MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

UNITED KINGDOM: NICE AND THE GLOBAL NUMERICAL STORYTELLING MEMEPLEX

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

LOGIT WORKING PAPER No 26 JANUARY 2026

www.maimonresearch.com

Tucson AZ

FOREWORD

HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

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

Health technology assessment did not become scientifically incoherent by accident. Its present dependence on constructed utilities, composite health state descriptions, and long-horizon simulation models can be traced to a specific institutional turning point. That turning point occurred in the United Kingdom with the emergence of the National Institute for Health and Care Excellence (NICE) and the formal adoption of cost-utility analysis as the dominant framework for evaluating therapies in the late 1990s. What followed was not merely the diffusion of a method, but the global institutionalization of a way of thinking about evidence in which numerical output replaced measurement as the criterion for decision making. NICE did not invent the QALY, but it transformed it from a contested academic construct into a regulatory instrument, thereby providing the foundation for what has since evolved into a global memeplex of numerical storytelling.

The central innovation of the NICE framework was not methodological sophistication but administrative closure. By embedding health state preference values, algorithmic scoring systems, and reference-case simulation models into a standardized decision process, NICE created a system capable of producing definitive reimbursement judgments in the absence of directly evaluable claims. Once a modeled cost per QALY estimate fell beneath an accepted threshold, the assessment was complete. This approach offered policy makers a powerful advantage: it eliminated the need for ongoing empirical testing, replication, or falsification of therapy impact claims over time. Decisions could be made once, numerically justified, and defended procedurally rather than scientifically. In doing so, NICE shifted HTA away from the logic of normal science and toward a framework grounded in plausibility rather than measurement.

This transformation had consequences far beyond the United Kingdom. NICE rapidly became the international reference authority for HTA practice. Its methodological guidelines were emulated, adapted, and institutionalized across jurisdictions, including Australia, Canada, Europe, and later the United States. Journals, academic programs, and consulting organizations aligned themselves with the NICE paradigm, reinforcing its assumptions and extending its reach. Over time, what began as a national administrative solution hardened into an unquestioned global orthodoxy. The arithmetic of cost-effectiveness came to be treated as evidence itself, even though the underlying constructs—utilities, quality-adjusted life years, and composite preference scores—failed the axioms required for measurement.

The result is the contemporary HTA landscape: a system rich in numbers yet barren of measures. The defining feature of this landscape is not disagreement about parameters or modeling technique, but the near-total absence of scrutiny regarding whether the quantities being manipulated possess the properties necessary for meaningful arithmetic. Measurement precedes arithmetic in every empirical science. In HTA, that ordering was reversed. NICE stands at the origin of this inversion. Its legacy is not simply a set of guidelines, but the normalization of numerical storytelling as a substitute for evaluable, falsifiable claims about therapy impact.

The objective of this assessment is to examine the extent to which we can determine quantitatively the extent to which NICE operates within a framework that satisfies the axioms of representational measurement, which define the conditions under which numerical claims may legitimately support scientific inference and policy decision making. Using a canonical 24-item diagnostic instrument, this analysis evaluates the extent to which NICE's methods acknowledge the precedence of measurement over arithmetic, distinguish between manifest and latent attributes, and apply admissible scale types when constructing and interpreting therapy impact claims. The purpose is not to critique individual appraisal decisions, but to interrogate the epistemic structure that governs NICE's evaluative architecture and determines what forms of evidence are considered admissible.

The findings demonstrate a consistent and highly structured pattern. Propositions that would require demonstrable measurement status prior to arithmetic are rejected at or near the floor of the logit scale, while propositions that enable the continued operation of the NICE reference case model are endorsed at the ceiling. In particular, linear ratio measurement for manifest attributes and Rasch logit ratio measurement for latent traits are absent as gatekeeping requirements. In their place, NICE endorses valuation-based constructs, utility algorithms, QALY aggregation, and simulation-based inference as legitimate quantitative evidence. The result is an evaluative system that is internally coherent yet externally indefensible in measurement terms, one in which arithmetic is permitted in advance of, and independent from, demonstrable measurement.

