MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

UNITED KINGDOM: THE ABSENCE OF MEASUREMENT IN HEALTH TECHNOLOGY ASSESSMENT

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

LOGIT WORKING PAPER No 8001 FEBRUARY 2026

www.maimonresearch.com

Tucson AZ

FOREWORD

HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

If the United Kingdom is widely regarded as the birthplace of the modern framework for evaluating therapy impact within health technology assessment, then by the same token it must also be recognized as the birthplace of the global HTA false measurement memeplex that now dominates the field. The methodological architecture developed through institutions such as the National Institute for Health and Care Excellence established the template that many other jurisdictions subsequently adopted. Central to this framework was the use of quality-adjusted life years, cost-effectiveness thresholds, and reference case modeling as the principal tools for comparing therapies and guiding reimbursement decisions. Over time this approach achieved extraordinary institutional success. It became embedded in national HTA agencies, academic research programs, journal publications, and professional training, eventually spreading across Europe, North America, Australia, and parts of Asia.

Yet the same framework that achieved this international influence rests on measurement assumptions that have never been examined against the axioms of representational measurement. Utility scores derived from ordinal preference data are routinely treated as interval quantities, multiplied by time to generate QALYs, and incorporated into simulation models projecting therapy impact far beyond observed evidence. The result is an analytical system that appears quantitatively rigorous but whose core constructs lack demonstrated measurement properties. The present logit analysis therefore begins where the global HTA framework itself began: with the United Kingdom knowledge base that first institutionalized the numerical conventions now reproduced throughout the international HTA literature.

The objective of this study is to examine whether the United Kingdom health technology assessment (HTA) knowledge base recognizes and applies the fundamental principles required for valid measurement. The analysis applies a 24-item canonical statement measurement diagnostic instrument to the HTA literature associated with England, Scotland, Wales, and Northern Ireland. These statements capture the core axioms of representational measurement theory as well as a set of propositions that are demonstrably false from a measurement perspective but widely embedded in HTA practice. Using AI large language model (LLM) interrogation, endorsement probabilities are generated for each statement and transformed into normalized logits. The resulting logit profile provides a diagnostic representation of whether the HTA knowledge base reinforces true measurement principles or instead reinforces analytical conventions that violate the axioms of measurement. The purpose of the study is not to evaluate individual decisions or agencies but to assess whether the conceptual structure of the UK HTA framework recognizes the conditions required for measurement-valid therapy impact assessment.

The results reveal a consistent pattern that has appeared across other jurisdictions examined in this series of interrogations. Statements that violate the axioms of representational measurement receive strong positive reinforcement within the United Kingdom HTA knowledge base, while statements expressing fundamental measurement principles appear largely absent. In particular, propositions consistent with the conventional QALY framework such as treating preference-weighted utility scores as interval measures and combining them multiplicatively with time are strongly endorsed. At the same time, statements asserting that measurement must precede arithmetic, that multiplication requires ratio scale properties, or that ordinal responses require Rasch transformation for valid measurement receive strongly negative logits. The resulting pattern represents what can be described as a measurement inversion: false measurement propositions are structurally embedded within the knowledge base while true measurement axioms are not. The implication is that the quantitative framework used in UK HTA relies heavily on numerical conventions whose measurement status is only appropriate as numerical storytelling for therapy impact claims..

The starting point is simple and inescapable: *measurement precedes arithmetic*. This principle is not a methodological preference but a logical necessity. One cannot multiply what one has not measured, cannot sum what has no dimensional homogeneity, cannot compare ratios when no ratio scale exists. When HTA multiplies time by utilities to generate QALYs, it is performing arithmetic with numbers that cannot support the operation. When HTA divides cost by QALYs, it is constructing a ratio from quantities that have no ratio properties. When HTA aggregates QALYs across individuals or conditions, it is combining values that do not share a common scale. These practices are not merely suboptimal; they are mathematically impossible.

The modern articulation of this principle 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 NATIONAL UNITED KINGDOM HTA KNOWLEDGE BASE

The United Kingdom occupies a central position in the development and global diffusion of health technology assessment. The institutional framework for HTA in the UK is anchored primarily by the National Institute for Health and Care Excellence (NICE), which produces guidance for England and whose methodological framework has strongly influenced HTA systems worldwide. Complementary structures exist in the devolved administrations through the Scottish Medicines Consortium (SMC) in Scotland and the All Wales Medicines Strategy Group (AWMSG) in Wales, with Northern Ireland generally drawing upon NICE guidance. Together these institutions form a coordinated policy environment in which the evaluation of new medicines and health technologies is conducted through formal pseudo-economic and clinical assessment.

