MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

UNITED KINGDOM: THE NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE AND THE ABSENCE OF REPRESENTATIONAL MEASUREMENT

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

LOGIT WORKING PAPER No 701 FEBRUARY 2026

www.maimonresearch.com

Tucson AZ

FOREWORD

HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

The National Institute for Health and Care Excellence (NICE) was established to provide authoritative guidance on the clinical and cost-effectiveness of health technologies within the National Health Service (NHS) in England and Wales. Its central aim in the context of health technology assessment (HTA) is to evaluate new pharmaceuticals, medical devices, diagnostics, and clinical interventions and to determine whether they represent value for money within a publicly funded health system operating under budget constraints. NICE does this primarily through its reference case framework, which requires the estimation of incremental cost-effectiveness ratios—most commonly expressed as cost per quality-adjusted life year (QALY) gained.

Beyond individual reimbursement decisions, NICE has had a profound systemic impact. Its methodological guidance has become a global benchmark, influencing HTA agencies across Europe, Australasia, and parts of Asia and North America. Academic programs in health economics routinely train students using NICE's framework. Pharmaceutical manufacturers structure submissions around its evidentiary and modeling requirements. In this way, NICE has not only shaped NHS decision-making but has also played a central role in institutionalizing the dominant paradigm of cost-utility analysis as the standard quantitative approach to evaluating therapy impact in modern HTA.

The objective of this study was to interrogate the NICE knowledge base using the standardized 24-item canonical statement instrument grounded in the axioms of representational measurement theory. NICE occupies a defining role in global health technology assessment (HTA), and its reference case framework for cost-effectiveness analysis has shaped reimbursement decisions, academic training, and methodological standards across multiple jurisdictions. The purpose of this interrogation was not to evaluate individual appraisal decisions, but to assess whether the underlying quantitative architecture embedded in NICE's guidance explicitly recognizes and enforces the foundational requirements of lawful measurement. In particular, the study sought to determine whether NICE's framework establishes unidimensionality, dimensional homogeneity, and ratio-scale admissibility prior to arithmetic operations such as multiplication, aggregation, and incremental comparison, and whether latent constructs are transformed through invariant measurement models before entering economic evaluation.

The logit profile demonstrates systematic weak reinforcement of foundational measurement axioms and collapse of Rasch-related statements to floor-level logits, indicating structured non-possession of invariant latent trait measurement principles within the NICE knowledge base. At the same time, manifest ratio properties, simulation modeling conventions, and QALY-based arithmetic receive strong reinforcement. The pattern reveals not random inconsistency but structural asymmetry: ratio constraints are recognized for observable quantities such as time, yet not enforced for latent constructs central to cost-utility analysis. Arithmetic operations are institutionalized within the reference case framework without prior demonstration that the constructs entering those operations satisfy the scale properties required by representational measurement. The findings therefore indicate that NICE's quantitative architecture is modeling-centered and administratively disciplined but weakly anchored in measurement-first principles.

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 NICE HTA KNOWLEDGE BASE

The knowledge base of NICE is defined by its reference case methodology for economic evaluation. Central to this framework is the requirement that technology appraisals estimate incremental cost-effectiveness ratios, typically expressed as cost per quality-adjusted life year (QALY) gained. Manufacturers submitting technologies for appraisal must construct decision-analytic models—often Markov state-transition models or patient-level simulations—that integrate clinical effectiveness data, survival estimates, healthcare resource utilization, and preference-based utility weights. These models project costs and outcomes over defined time horizons, frequently extending beyond observed trial data, and incorporate probabilistic sensitivity analysis to characterize parameter uncertainty.

Utility values used within this framework are commonly derived from preference-based instruments such as the EQ-5D. These instruments collect patient-reported responses across multiple health dimensions and apply societal preference weights derived through time trade-off or related elicitation techniques. The resulting index values are treated as quantitative parameters suitable for multiplication with time to generate QALYs. QALYs are then aggregated across individuals and compared incrementally across treatment alternatives. NICE's appraisal committees consider these incremental ratios in relation to implicit or explicit cost-effectiveness thresholds when formulating recommendations for NHS reimbursement.

The methodological guidance emphasizes transparency, consistency, and comparability across submissions. Sensitivity analyses, scenario analyses, and structural model validation are required. Assumptions must be justified, and alternative model structures explored. The reference case defines preferred data sources, discount rates, and analytical perspectives. In this sense, the NICE knowledge base is highly structured. It establishes clear procedural rules and enforces analytic discipline in model construction and reporting.

