measurement does not require abandoning current processes overnight. It requires reordering them. The essential change is not procedural but conceptual: measurement must become the gatekeeper for arithmetic, not its byproduct.

The first step is formal recognition that not all numerical outputs constitute evidence. Health systems must explicitly distinguish between descriptive analyses and evaluable claims. Numbers that do not meet measurement requirements may inform discussion but cannot anchor decisions.

The second step is restructuring submissions around explicit claims rather than models. Each submission should identify a limited number of therapy impact claims, each defined by attribute, population, timeframe, and comparator. Claims must be unidimensional by design.

Third, each claim must be classified as manifest or latent. This classification determines the admissible measurement standard and prevents inappropriate mixing of scale types.

Fourth, measurement validity must be assessed before any arithmetic is permitted. For manifest claims, this requires confirmation of ratio properties. For latent claims, this requires Rasch-based measurement with demonstrated invariance.

Fifth, claims must be supported by prospective or reproducible protocols. Evidence must be capable of reassessment, not locked within long-horizon simulations designed to frustrate falsification.

Sixth, committees must be supported through targeted training in representational measurement principles, including Rasch fundamentals. Without this capacity, enforcement cannot occur consistently.

Finally, evaluation must be iterative. Claims are not accepted permanently. They are monitored, reproduced, refined, or rejected as evidence accumulates.

These steps do not reduce analytical rigor. They restore it.

TRANSITION REQUIRES TRAINING

A transition to meaningful measurement cannot be achieved through policy alone. It requires a parallel investment in training, because representational measurement theory is not intuitive and has never been part of standard professional education in health technology assessment, pharmacoeconomics, or formulary decision making. For more than forty years, practitioners have been taught to work within frameworks that assume measurement rather than demonstrate it. Reversing that inheritance requires structured learning, not informal exposure.

At the center of this transition is the need to understand why measurement must precede arithmetic. Representational measurement theory establishes the criteria under which numbers can legitimately represent empirical attributes. These criteria are not optional. They determine whether addition, multiplication, aggregation, and comparison are meaningful or merely symbolic. Without this foundation, committees are left evaluating numerical outputs without any principled way to distinguish evidence from numerical storytelling.

Training must therefore begin with scale types and their permissible operations. Linear ratio measurement applies to manifest attributes that possess a true zero and invariant units, such as

time, counts, and resource use. Latent attributes, by contrast, cannot be observed directly and cannot be measured through summation or weighting. They require formal construction through a measurement model capable of producing invariant units. This distinction is the conceptual fulcrum of reform, because it determines which claims are admissible and which are not.

For latent trait claims, Rasch measurement provides the only established framework capable of meeting these requirements. Developed in the mid–twentieth century alongside the foundations of modern measurement theory, the Rasch model was explicitly designed to convert subjective observations into linear logit ratio measures. It enforces unidimensionality, tests item invariance, and produces measures that support meaningful comparison across persons, instruments, and time. These properties are not approximations; they are defining conditions of measurement.

Importantly, Rasch assessment is no longer technically burdensome. Dedicated software platforms developed and refined over more than four decades make Rasch analysis accessible, transparent, and auditable. These programs do not merely generate statistics; they explain why items function or fail, how scales behave, and whether a latent attribute has been successfully measured. Measurement becomes demonstrable rather than assumed.

Maimon Research has developed a two-part training program specifically to support this transition. The first component provides foundational instruction in representational measurement theory, including the historical origins of scale theory, the distinction between manifest and latent attributes, and the criteria that define admissible claims. The second component focuses on application, detailing claim types, protocol design, and the practical use of Rasch methods to support latent trait evaluation.

Together, these programs equip health systems, committees, and analysts with the competence required to enforce measurement standards consistently. Training does not replace judgment; it enables it. Without such preparation, the transition to meaningful measurement cannot be sustained. With it, formulary decision making can finally rest on claims that are not merely numerical, but measurable.

A NEW START IN MEASUREMENT FOR HEALTH TECHNOLOGY ASSESSMENT

For readers who are looking for an introduction to measurement that meets the required standards, Maimon Research has just released two distance education programs. These are:

- Program 1: Numerical Storytelling – Systematic Measurement Failure in HTA.
- Program 2: A New Start in Measurement for HTA, with recommendations for protocol-supported claims for specific objective measures as well as latent constructs and manifested traits.