The analytical framework used across this system is built around cost-effectiveness analysis using quality-adjusted life years. The QALY serves as the central metric for comparing therapies and informing reimbursement and access decisions. Clinical trial data, observational evidence, and epidemiological studies are synthesized to estimate the incremental health benefits associated with new technologies. These benefits are expressed as gains in QALYs relative to existing treatments and compared against estimated costs to produce incremental cost-effectiveness ratios. Policy decisions are then guided by threshold values that reflect judgments regarding the acceptable cost of generating additional health gains within the healthcare system.

A defining characteristic of the UK HTA knowledge base is the reliance on preference-based measures of health-related quality of life to generate composite ordinal utility values. Instruments such as the EQ-5D are widely used to derive preference weights that represent the perceived value of different health states. These weights are obtained from population-based preference studies using techniques such as time trade-off or standard gamble. The resulting preference scores are treated as interval quantities that can be combined with time to produce QALYs. This approach allows health outcomes across different diseases to be expressed within a common evaluative framework, facilitating comparisons between therapies that affect different aspects of health.

The methodological framework supporting this approach is reinforced through guidance documents, academic research programs, and journal publications. NICE methodological guides provide detailed instructions regarding the construction of cost-effectiveness models, the use of preference-based utility measures, and the application of discounting and sensitivity analysis. Academic centers specializing in health economics contribute to the development of modeling techniques and the analysis of cost-effectiveness data. The resulting body of literature forms a highly coherent knowledge base in which the principles of cost-utility analysis are widely accepted while the axioms of representational measurement remain unknown.

Simulation modeling plays a central role in this analytical environment. Because clinical trials rarely extend over the full duration of a patient’s lifetime, models are used to extrapolate outcomes beyond the observed data. Markov models and related simulation techniques estimate long-term survival, disease progression, and quality-of-life trajectories. These projections generate estimates of lifetime QALY gains and associated costs, which are then used to calculate cost-effectiveness ratios. Sensitivity analyses are typically conducted to examine the robustness of results under alternative assumptions.

The coherence of the UK HTA knowledge base reflects decades of methodological development and institutional reinforcement of false measurement. Training programs in health economics reproduce the same analytical framework, and journals specializing in pharmacoeconomics and HTA provide outlets for research employing similar methods. As a result, the NICE reference case has become a template adopted by numerous HTA agencies internationally. Many countries have implemented guidelines that closely mirror the UK approach, thereby extending its influence beyond national boundaries.

Despite this institutional coherence, the measurement foundations of the framework are rarely examined explicitly within the HTA literature. The arithmetic operations used to construct QALYs, multiplying utility scores by time and aggregating the resulting values, require that the underlying scales satisfy the properties of ratio measurement and dimensional homogeneity. Yet preference-based utility scores are derived from ordinal response structures and composite health state descriptions. Their scale properties are typically assumed rather than demonstrated. As a consequence, the false measurement status of the quantities used in cost-effectiveness analysis remains implicit.

The UK HTA knowledge base represents what advocates would describe an analytically sophisticated but in fact is a conceptually meaningless framework. Certainly, there is a systematic organization of evidence within the reference case framework and the transparent procedures used to evaluate therapies. At the same time, the quantitative constructs on which the framework depends have evolved independently of the principles of representational measurement. The logit analysis provides a means of examining this tension by identifying whether the knowledge base itself recognizes the axioms required for measurement. The resulting profile suggests that while the UK HTA system excels in institutional coordination and methodological consistency of false measurement, its quantitative framework remains detached from the measurement foundations necessary for representing therapy impact as a measurable attribute. A wasted 40 years.

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: ABSOLUTE MEASUREMENT FAILURE

The United Kingdom occupies a central position in the historical development of health technology assessment. Institutions such as the National Institute for Health and Care Excellence (NICE), together with the Scottish Medicines Consortium (SMC), the All Wales Medicines Strategy Group (AWMSG), and the advisory structures operating in Northern Ireland, have played a formative role in shaping the global HTA framework. The analytical methods associated with these institutions, particularly cost-effectiveness analysis based on quality-adjusted life years (QALYs) have become the dominant paradigm adopted across many jurisdictions. Because of this influence, the United Kingdom provides an especially important case for examining whether the conceptual foundations of HTA recognize the principles required for valid measurement. The logit assessment is presented in Table 1.