However, the knowledge base prioritizes modeling coherence and policy relevance rather than explicit demonstration of measurement validity for latent constructs. Utility scores derived from ordinal questionnaire responses are accepted as quantitative inputs once processed through preference algorithms. The framework does not require demonstration that these utility values possess invariant unit structure or ratio-scale properties prior to arithmetic use. The multiplication of utility weights by time to generate QALYs is treated as methodologically standard rather than as an operation requiring prior validation of scale admissibility.

Similarly, while NICE guidance emphasizes internal and external validity of clinical data, uncertainty analysis, and transparency of assumptions, it does not embed representational measurement axioms as explicit constraints. Unidimensionality of latent constructs is not systematically enforced as a prerequisite for aggregation. Dimensional homogeneity is presumed within QALY construction rather than demonstrated through measurement modeling. Rasch transformation or invariant logit scaling does not function as a required step in latent trait quantification.

The knowledge base therefore reflects a mature decision-analytic culture built around cost-utility modeling and threshold-based reimbursement decisions. It is procedurally rigorous, internationally influential, and administratively stable. Yet its quantitative foundation rests on the conventional treatment of composite preference scores as if they were admissible for ratio-level arithmetic without requiring demonstration of lawful measurement structure. The interrogation does not suggest absence of methodological sophistication; rather, it shows that the conceptual ordering of modeling and measurement places arithmetic first and measurement validation second, or not at all.

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: NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE HTA KNOWLEDGE BASE

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.

NICE: THE ABSENCE OF REPRESENTATIONAL MEASUREMENT AND THE ENDORSEMENT OF FALSE MEASUREMENT

The National Institute for Health and Care Excellence (NICE) is widely regarded as one of the most influential health technology assessment bodies in the world. Its methodological guidance shapes reimbursement decisions, informs international HTA agencies, and defines the structure of cost-effectiveness submissions. NICE is not peripheral to the global HTA architecture; it is one of its central architects. For that reason, interrogation of its knowledge base using the 24-item canonical measurement instrument carries more than structural significance; it has global implications for the promotion of false measurement in health care decision making (Table 1).

TABLE 1: ITEM STATEMENT, RESPONSE, ENDORSEMENT AND NORMALIZED LOGITS NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE HTA KNOWLEDGE BASE

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	-1.75
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	0.85	+1.75
THE QALY IS A RATIO MEASURE	0	0.80	+1.40
TIME IS A RATIO MEASURE	1	0.95	+2.50
MEASUREMENT PRECEDES ARITHMETIC	1	0.10	-2.20
SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES	0	0.85	+1.75
MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC	1	0.10	-2.20
THERE ARE ONLY TWO CLASSES OF MEASUREMENT LINEAR RATIO AND RASCH LOGIT RATIO	1	0.05	-2.50
TRANSFORMING SUBJECTIVE RESPONSES TO INTERVAL MEASUREMENT IS ONLY POSSIBLE WITH RASH RULES	1	0.05	-2.50
SUMMATION OF LIKERT QUESTION SCORES CREATES A RATIO MEASURE	0	0.85	-1.75
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	0.85	+1.75

NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.60	+0.40
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.80	+1.40
THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.70	+0.85
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	0.70	+0.85
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 logit profile reveals a pattern that is neither accidental nor ambiguous. It shows systematic inversion of representational measurement principles at the core of NICE's evaluative framework.

Foundational axioms receive strongly negative logits. "Measurement precedes arithmetic" registers −2.20. "Meeting the axioms of representational measurement is required for arithmetic" also −2.20. "Multiplication requires a ratio measure" collapses to −2.20. "Measures must be unidimensional" falls to −1.73. These are not obscure technical propositions; they are the structural rules that govern lawful quantitative reasoning. Their consistent negative reinforcement indicates that within NICE's knowledge base these principles are not embedded as binding methodological constraints.

The Rasch-related statements collapse completely. "There are only two classes of measurement: linear ratio and Rasch logit ratio" receives −2.50. "Transforming subjective responses to interval measurement is only possible with Rasch rules" −2.50. "The Rasch logit ratio scale is the only basis for assessing therapy impact for latent traits" −2.50. "The Rasch rules for measurement are identical to the axioms of representational measurement" −2.50. These floor-level logits indicate structured non-possession. Rasch measurement is not marginal within NICE's framework; it is absent.

In contrast, manifest ratio properties are strongly reinforced. "Time is a ratio measure" receives +2.50. NICE's framework fully recognizes the ratio structure of survival time and observable resource use. The asymmetry between manifest and latent domains is therefore sharp and unmistakable. Where quantities are classical and physical, scale properties are respected. Where constructs are latent scale constraints are not structurally enforced.