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

The foundational theory that establishes *when* and *whether* a set of numbers can be interpreted as measurements came with the publication of Krantz, Luce, Suppes, and Tversky's *Foundations of*

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

Whereas representational measurement theory articulates the axioms for interval measurement, Georg Rasch's 1960 model provides the only scientific method for transforming ordered categorical responses into interval measures for latent traits [3]. Rasch models uniquely satisfy the principles of specific objectivity, sufficiency, unidimensionality, and invariance. For any construct such as pain, fatigue, depression, mobility, or need, Rasch analysis is the only legitimate means of producing an interval scale from ordinal item responses. Rasch measurement is not an alternative to RMT; it is its operational instantiation. The equivalence of Rasch's axioms and the axioms of representational measurement was demonstrated by Wright, Andrich and others as early as the 1970s. In the latent-trait domain, the very domain where HTA claims to operate; Rasch is the only game in town [4].

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

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

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

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

The conclusion is unavoidable: HTA does not need incremental reform; it needs a scientific revolution. Measurement must precede arithmetic. Representational axioms must precede valuation rituals. Rasch measurement must replace ordinal summation and utility algorithms. Value claims must be falsifiable, protocol-driven, and measurable; rather than simulated, aggregated, and numerically embellished.

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

Paul C Langley, Ph.D

Email: langleylapaloma@gmail.com

DISCLAIMER

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

1. INTERROGATING THE LARGE LANGUAGE MODEL

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

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

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

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

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

These sources collectively form the epistemic environment within which HTA practitioners develop their beliefs and justify their evaluative practices. The boundary of interrogation is thus not the whole of medicine, economics, or public policy, but the specific textual ecosystem that sustains HTA reasoning. . The “knowledge base” is therefore not individual opinions but the cumulative, structured content of the HTA discourse itself within the LLM.

THE NICE KNOWLEDGE BASE

The knowledge base of NICE can be characterized as a closed, self-reinforcing system organized around valuation rather than measurement. Its foundational assumption is that health states can be numerically represented through preference elicitation and that these numbers may then be manipulated arithmetically to support comparative judgments. This assumption defines the boundaries of admissible inquiry. Questions concerning the empirical structure of attributes, the existence of invariant units, or the permissibility of arithmetic operations are not treated as prerequisites but as settled matters resolved by convention.

Within this knowledge base, numerical representation is equated with quantification. The presence of numbers is taken as sufficient evidence that measurement has occurred. Consequently, the distinction between ordering and measuring is effectively erased. Ordinal preference responses are transformed through scoring algorithms and treated as interval or ratio measures without demonstration of equal units or true zero properties. The resulting values are then multiplied by time, aggregated across individuals, and compared across disease areas as if they belonged to a common quantitative scale.

Latent attributes occupy a central role in NICE evaluations, yet they are never formally measured. Constructs such as health-related quality of life, wellbeing, and burden are invoked as if they were quantities, but no measurement model capable of generating invariant units is required. Rasch measurement, which alone provides a logit ratio scale representing possession of a latent trait, is absent from NICE’s methodological core. This absence is decisive. Without Rasch transformation, subjective responses remain ordinal, regardless of the sophistication of subsequent statistical treatment.

Instead of measurement, NICE relies on valuation. Health state descriptions are valued by external populations, converted into preference weights, and treated as quantitative modifiers of survival. This process does not construct a measure of health; it constructs a numerical index reflecting social preferences over descriptive states. Yet within the NICE knowledge base, this distinction is never operationalized. Valuation outputs are treated as if they were measures of patient experience, enabling arithmetic operations that measurement theory explicitly disallows.

The reference case model functions as the organizing mechanism of this system. It provides a standardized template within which all submissions must be expressed. This template does not test claims against observable outcomes; it generates conditional projections based on assumptions. Sensitivity analyses explore internal variation, not empirical refutation. Nevertheless, these outputs are treated as evidence sufficient for coverage and pricing decisions.

What defines the NICE knowledge base most clearly is not its explicit methodological commitments, but its patterned exclusions. Representational measurement theory is absent as a governing framework. Scale-type admissibility is not treated as a threshold requirement. The question of whether an outcome is measurable is never posed prior to analysis. This silence allows the system to maintain apparent rigor while insulating its core constructs from invalidation.

Over time, this architecture has produced stability rather than knowledge growth. Appraisal procedures have become more elaborate, models more complex, and guidance more detailed, yet none of these developments address the foundational issue of whether the quantities being manipulated exist as measures at all. The system evolves procedurally while remaining epistemically static.

NICE's knowledge base therefore does not operate as a mechanism of normal science. It does not generate provisional claims subject to falsification through measurement. Instead, it produces administratively complete narratives that permit closure. Decisions are made, thresholds applied, and cases resolved without the possibility of empirical refutation over time.