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	0.15	-1.60
MULTIPLICATION REQUIRES A RATIO MEASURE	1	0.10	-2.20
TIME TRADE-OFF PREFERENCES ARE UNIDIMENSIONAL	0	0.75	+1.10
RATIO MEASURES CAN HAVE NEGATIVE VALUES	0	0.80	+1.40
EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES	0	0.80	+1.40
THE QALY IS A RATIO MEASURE	0	0.75	+1.10
TIME IS A RATIO MEASURE	1	0.95	+2.50
MEASUREMENT PRECEDES ARITHMETIC	1	0.15	-1.60
SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES	0	0.80	+1.40
MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC	1	0.15	-1.60
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	0.80	+1.40
THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE	0	0.75	+1.10
CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.10	-2.20
QALYS CAN BE AGGREGATED	0	0.75	+1.10

NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.70	+0.85
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.70	+0.85
THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.60	+0.40
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	0.20	-1.40
THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.05	-2.50

.

The present analysis applies the 24-item canonical measurement diagnostic instrument to the United Kingdom HTA knowledge base, combining the literature and methodological guidance emerging from England, Scotland, Wales, and Northern Ireland. The objective is not to evaluate individual policy decisions or agency procedures. Rather, the purpose is to determine whether the knowledge base that supports UK HTA recognizes the axioms required for measurement within the framework of representational measurement theory. The resulting logit profile reveals a pattern that has appeared repeatedly across other jurisdictions examined in this series of interrogations. The reinforcement structure of the knowledge base does not align with the requirements of measurement. Instead, false measurement propositions receive strong endorsement, while statements reflecting foundational measurement principles appear largely absent.

The central organizing construct of the United Kingdom HTA framework is the quality-adjusted life year. Within the NICE reference case, therapeutic value is expressed as the incremental cost per QALY gained relative to an alternative intervention. QALYs are derived by combining time with preference-based utility scores obtained from instruments such as the EQ-5D. From a policy perspective this structure offers an attractive administrative convenience. It allows diverse clinical outcomes to be summarized in a single metric and compared across disease areas using explicit cost-effectiveness thresholds. This approach has contributed significantly to the global influence of the UK HTA system and the role of false measurement.

The measurement implications of this framework, however, are rarely examined. The QALY assumes that utility scores represent quantities that can legitimately be multiplied by time and aggregated across individuals. This assumption requires that the utility scale possess ratio properties and dimensional homogeneity. If these properties are absent, which they are, the arithmetic operations used to construct QALYs become mathematically indefensible. The logit profile indicates that the United Kingdom HTA knowledge base strongly reinforces statements consistent with the conventional QALY framework. Propositions suggesting that preference-weighted health state values generate interval measurement receive strong positive logits, while the statement that the QALY functions as a ratio measure is also widely reinforced. These results reflect the deep institutional embedding of the QALY or false measurement paradigm within UK HTA.

At the same time, the logit assessment reveals weak reinforcement of fundamental measurement principles. Statements asserting that measurement must precede arithmetic receive strongly negative logits, as do statements indicating that multiplication requires ratio scale properties. These propositions represent elementary axioms of representational measurement theory. Their absence from the knowledge base suggests that the arithmetic operations used in cost-effectiveness analysis are rarely examined in relation to the scale properties of the variables involved. The knowledge base therefore operates with implicit assumptions about measurement that are not explicitly recognized or even justified.

A similar pattern appears in relation to the transformation of subjective responses into quantitative measures. Statements indicating that ordinal responses require Rasch transformation to produce interval measurement receive the lowest possible logit values, indicating effective non-possession within the knowledge base. Rasch measurement, which provides the only method for constructing invariant measurement scales for latent attributes, is almost entirely absent from the analytical framework of HTA. Instead, ordinal preference scores derived from questionnaires are routinely treated as if they possessed interval or ratio properties. These scores are then manipulated through arithmetic operations to generate QALYs and cost-effectiveness ratios. The fact that summed scores from observations, a mainstay of psychometrics, are inevitably ordinal is never mentioned.

The combined effect of these patterns is the emergence of what can only be described as measurement inversion. False measurement propositions, such as the assumption that ordinal scores can support arithmetic operations, are strongly reinforced within the literature. At the same time, the axioms required for lawful measurement are largely absent. The resulting analytical framework therefore applies arithmetic operations to constructs whose measurement status has never been established. This inversion does not arise from deliberate error on the part of individual researchers. Rather, it reflects the historical evolution of an analytical paradigm that has gradually become institutionalized across agencies, journals, and academic programs. A paradigm built on the false premise that preferences for health statedescriptions must be seen as single attributes.