The positive logits for several scale-related statements reveal a further structural feature. “Summation of Likert question scores creates a ratio measure” receives +1.73. “EQ-5D preference algorithms create interval measures” +1.73. “QALYs can be aggregated” +1.73. “The QALY is a ratio measure” +1.39. These logits do not imply alignment with representational measurement. They indicate reinforcement of NICE’s internal interpretive framework. NICE does not openly assert crude mathematical impossibilities; it operates within a convention in which utility scores are treated as suitable inputs for arithmetic modeling once embedded within the reference case structure.

The key structural contradiction lies here: multiplication of time by utility is the defining operation of the QALY. Yet “Multiplication requires a ratio measure” receives −2.20. This is not a minor inconsistency. It is a direct violation of arithmetic admissibility. If ratio properties are not demonstrably present in the utility component, multiplication with time lacks measurement legitimacy. The logit profile shows that NICE’s knowledge base does not embed the ratio requirement as a constraint.

“Measurement precedes arithmetic” at −2.20 reveals the deeper inversion. NICE’s methodology begins with modeling. Utility weights are inserted into state-transition models. Time horizons extend decades beyond observed data. Incremental cost-effectiveness ratios are generated through simulation. The arithmetic is elaborate, disciplined, and standardized. Yet the scale properties of the latent constructs driving the arithmetic are not validated within a representational measurement framework. Arithmetic precedes measurement.

The positive logit for “Reference case simulations generate falsifiable claims” (+1.39) indicates reinforcement of NICE’s modeling architecture. Simulation outputs are treated as decision-relevant quantities. Yet simulations are projections conditioned on structural assumptions. They are not directly falsifiable in the Popperian sense because they extend beyond empirical observation. The knowledge base recognizes falsifiability in principle—“Non-falsifiable claims should be rejected” receives +0.41—but does not reconstruct the modeling framework around that requirement.

The logit pattern demonstrates that NICE’s quantitative authority rests on three pillars: manifest ratio quantities, composite utility scoring, and simulation modeling. Only the first of these pillars satisfies representational measurement. The second lacks invariant unit structure. The third generates non-falsifiable projections. The combination produces a numerically stable but measurement-detached system. It is numerical storytelling.

The absence of Rasch measurement is decisive. Without Rasch transformation, ordinal responses remain ordinal. Preference algorithms applied to ordinal responses do not generate invariant measurement scales. They produce value indices. Indices may be internally consistent, but they are not ratio measures. Without ratio properties, aggregation and multiplication lack admissibility. The floor-level logits across Rasch statements confirm that NICE’s framework does not recognize this transformation requirement.

The result is structural quantitative incoherence. NICE enforces strict modeling standards, probabilistic sensitivity analysis, and threshold rules. It requires incremental ratios expressed as cost per QALY. Yet it does not enforce the measurement axioms that would make the QALY a lawful ratio quantity. The arithmetic is precise. The measurement foundation is absent. This is not confusion. It is institutionalized non-measurement.

The global influence of NICE amplifies the significance of this profile. Other agencies model their guidance on NICE's reference case. Academic journals train authors to meet NICE standards. Industry submissions are structured around NICE's methodological framework. The logit interrogation therefore does not identify a localized anomaly. It exposes the core quantitative architecture of modern HTA.

The implications are severe. Therapy impact claims expressed through incremental cost per QALY ratios are treated as quantitative magnitudes guiding resource allocation. Yet the logit evidence shows that the axioms licensing ratio arithmetic are not embedded within the knowledge base. The central currency of reimbursement decision-making rests on constructs whose measurement status is weakly reinforced or absent.

The NICE logit profile does not depict a mature measurement science awaiting refinement. It depicts a stabilized arithmetic system detached from representational constraints. Manifest quantities are measured. Latent constructs are scored. The two are multiplied and simulated as if commensurable. That is not a coherent scientific framework under the rules governing quantitative measurement. It is a stabilized quantitative error sustained by institutional authority, policy repetition, and methodological habit.

The interrogation makes the structural condition visible. Whether reconstruction occurs depends on whether NICE is willing to enforce the rule that measurement must precede arithmetic and that latent traits must be transformed through invariant measurement models before entering economic evaluation. Until then, the arithmetic will remain elaborate, the modeling sophisticated, and the quantitative foundation absent.

THE EVOLUTION OF OBJECTIVE KNOWLEDGE AND THE CLOSED HTA PARADIGM

Scientific progress depends on the evolution of objective knowledge. That evolution requires two structural conditions: falsifiability and measurement validity. Claims must be testable, and the quantities used to express those claims must satisfy the axioms governing arithmetic operations. Without these conditions, numerical outputs may accumulate, but objective knowledge does not evolve.

