MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

UNITED KINGDOM: THE BIRTHPLACE OF FALSE MEASUREMENT IN HEALTH TRCHNOLOGY ASSESSMENT

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

LOGIT WORKING PAPER No 24 JANUARY 2026

www.maimonresearch.com

Tucson AZ

FOREWORD

HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

The United Kingdom occupies a singular position in the modern history of health technology assessment. It was in the UK that cost-utility analysis, preference-based multiattribute instruments, and reference-case economic evaluation were first assembled into a coherent policy framework and endowed with institutional authority. Through the establishment of national appraisal processes, most visibly under the auspices of National Institute for Health and Care Excellence, the UK did not merely adopt these constructs; it normalized them. What began as methodological expedients to support centralized decision making were elevated into putative scientific standards, exported internationally, and reproduced through guidelines, journals, and training programs. In this sense, the UK is not simply one jurisdiction among many, but the birthplace of a global evaluative architecture.

The purpose of this first Logit Working Paper for the UK is to set the analytical framework to review this legacy through the lens of representational measurement. The central claim is not that UK HTA has misapplied otherwise sound tools, but that it institutionalized a framework grounded in false measurement from the outset. Multiattribute utility instruments, QALYs, and reference-case simulations were never shown to meet the axioms required for quantitative claims, yet they were rapidly embedded as if they did. This paper applies a canonical logit assessment to the UK HTA knowledge base to demonstrate that foundational measurement principles of unidimensionality, admissible arithmetic, ratio scale requirements, and falsifiability do not operate as binding constraints. The UK case therefore matters not because it is now not uniquely flawed, but because it established and legitimized the evaluative false measurement memeplex that now defines HTA worldwide.

The objective of this first study is to evaluate the United Kingdom's health technology assessment knowledge base using the 24-item canonical measurement diagnostic, with the aim of determining whether the axioms of representational measurement operate as binding constraints on quantitative claims. The assessment is not concerned with the technical sophistication of UK HTA methods, nor with their institutional influence, but with their epistemic status. Specifically, the study examines whether unidimensionality, admissible arithmetic, ratio scale requirements, dimensional homogeneity, Rasch measurement for latent traits, and falsifiability are recognized as necessary conditions for evaluable claims, or whether they are systematically bypassed in favor of administratively convenient constructs such as QALYs, cost-effectiveness ratios, and reference-case simulations.

The canonical assessment reveals a uniform pattern of non-possession of foundational measurement principles across the UK HTA knowledge base. Statements asserting axiomatic requirements—such as the priority of measurement over arithmetic, the necessity of ratio scales for multiplication, the unidimensionality of measures, and the role of Rasch rules in transforming

ordinal responses—collapse to floor or near-floor logit values. In contrast, false propositions central to the UK evaluative framework, including the treatment of EQ-5D indices as interval or ratio measures, the dimensional coherence and aggregability of QALYs, and the falsifiability of reference-case simulation outputs, are strongly endorsed. The findings indicate that UK HTA does not merely misapply measurement theory; it operates within a closed evaluative framework in which measurement axioms do not function as admissible constraints on what counts as evidence.

The foundation of representational measurement is the principle that measurement precedes arithmetic. This 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 UNITED KINGDOM NOWLEDGE BASE

The UK health technology assessment knowledge base represents the most fully articulated and institutionally entrenched expression of modern HTA practice. It integrates national guidelines, academic HTA centers, and a substantial journal literature into a coherent evaluative architecture that has been widely emulated internationally. This architecture is frequently presented as methodologically rigorous, transparent, and evidence based. However, when examined through the lens of representational measurement, its defining characteristic is not rigor but closure.

At the core of the UK framework lies a commitment to multiattribute utility instruments, QALYs, and reference-case simulation models as the primary means of expressing therapeutic value. These constructs were adopted to address a policy problem—allocating finite resources under uncertainty—by producing a single quantitative index capable of supporting comparative decisions. The framework’s success in delivering decisional closure has often been mistaken for scientific validity. The canonical assessment shows that this closure was achieved by systematically bypassing the axioms required for measurement.

Within the UK knowledge base, unidimensionality does not operate as a prerequisite for quantification. Health is decomposed into multiple attributes, preferences over composite states are elicited, and the resulting scores are treated as if they represented a single measurable quantity. This conflation of description, preference, and measurement is not interrogated. Instead, it is normalized through guidelines and reinforced through publication practice. As a result, the question “what exactly is being measured?” is rarely posed, and never answered.

Arithmetic within the UK framework proceeds independently of scale type. Ordinal responses are summed, weighted, multiplied by time, and aggregated across individuals without demonstration that the underlying constructs possess ratio properties. The canonical logit profile shows that the proposition “measurement precedes arithmetic” is effectively absent from the UK evaluative vocabulary. Arithmetic is treated as an analytic convenience rather than as an operation constrained by the properties of the scale. This inversion is incompatible with normal science, yet it is foundational to UK HTA practice.