In this sense, NICE exemplifies a mature HTA memeplex. Its durability rests not on empirical validation, but on institutional reinforcement. Journals, agencies, consultants, and manufacturers reproduce the same assumptions because deviation would render participation impossible. Measurement is not rejected explicitly; it is rendered irrelevant.

The consequence is that NICE evaluates therapies through numerical storytelling rather than quantified evidence. Its outputs appear scientific, yet they lack the conditions required for measurement-based inference. The 24-item diagnostic makes this unavoidable: NICE's authority derives not from measurement, but from the normalization of arithmetic without it.

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 INSTUITUTE OF HEALTH AND CARE EXCELLENCE

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 ADMINISTRATIVE PERFECTION OF ARITHMETIC WITHOUT MEASUREMENT

If you want a single jurisdiction that demonstrates the HTA memeplex in its most disciplined, institutionalized form, it is the United Kingdom. Not because the UK is uniquely careless, but because it is uniquely confident. It has converted a set of mathematically inadmissible constructs utilities, QALYs, and reference-case simulations into an administrative machine for decision closure. The canonical profile in Table 1 shows the UK system does not merely tolerate the inversion of representational measurement; it operationalizes it as governance. That is why this table matters. It is not a list of technical quibbles. It is a map of what the system must believe in order to function.

TABLE 1: ITEM STATEMENT, RESPONSE, ENDORSEMENT AND NORMALIZED LOGITS NATIONAL INSTITUTE OF HEALTH AND CARE EXCELLENCE (NICE)

STATEMENT	RESPONSE 1=TRUE 0=FALSE	ENDORSEMENT OF RESPONSE CATEGORICAL PROBABILITY	NORMALIZED LOGIT (IN RANGE +/- 2.50)
INTERVAL MEASURES LACK A TRUE ZERO	1	0.30	-0.85
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	0.90	+2.20
RATIO MEASURES CAN HAVE NEGATIVE VALUES	0	0.90	+2.20
EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES	0	0.95	+2.50
THE QALY IS A RATIO MEASURE	0	0.95	+2.50
TIME IS A RATIO MEASURE	1	0.95	+2.50
MEASUREMENT PRECEDES ARITHMETIC	1	0.10	-2.20
SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES	0	0.90	+2.20
MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC	1	0.10	-2.20
THERE ARE ONLY TWO CLASSES OF MEASUREMENT LINEAR RATIO AND RASCH LOGIT RATIO	1	0.05	-2.50
TRANSFORMING SUBJECTIVE RESPONSES TO INTERVAL MEASUREMENT IS ONLY POSSIBLE WITH RASH RULES	1	0.05	-2.50
SUMMATION OF LIKERT QUESTION SCORES CREATES A RATIO MEASURE	0	0.90	+2.20
THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE	0	0.85	+1.75
CLAIMS FOR COST-EFFECTIVENESS FAIL THE	1	0.15	-1.75

AXIOMS OF REPRESENTATIONAL MEASUREMENT			
QALYS CAN BE AGGREGATED	0	0.95	+2.50
NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.70	+0.85
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.95	+2.50
THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.65	+0.60
THE RASCH LOGIT RATIO SCALE IS THE ONLY BASIS FOR ASSESSING THERAPY IMPACT FOR LATENT TRAITS	1	0.05	-2.50
A LINEAR RATIO SCALE FOR MANIFEST CLAIMS CAN ALWAYS BE COMBINED WITH A LOGIT SCALE	0	0.60	+0.40
THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT	1	0.25	-1.10
THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.05	-2.50

Start with the gatekeeper proposition: measurement precedes arithmetic. In a measurement-literate system this would be non-negotiable. Here it is crushed to p = 0.10 with a canonical logit of −2.20. The companion proposition, meeting the axioms of representational measurement is required for arithmetic, sits at the same floor. These two results alone are enough to convict the entire UK HTA architecture. They tell you that the system does not treat measurement as a prerequisite for calculation. It treats calculation as a method for manufacturing decision outputs, with meaning assumed after the fact.

This is why the UK can simultaneously endorse, at or near the ceiling, every false proposition required to sustain cost-utility analysis. The claim that EQ-5D preference algorithms create interval measures sits at p = 0.95 (+2.50). That is the system announcing, in effect, that an algorithm can confer measurement properties that the empirical attribute does not possess. It is also the system declaring that preference elicitation and scoring rules can substitute for the existence of equal units and invariance. Once that belief is installed, the rest is automatic. The QALY becomes a “ratio measure” at p = 0.95 (+2.50). QALYs can be aggregated at p = 0.95 (+2.50). Reference-case simulations generate falsifiable claims at p = 0.95 (+2.50). Every one of these is a load-bearing falsehood, and the UK system reinforces them not modestly but maximally.