The logit profiles demonstrate that within the NICE knowledge base and across global HTA more broadly representational measurement is not embedded as a foundational constraint. The arithmetic of cost-utility modeling proceeds without prior validation of scale properties. Simulation models generate outputs that cannot be directly falsified because they extend beyond observable data through structural assumptions. In such a framework, disagreement leads to alternative modeling assumptions, not to decisive empirical refutation. This produces epistemic closure.

When a field operates with constructs that are not measurement-validated, empirical challenge becomes structurally weakened. Competing models can always adjust parameters. Thresholds can be reinterpreted. Sensitivity analyses can expand uncertainty bands. But the foundational question of whether the quantities themselves are lawful measures is never addressed.

The absence of Rasch measurement is particularly significant. Without invariant latent trait measurement, subjective responses remain ordinal. Ordinal scores do not support arithmetic comparison across persons or over time in the manner required for cumulative quantitative science. When arithmetic is performed on such scores, the outputs may be numerically precise but not scientifically grounded.

The evolution of objective knowledge requires that measurement constraints precede modeling architecture. In physics, chemistry, and engineering, arithmetic operations are licensed only after quantities are defined within lawful scale structures. In HTA, the sequence is reversed. Modeling conventions define admissible outputs, and measurement theory is peripheral or absent. The logit interrogation does not show incremental oversight. It shows systemic exclusion of representational measurement principles. Under these conditions, HTA does not evolve toward measurement-valid quantification. It stabilizes around convention.

This is why decades of increasingly sophisticated simulation have not resolved foundational criticism. The structure is closed. The arithmetic continues. The measurement question is not admitted. To reopen the evolution of objective knowledge in HTA requires reinstating measurement as the gatekeeper of arithmetic. Latent constructs must be transformed into invariant measures before entering economic evaluation. Claims must be structured so that empirical refutation is possible, not indefinitely deferred through model revision.

Without that shift, HTA remains numerically elaborate but epistemically static. The appearance of quantitative rigor masks the absence of measurement foundation. Scientific evolution halts when arithmetic is detached from lawful measurement. The logit evidence makes this visible. Whether the field chooses to respond determines whether HTA becomes a measurement science or remains a modeling convention sustained by institutional repetition.

DUTY OF CARE AND THE LEGITIMACY OF THERAPY CHOICE

The consequences of measurement failure in HTA are not abstract. They are embedded in therapy choice, reimbursement restriction, and patient access. When an agency such as NICE recommends for or against adoption of a therapy based on incremental cost per QALY ratios, it is not engaging in academic exercise. It is shaping real clinical pathways, determining which treatments will be available, and which will be denied or delayed. Under those conditions, the quantitative framework used to support such decisions must meet the highest standards of scientific validity.

Duty of care is not limited to clinical competence at the bedside. It extends to the evidentiary standards governing system-level decisions. If arithmetic operations are performed on constructs whose scale properties are not demonstrably ratio-valid, then the resulting quantitative claims cannot be considered lawful measures of therapy impact. They become numerical artifacts. To base treatment access decisions on such artifacts raises a profound ethical issue. It substitutes convention for measurement and modeling for empirical quantity.

The logit interrogation shows that NICE’s knowledge base does not embed representational measurement as a binding constraint. Multiplication is institutionalized without ratio validation. Latent constructs are treated as quantitative magnitudes without invariant transformation. Simulation outputs are projected beyond observable data without falsifiable structure. Under these conditions, cost per QALY thresholds function as decision heuristics rather than as ratios grounded in measurement law.

Duty of care in this context requires that therapy impact claims be falsifiable, replicable, and based on quantities whose arithmetic is admissible. If latent traits cannot be lawfully measured, then arithmetic operations involving them cannot legitimately govern access decisions. To proceed otherwise is to replace measurement with administrative convenience.

This is not an argument against resource allocation. Nor is it an argument against economic analysis. It is an argument that therapy choice must rest on lawful quantification if it is to claim scientific legitimacy. When measurement is absent, decision frameworks may still operate, but they operate without the epistemic safeguards that protect against structural error. If a health system denies a therapy because its cost per QALY exceeds a threshold derived from measurement-invalid arithmetic, the ethical burden is heavy. The issue is not whether decisions must be made. It is whether those decisions are supported by quantities that meet the standards required in any other quantitative discipline. A measurement-valid framework would require that manifest outcomes be assessed on linear ratio scales and latent constructs be transformed through invariant Rasch logit measurement prior to arithmetic. Without that foundation, therapy choice is governed by numerical convention rather than by lawful measurement.

Duty of care therefore demands reconstruction. Not refinement. Not incremental adjustment. Reconstruction. This is the subject of the next section.

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.

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