The treatment of QALYs is particularly revealing. The knowledge base strongly endorses the claims that QALYs are ratio measures, dimensionally homogeneous, and aggregable. These endorsements persist despite the absence of a demonstrated true zero, the presence of negative values, and the conflation of time with preference-weighted scores. Rather than addressing these contradictions, the UK framework absorbs them by redefining what counts as acceptable quantification. QALYs function not as measures, but as policy tokens that enable comparison while remaining insulated from refutation.

Latent traits are handled with similar disregard for measurement requirements. Patient-reported outcomes and quality-of-life constructs are routinely invoked, yet the only established method for

constructing linear measures from ordinal responses—the Rasch model—is effectively absent. All Rasch-related propositions collapse to floor logit values, indicating that Rasch measurement does not operate as a recognized or binding principle within the UK knowledge base. Subjective data are treated as directly quantifiable, and their numerical outputs are granted evidentiary status without lawful transformation.

Falsifiability, a cornerstone of scientific inquiry, is rhetorically acknowledged but operationally neutralized. While there is limited endorsement of rejecting non-falsifiable claims, the framework simultaneously endorses the proposition that reference-case simulations generate falsifiable claims. In practice, simulation outputs are not tested against observed outcomes in a way that would permit decisive failure. Instead, uncertainty is managed through sensitivity analysis, scenario exploration, and recalibration. Error is accommodated rather than eliminated. Learning is defined as refinement within an accepted model rather than as rejection of a failed claim.

The cumulative effect is a knowledge base that is internally coherent but epistemically closed. Quantitative claims are produced in abundance, refined in detail, and embedded in decision processes, yet they are shielded from the possibility of being wrong. Measurement axioms do not constrain belief; they are simply not part of the evaluative grammar. This explains the resilience of the UK framework in the face of repeated conceptual critique. Challenges grounded in representational measurement do not register as relevant objections because the framework no longer recognizes their authority.

The UK case is significant not because it is uniquely flawed, but because it established and legitimized this mode of evaluation. By codifying false measurement into national guidance and exporting it through training, consultancy, and publication, the UK created a global HTA memeplex. Other jurisdictions inherited this framework largely intact. The canonical assessment makes clear that the problem is not local or accidental; it is systemic and foundational.

Until the UK HTA knowledge base recognizes measurement axioms as non-negotiable constraints, it cannot function as a scientific enterprise. It will continue to produce numbers, comparisons, and decisions, but not evaluable claims. The issue, therefore, is not whether UK HTA can be improved at the margins, but whether it can relinquish an evaluative framework built on false measurement and re-enter the tradition of normal science in which claims are exposed to the risk of being wrong.

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: UNITED KINGDOM

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.

UNITED KINGDOM: A CANONICAL LOGIT ASSESSMENT OF THE HTA KNOWLEDGE BASE

The United Kingdom occupies a foundational position in the global development of health technology assessment. It is here that cost-utility analysis, preference-based multiattribute instruments, and reference-case modeling were first assembled into a coherent national evaluative framework and endowed with formal decision authority. Through the establishment of centralized appraisal processes, the UK did not simply adopt these tools as pragmatic aids; it institutionalized them as scientific standards. The canonical assessment presented here demonstrates that this institutionalization occurred in the absence of adherence to the axioms of representational measurement, and that this absence has persisted, unchallenged and largely unrecognized, across four decades of HTA practice.

TABLE 1: ITEM STATEMENT, RESPONSE, ENDORSEMENT AND NORMALIZED LOGITS UNITED KINGDOM

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	0.10	-2.20
MULTIPLICATION REQUIRES A RATIO MEASURE	1	0.05	-2.50
TIME TRADE-OFF PREFERENCES ARE UNIDIMENSIONAL	0	0.90	+2.20
RATIO MEASURES CAN HAVE NEGATIVE VALUES	0	0.95	+2.50
EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES	0	0.95	+2.50
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	0.05	-2.50
SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES	0	0.90	+2.20
MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC	1	0.05	-2.50
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	0.95	+2.50
CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.05	-2.50
QALYS CAN BE AGGREGATED	0	0.95	+2.50

NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.25	-0.95
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.90	+2.50
THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.40	-0.45
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	0.80	+1.40
THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT	1	0.05	-2.50
THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.05	-2.50

The logit profile for the United Kingdom is striking not for its variability but for its consistency. Statements that define what measurement requires collapse uniformly to floor or near-floor values. Unidimensionality, the requirement that measurement precede arithmetic, the necessity of ratio scales for multiplication, and the role of Rasch rules in transforming ordinal responses all register at −2.50 or adjacent values. These results indicate effective non-possession. The propositions do not function as live constraints within the UK HTA knowledge base; they are not debated, refined, or selectively applied. They are absent.