Another defining feature of the United Kingdom HTA knowledge base is the extensive reliance on simulation modeling. Cost-effectiveness analyses frequently employ Markov models or other simulation frameworks to project long-term costs and health outcomes beyond the time horizon of clinical trials. These models generate estimates of lifetime costs and QALYs based on a combination of empirical data and structural assumptions. The outputs are then used to inform policy decisions regarding pricing, reimbursement, and access to therapies.

From a methodological perspective, however, the reliance on simulation modeling raises questions regarding empirical testability. A fundamental principle of scientific inquiry is that claims must be capable of falsification through observation. Simulation models projecting outcomes decades into the future cannot be directly verified within the timeframe of policy decision making. The logit profile reflects this ambiguity. While the knowledge base strongly reinforces the role of simulation modeling in HTA, the principle that non-falsifiable claims should be rejected receives only moderate endorsement. The analytical framework therefore operates with a relaxed interpretation of falsifiability outside of any notion of the evolution of objective knowledge.

One of the most striking aspects of the United Kingdom HTA system is its institutional coherence. Methodological guidance issued by NICE is reinforced through academic research, professional training programs, and journal publications. The resulting framework is internally consistent and widely accepted within the policy community. This institutional coherence has undoubtedly contributed to the global influence of the UK HTA false measurement memeplex model. Many countries have adopted methodological guidelines closely aligned with the NICE reference case, thereby reproducing the same analytical structure across multiple jurisdictions. This is demonstrated without exception for the over 100 logit LLM interrogations reported on here indicating a global lack of awareness of the axioms of representational measurement.

Institutional coherence, however, should not be confused with measurement validity. The logit assessment indicates that the analytical conventions embedded within the United Kingdom HTA framework are largely independent of the principles required for lawful measurement. Once these conventions became embedded within policy institutions, they acquired authority through repetition and administrative utility rather than through measurement foundations. The resulting knowledge base therefore reinforces a set of analytical practices that appear quantitative but lack the scale properties required for measurement.

The pattern observed in the United Kingdom closely resembles the logit profiles obtained for other HTA jurisdictions. Countries such as Australia, Canada, and several European nations display similar reinforcement of QALY-based frameworks and similar absence of measurement axioms. This convergence reflects the global diffusion of the NICE reference case false measurement memeplex. As agencies adopted similar methodological guidelines, the same measurement assumptions became embedded within the international HTA literature.

The implications of these findings extend beyond methodological debates. Health technology assessment plays a central role in determining patient access to therapies and guiding the allocation of healthcare resources. If the measurement foundations of the analytical framework are false, the numerical outputs used to support policy decisions will lack empirical meaning. The quantitative framework used to summarize these inputs may provide an appearance of precision that is not supported by measurement.

The logit assessment therefore raises a fundamental question regarding the acceptance of the lack of scientific foundations of HTA. A discipline that defines itself through quantitative evaluation must ultimately demonstrate that the quantities it manipulates represent measurable attributes. When arithmetic operations are applied to constructs lacking valid scale properties, the resulting outputs cannot be interpreted as measurements of therapy impact. They function instead as numerical conventions or storytelling within an administrative decision process that has no concept of representational measurement.

Reconstructing HTA on measurement-valid foundations would require distinguishing clearly between manifest and latent attributes. Manifest attributes such as survival time, hospital utilization, or treatment duration can be represented using linear ratio scales that preserve empirical relationships. Latent attributes such as pain severity or functional impairment require the construction of invariant measurement scales through Rasch modeling. Only within such a framework can therapy impact be represented quantitatively in a manner consistent with the axioms of representational measurement.

The United Kingdom HTA knowledge base remains one of the most influential analytical frameworks in global health policy. Yet the logit profile shows that the conceptual foundations of this framework have developed independent of measurement theory. The resulting pattern of measurement inversion with strong reinforcement of false measurement propositions combined with the absence of fundamental measurement principles raises unavoidable questions about the pseudo-scientific status of the quantities used in HTA. Whether the field chooses to address these questions remains uncertain. What the logit evidence demonstrates, however, is that the measurement assumptions underlying contemporary HTA ensure it remains a non-science

III. 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