The multiplication rule makes the inversion grotesque rather than merely wrong. Multiplication requires a ratio measure is pushed to p = 0.10 (−2.20). Yet the defining act of the QALY is multiplication: time multiplied by a preference weight. The UK system therefore rejects the rule and mandates the operation. It cannot do otherwise. If the multiplication rule were enforced, the QALY collapses. Cost-per-QALY collapses. Thresholds collapse. The entire decision apparatus loses its numerical backbone. To survive, the rule must be suppressed, and Table 1 shows exactly that suppression.

Time is treated correctly, with p = 0.95 (+2.50). This matters because it proves the system is capable of recognizing ratio measurement when it is inconvenient to deny it. The UK knows exactly what ratio measurement looks like in the physical world. Its failure is not conceptual incapacity; it is selective exemption. Measurement discipline is applied where it does not threaten the memeplex and suspended where it would dissolve the central constructs.

The next major fault line is unidimensionality. Measures must be unidimensional is weakly endorsed at p = 0.15 (−1.75), while composite health state descriptions as time trade-off preferences are asserted to be unidimensional by endorsing the false statement at p = 0.90 (+2.20). That is the UK's characteristic move: reject unidimensionality as a requirement, then assume unidimensionality when needed. In other words, unidimensionality is not treated as a property to be demonstrated; it is treated as a convenience label applied to preference outputs because the arithmetic requires a single continuum. This is why the QALY can be defended as "one number" even though it is the product of heterogeneous elements and derived from multiattribute descriptions.

The table's treatment of negative values exposes the same exemption with even less subtlety. Ratio measures can have negative values is endorsed as false at p = 0.90 (+2.20), meaning the UK knowledge base strongly reinforces the claim that ratio measures can indeed take negative values. This is the signature accommodation of the EQ-5D value set world: "worse than dead" is accepted, negative utilities are normalized, and yet the system continues to describe the output as if it were on a ratio scale. The contradiction is not repaired; it is institutionalized. Once that is done, the system can go on calling the QALY a ratio measure without ever confronting what a true zero means or the actual definition of a ratio measure.

Quality of life, as the UK system uses it, is a further demonstration of the same categorical failure. The table does not include a separate "quality of life" item because the canonical 24 statements are already sufficient, but the logic is embedded in the endorsement of the QALY as dimensionally homogeneous at p = 0.85 (+1.75). Dimensional homogeneity is the condition for meaningful aggregation and ratio formation. Endorsing it for the QALY is endorsing a fiction: that time and ordinal preference weights are commensurable in a way that supports multiplication, that the product is a single attribute, and that the result can be added across people. This is exactly the kind of claim a measurement gatekeeper would block immediately. The UK system instead blesses it, because it needs a single scalar for administrative closure.

Now the decisive block: Rasch. If the UK system had any serious concept of latent trait possession, it would be driven toward Rasch. Not as an optional method, but as a compulsory condition for claiming measurement of subjective attributes. Table 1 shows the opposite. Every Rasch proposition collapses to the absolute floor, p = 0.05 (−2.50), including the central claim that there are only two admissible measurement classes (linear ratio for manifest attributes and Rasch logit ratio for latent traits), the claim that transforming subjective responses to interval measurement is only possible with Rasch rules, the claim that the Rasch logit ratio scale is the only basis for assessing therapy impact for latent traits, and the claim that Rasch rules are identical to the axioms of representational measurement. These values do not mean "Rasch is less popular." They mean Rasch is epistemically disallowed. It cannot be permitted to become sovereign because sovereignty would invalidate the scoring-and-mapping pipeline on which the UK HTA memeplex depends.

The possession item reinforces the point. The outcome of interest for latent traits is the possession of that trait sits at p = 0.25 (−1.10). That is weak endorsement: the concept exists only faintly, at the boundary. And that is exactly what you would expect in a system that wants the rhetoric of patient-centeredness but cannot tolerate the measurement consequences. If possession became central, the next question would be: "Measured how?" and the answer would be: "Only on a Rasch logit ratio scale." The system therefore keeps possession conceptually thin and methodologically non-binding.