This absence is not accidental. The UK framework was constructed to solve a policy problem—how to allocate finite NHS resources under conditions of uncertainty—rather than a measurement problem. Multiattribute utility instruments such as EQ-5D provided a convenient means of collapsing heterogeneous health descriptions into a single index. That index, when multiplied by time, produced QALYs, which could then be compared against costs through a single ratio. This architecture delivered decisional closure. It did not deliver measurement.

The canonical table shows that false propositions central to this architecture are strongly endorsed. The claims that EQ-5D preference algorithms create interval measures, that the QALY is a ratio measure, that QALYs are dimensionally homogeneous, and that QALYs can be aggregated all register at the upper end of the logit scale. These are not marginal assumptions; they are the load-bearing beams of the UK evaluative framework. Their endorsement reflects a settled belief system rather than an empirically defended position.

The asymmetry between rejected true statements and endorsed false statements is epistemically decisive. It shows that UK HTA does not merely overlook measurement axioms; it operates as if

they are irrelevant. Arithmetic is performed first, and questions of measurement validity are deferred indefinitely or reframed as methodological preferences. This inversion of arithmetic preceding measurement is incompatible with any conception of normal science.

The treatment of preference-based utilities illustrates this inversion clearly. Time trade-off preferences are treated as if they were unidimensional measures of health, despite being elicited over complex, multiattribute health states. The logit value for the statement "time trade-off preferences are unidimensional" sits at +2.20, indicating strong endorsement of a proposition that has no foundation in measurement theory. Preferences are not measures of health; they are judgments about states of health. Treating them as interchangeable is a categorical error, yet it is one that the UK framework has normalized.

Negative utilities ("worse than dead") are similarly revealing. The strong endorsement of the false statement that ratio measures can have negative values reflects a confusion between scale permissibility and scale existence. The issue is not whether negative values can appear on some ratio scales in principle, but whether the construct being measured has been shown to possess ratio properties at all. In the UK framework, the answer is assumed rather than demonstrated.

The Rasch-related items provide the clearest diagnostic. All Rasch statements collapse to −2.50. This is not a minor technical omission. Rasch measurement provides the only established means of constructing linear measures from ordinal responses for latent traits while preserving invariance. The complete absence of Rasch principles from the UK HTA knowledge base indicates that subjective outcomes are treated as quantifiable without transformation. The result is a proliferation of scores, indices, and utilities that look numerical but lack measurement status.

The handling of falsifiability completes the picture. While there is some rhetorical endorsement of rejecting non-falsifiable claims, this endorsement is weak (−0.95) and overridden by strong endorsement of the false claim that reference-case simulations generate falsifiable claims (+2.20). This contradiction is not resolved within the UK framework because falsifiability has been redefined. A model is treated as informative if it is transparent, sensitivity-tested, and reproducible, even if its outputs cannot be empirically refuted. Error is absorbed into recalibration; disagreement is managed procedurally.

This is where the UK's global influence matters. By embedding this framework in national guidance and exporting it through training, consultancy, and academic publication, the UK effectively set the template for HTA worldwide. Other jurisdictions did not independently rediscover false measurement; they inherited it. The UK case is therefore not merely one national example among many. It is the origin point of a memeplex that has become globally stabilized.

The knowledge base that sustains this memeplex spans policy documents, academic HTA centers, textbooks and journals aligned with UK practice. Across this ecosystem, measurement axioms are not enforced as constraints. Instead, they are treated as abstract concerns peripheral to "real-world decision making." This framing is itself revealing. It presupposes that scientific validity and practical relevance are separable, and that the latter can be pursued without the former. That presupposition is incompatible with the history of science.

It is important to stress that this critique is not retrospective moralizing. The axioms of representational measurement were well established long before the UK HTA framework was consolidated. Stevens' scale typology, conjoint measurement theory, and Rasch measurement were all available. The failure to engage them reflects a choice, often implicit, sometimes explicit to prioritize administrative closure over epistemic discipline.

The canonical assessment makes clear that this choice has consequences. A framework that does not recognize measurement constraints cannot generate cumulative knowledge. It can generate consensus, comparability, and an appearance of rigor, but it cannot learn in the Popperian sense. Claims are not exposed to the risk of being wrong; they are refined within a closed system.

In this respect, the United Kingdom's HTA framework represents a unique episode in the annals of applied science. It is not unique because it made errors, many sciences do, but because it institutionalized an evaluative architecture that accommodates multiple independent violations of measurement simultaneously and then treated that architecture as a scientific gold standard. The result is a vast literature of therapy impact claims whose numerical sophistication masks an absence of measurement.