Each program consists of five modules (approx. 5,500 words each), with extensive questions and answers. Each program is priced at US$65.00. Invitations to participate in these programs will be distributed in the first instance to 8,700 HTA professionals in 40 countries.

More detail on program content and access, including registration and on-line payment, is provided with this link: https://maimonresearch.com/distance-education-programs/

DESIGNED FOR CLOSURE

For those who remain unconvinced that there is any need to abandon a long-standing and widely accepted HTA framework, it is necessary to confront a more fundamental question: why was this system developed and promoted globally in the first place?

The most plausible explanation is administrative rather than scientific. Policy makers were searching for an assessment framework that could be applied under conditions of limited empirical data while still producing a determinate conclusion. Reference-case modeling offered precisely this convenience. By constructing a simulation populated with assumptions, surrogate endpoints, preference weights, and extrapolated time horizons, it became possible to generate a numerical result that could be interpreted as decisive. Once an acceptable cost-effectiveness ratio emerged, the assessment could be declared complete and the pricing decision closed. This structure solved a political and administrative problem. It allowed authorities to claim that decisions were evidence-based without requiring the sustained empirical burden demanded by normal science. There was no requirement to formulate provisional claims and subject them to ongoing falsification. There was no obligation to revisit conclusions as new data emerged. Closure could be achieved at launch, rather than knowledge evolving over the product life cycle.

By contrast, a framework grounded in representational measurement would have imposed a very different obligation. Claims would necessarily be provisional. Measurement would precede arithmetic. Each therapy impact claim would require a defined attribute, a valid scale, a protocol, and the possibility of replication or refutation. Evidence would accumulate rather than conclude. Decisions would remain open to challenge as real-world data emerged. From an administrative standpoint, this was an unreasonable burden. It offered no finality.

The reference-case model avoided this problem entirely. By shifting attention away from whether quantities were measurable and toward whether assumptions were plausible, the framework replaced falsification with acceptability. Debate became internal to the model rather than external to reality. Sensitivity analysis substituted for empirical risk. Arithmetic proceeded without prior demonstration that the objects being manipulated possessed the properties required for arithmetic to be meaningful.

Crucially, this system required no understanding of representational measurement theory. Committees did not need to ask whether utilities were interval or ratio measures, whether latent traits had been measured or merely scored, or whether composite constructs could legitimately be multiplied or aggregated. These questions were never posed because the framework did not require

them to be posed. The absence of measurement standards was not an oversight; it was functionally essential.

Once institutionalized, the framework became self-reinforcing. Training programs taught modeling rather than measurement. Guidelines codified practice rather than axioms. Journals reviewed technique rather than admissibility. Over time, arithmetic without measurement became normalized as "good practice," while challenges grounded in measurement theory were dismissed as theoretical distractions. The result was a global HTA architecture capable of producing numbers, but incapable of producing falsifiable knowledge. Claims could be compared, ranked, and monetized, but not tested in the scientific sense. What evolved was not objective knowledge, but institutional consensus.

This history matters because it explains why the present transition is resisted. Moving to a real measurement framework with single, unidimensional claims does not merely refine existing methods; it dismantles the very mechanism by which closure has been achieved for forty years. It replaces decisiveness with accountability, finality with learning, and numerical plausibility with empirical discipline. Yet that is precisely the transition now required. A system that avoids measurement in order to secure closure cannot support scientific evaluation, cumulative knowledge, or long-term stewardship of healthcare resources. The choice is therefore unavoidable: continue with a framework designed to end debate, or adopt one designed to discover the truth.

Anything else is not assessment at all, but the ritualized manipulation of numbers detached from measurement, falsification, and scientific accountability.

ACKNOWLEDGEMENT

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

REFERENCES

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

[2] Krantz D, Luce R, Suppes P, Tversky A. Foundations pf Measurement Vol 1: Additive and Polynomial Representations. New York: Academic Press, 1971

[3] Rasch G, Probabilistic Models for some Intelligence and Attainment Tests. Chicago: University of Chicago Press, 1980 [An edited version of the original 1960 publication]

[4] Wright B. Solving measurement problems with the Rasch Model. *J Educational Measurement.* 1977;14(2):97-116