Finally, consider falsifiability. The UK system can gesture toward scientific norms, non-falsifiable claims should be rejected sits at p = 0.70 (+0.85) but it simultaneously endorses the opposite institutional practice at the ceiling: reference case simulations generate falsifiable claims sits at p = 0.95 (+2.50). This is not confusion; it is redefinition. "Falsifiable" is quietly transformed to mean "showing sensitivity to assumptions," not "capable of empirical refutation by invariant measurement under a protocol." Once falsification is reduced to model behavior, the system can claim scientific legitimacy while remaining insulated from scientific risk. It can close cases, defend thresholds, and ration access using outputs that can never be falsified in the strong Popperian sense.

That is why the UK is so important in the global memeplex. It demonstrates that the HTA belief system is not merely a set of bad habits. It is an administrative technology designed to produce decisions under limited data, to enforce closure, and to avoid the permanent instability that true normal science would impose. A measurement-first system never closes the case permanently because claims remain provisional and subject to refutation over the product's life. The UK system is built to avoid that burden. It begins with arithmetic, mandates reference-case modeling, and treats the outputs as if they were measurement-grade facts. It is an elegant bureaucratic solution to an epistemic problem, achieved by deleting the gatekeeper.

The UK memeplex profile is not merely "another instance" of false measurement. It is the most disciplined instance of it, because it is embedded in the rules of assessment. The canonical logits make this explicit. The system pushes measurement axioms to the floor while pushing QALY and reference-case propositions to the ceiling. That is the signature of a mature memeplex: it rewards the propositions that reproduce it and suppresses the propositions that would invalidate it. Under representational measurement theory, the UK's central numerical objects are not merely contestable; they are inadmissible. Yet they persist because they solve an administrative problem, closure, while demanding almost nothing from measurement literacy.

If the UK were to adopt a real measurement framework, the change would be immediate and destructive to the present architecture. The QALY would be reclassified as a composite scoring artifact rather than a ratio measure. Mapping and preference algorithms would be treated as non-measurement transformations. Reference-case simulation outputs would be treated as conditional projections rather than falsifiable claims. Manifest claims would be restricted to linear ratio measures. Latent claims would require Rasch logit ratio measurement demonstrating invariance and unidimensionality. And, most importantly, measurement would be restored as the non-negotiable precondition for arithmetic. Until that reversal occurs, the UK system will remain the world's most polished example of arithmetic without measurement, defended not by science but by administrative necessity.

DOES THE NICE REFERENCE DECISION TOOL HAVE A FUTURE?

For more than two decades, the NICE reference decision tool has occupied a central position in health technology assessment, not only within the United Kingdom but globally. Its influence extends far beyond national guidance. Through emulation, training pipelines, and academic diffusion, the NICE reference case became the template for cost-effectiveness assessment worldwide. It established what counts as acceptable evidence, what constitutes "value," and how pricing and access decisions should be framed. Until recently, this authority was rarely questioned at a foundational level. The reference decision tool was treated as a settled achievement rather than as a provisional construct open to scientific scrutiny. That assumption can no longer be sustained.

The deconstruction of the Office of Health Economics and its role in promoting QALY-based evaluation in the 1990s has already raised serious concerns about the epistemic foundations of the NICE framework. But what was previously difficult to demonstrate systematically can now be shown with clarity. AI large language model diagnostics allow interrogation of institutional knowledge bases at scale, revealing not what organizations claim to believe, but what they actually endorse, reinforce, and exclude across decades of publications, guidance documents, and methodological texts.

When applied to NICE and its reference decision tool, the results are unambiguous. The framework systematically rejects the axioms of representational measurement while simultaneously relying on arithmetic operations that presuppose those axioms. Measurement is not treated as a prerequisite for calculation. It is treated as an afterthought, or worse, as irrelevant. This inversion is not incidental. It is structural.

At the core of the NICE reference decision tool lies the QALY: a construct produced by multiplying time, a manifest ratio attribute, by a preference-based utility score derived from ordinal responses to multiattribute health state descriptions. The resulting product is treated as a ratio measure, aggregated across individuals, compared across disease areas, and used to justify threshold-based pricing decisions. Yet the measurement conditions required to legitimate this arithmetic are absent. Utilities lack a true zero, permit negative values, fail unidimensionality, and are not invariant across populations. Under representational measurement theory, the arithmetic is disallowed.