The implication is not that UK HTA requires incremental reform. The implication is that its foundational evaluative framework is unsound. Repair is not possible because the framework's defining features, multiattribute utilities, QALYs, reference-case simulation are the sources of the failure. Replacing them with measurement-valid alternatives would dismantle the system's logic of closure.

A scientifically defensible successor would look very different. It would abandon the search for a single summary index and instead adopt a portfolio of single-attribute claims, each supported by lawful measurement and explicit protocols. Manifest outcomes would be reported on linear ratio scales; latent traits would be assessed using Rasch logit ratio measures. Claims would be provisional, evaluable, and revisable in light of evidence. Such a framework would be less tidy, less convenient, and more demanding, but it would be scientific.

The UK case therefore matters not only as a historical origin but as a cautionary example. It shows how easily the language of science can be appropriated to support a system of numerical belief, and how difficult it is to dislodge that system once it is institutionally entrenched. The canonical logit profile does not indict individuals or institutions; it diagnoses a knowledge base. And that diagnosis is unambiguous: foundational measurement principles do not operate as binding constraints in UK HTA. Until they do, the framework remains a paradigmatic example of false measurement elevated to policy orthodoxy.

A CLEANSING OF THE STABLES: THE GLOBAL PERSPECTIVE

The question that follows from the United Kingdom's role in shaping modern health technology assessment is unavoidable: to what extent can the global acceptance of the UK HTA model be challenged, and on what grounds? The metaphor of cleansing the stables is apt because the problem is not a single defect, or even a small number of methodological errors, but the accumulation of practices that have been normalized, institutionalized, and defended across jurisdictions for

decades. What began as a national evaluative framework has become a global orthodoxy, reproduced through guidelines, journals, academic training, and consultancy, largely without scrutiny of its measurement foundations.

The diffusion of the UK model occurred not because it solved a measurement problem, but because it offered decisional closure. Faced with rising costs and competing technologies, health systems required a standardized, seemingly objective method for ranking interventions. Cost-per-QALY analysis, anchored in preference-based multiattribute instruments and reference-case simulations, delivered precisely that. Its appeal lay in its administrative convenience and apparent neutrality. Once adopted by influential agencies and endorsed by international organizations, it acquired the status of best practice rather than a provisional construct.

The difficulty in challenging this global acceptance lies in the way the framework insulates itself from refutation. By grounding claims in simulation rather than observation, by treating composite indices as measures, and by redefining uncertainty as a parameter to be explored rather than a reason to reject a claim, the model avoids the conditions under which it could be shown to be wrong. Critiques grounded in representational measurement, scale theory, or falsifiability are dismissed as abstract, impractical, or irrelevant to policy needs. In effect, the framework has substituted procedural legitimacy for scientific accountability.

The canonical logit assessments across countries and journals reveal that this is not a localized failure. The repeated collapse of measurement axioms to floor values is observed in jurisdictions with very different health systems, political cultures, and institutional arrangements. What unites them is not shared evidence, but shared belief. The global HTA enterprise operates as a memeplex: a socially stabilized system of ideas that defines what counts as evidence and excludes alternatives by default. Within such a system, the question is no longer whether the framework is valid, but whether it is accepted.

Yet history suggests that such systems are not immune to challenge. Scientific revolutions do not begin with consensus; they begin with the recognition that foundational assumptions are untenable. The critique of false measurement strikes at precisely this level. It does not propose a better model or a more refined simulation; it questions the legitimacy of modeling as a substitute for measurement. By reasserting non-negotiable constraints—unidimensionality, admissible arithmetic, ratio scale requirements, and empirical falsifiability—it exposes the global HTA framework as a construction that cannot meet the standards it claims to embody.

The implications of this challenge are profound. If accepted, it would require health systems to abandon the search for a single summary metric and instead adopt portfolios of evaluable claims, each grounded in lawful measurement and explicit protocols. Decisions would become less tidy, less comparable, and more context-dependent. But they would also become empirically accountable. Claims could fail. Evidence could accumulate. Learning could occur.

Whether such a cleansing is possible depends less on technical feasibility than on institutional willingness. The global HTA community has invested heavily—intellectually, professionally, and politically—in the UK-derived model. Careers, journals, and decision processes are organized

around it. To relinquish it would require acknowledging that decades of authoritative-looking numbers lack measurement validity. That is a difficult admission for any field.

Nevertheless, the alternative is continued participation in what amounts to a global exercise in numerical storytelling. The stable can remain as it is: orderly, familiar, and fundamentally unsanitary. Or it can be cleared, not by replacing one orthodoxy with another, but by restoring the basic conditions of science. The question is no longer whether the UK HTA model can be challenged in principle. The evidence shows that it can. The question is whether the global HTA community is prepared to accept the consequences of that challenge.

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