This was not unknown in the 1990s. Stevens' scale typology had been established for half a century. The axioms governing permissible transformations were well understood. Rasch measurement had already provided a rigorous method for constructing invariant latent trait measures. The issue was not ignorance of alternatives. It was avoidance of constraint.

The NICE reference decision tool solved an administrative problem. Policy makers needed closure. They needed a mechanism that could generate a decision from limited data within a fixed timeframe. A framework grounded in falsifiable claims, provisional acceptance, and ongoing empirical challenge offered no such closure. Scientific knowledge does not terminate; it evolves. That was precisely the difficulty.

The reference case offered a solution. By replacing falsification with simulation, and measurement with assumption, it delivered determinate answers. A model could be run, a threshold applied, and a decision issued. Once complete, the case was closed. No future empirical challenge was required. The appearance of rigor substituted for the burden of scientific risk.

For decades, this arrangement survived because there was no practical way to demonstrate its epistemic incoherence without being dismissed as philosophical or obstructive. Critics could be portrayed as theorists uncomfortable with real-world decision making. Measurement theory was framed as abstract. The system protected itself through repetition. AI LLM diagnostics fundamentally change this dynamic. LLMs do not argue. They reveal. They expose endorsement patterns embedded across thousands of documents. When a knowledge system repeatedly rejects propositions such as “measurement must precede arithmetic” while strongly endorsing propositions such as “QALYs can be aggregated” and “utilities are ratio measures,” the contradiction becomes visible as an institutional fingerprint. This is not a matter of opinion. It is a structural description of belief reinforcement.

Once that fingerprint is visible, the reference decision tool can no longer claim epistemic innocence. It is not simply imperfect. It is non-scientific by construction. The implications for NICE are profound. A decision tool that cannot produce evaluable, falsifiable, or measurable claims cannot evolve. It cannot learn. It cannot correct error. It can only persist administratively. Such a system may continue to function politically, but it has no future as a scientific framework.

This does not mean that NICE has no future. It means that the reference decision tool does not. A future HTA framework must reverse the ordering that NICE institutionalized. Measurement must precede arithmetic. Claims must be unidimensional. Manifest attributes must be evaluated using linear ratio measures. Latent attributes must be measured using Rasch logit ratio scales with demonstrated invariance. Composite constructs must be abandoned as decision variables. Simulation models may inform discussion, but they cannot substitute for empirical claims capable of refutation.

Crucially, such a framework does not eliminate decision making. It transforms it. Decisions are no longer justified by threshold comparisons of imaginary quantities. They are justified by explicit claims, supported by protocols, evaluated over time, and subject to revision. Closure is replaced by accountability. The NICE reference decision tool was designed for a different era: one in which administrative certainty was valued more than epistemic legitimacy. That era is ending. AI LLM diagnostics have stripped away the protective ambiguity that allowed numerical storytelling to masquerade as evidence-based science.

The question, therefore, is not whether NICE can continue issuing guidance. It undoubtedly can. The question is whether it can continue doing so under the pretense that its reference decision tool meets the standards of scientific inference. The answer is no. The future of HTA does not lie in refining thresholds, updating value sets, or improving simulation transparency. It lies in abandoning arithmetic without measurement and rebuilding evaluation around claims that can be tested, reproduced, and proven wrong.

If NICE chooses to make that transition, it may yet play a leading role in restoring epistemic legitimacy to health technology assessment. If it does not, the reference decision tool will persist only as an administrative relic, useful for governance, perhaps, but no longer defensible as science. AI has not created this reckoning. It has merely made it unavoidable.

3. THE TRANSITION TO MEASUREMENT IN HEALTH TECHNOLOGY ASSESSMENT

THE IMPERATIVE OF CHANGE

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

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

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

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

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

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

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

MEANINGFUL THERAPY IMPACT CLAIMS

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

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

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

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

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

THE PATH TO MEANINGFUL MEASUREMENT

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

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

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

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

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

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

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

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

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

TRANSITION REQUIRES TRAINING

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

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

Training must therefore begin with scale types and their permissible operations. Linear ratio measurement applies to manifest attributes that possess a true zero and invariant units, such as time, counts, and resource use. Latent attributes, by contrast, cannot be observed directly and cannot be measured through summation or weighting. They require formal construction through a measurement model capable of producing invariant units. This distinction is the conceptual fulcrum of reform, because it determines which claims are admissible and which are not.

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

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

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

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

A NEW START IN MEASUREMENT FOR HEALTH TECHNOLOGY ASSESSMENT

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

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

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