MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

UNITED KINGDOM: THE YORK CENTRE FOR HEALTH ECONOMICS - A CANONICAL KNOWLEDGE BASE INTERROGATION AND THE ABSENCE OF MEASUREMENT

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

LOGIT WORKING PAPER No 780 APRIL 2026

www.maimonresearch.com

Tucson AZ

FOREWORD

HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

The Centre for Health Economics (CHE), University of York, founded in 1983, is one of the leading academic institutions dedicated to the application of economic analysis to healthcare and public policy. Its core function is to generate research that informs decisions about the allocation of limited healthcare resources, with a particular focus on improving efficiency, effectiveness, and equity within health systems.

A central activity of the CHE is the development and application of economic evaluation methods. These include cost-effectiveness analysis, cost-utility analysis, and cost–benefit approaches used to compare healthcare interventions in terms of their costs and outcomes. The CHE has played a key role in establishing the use of standardized outcome measures, particularly quality-adjusted life years (QALYs), as a common metric for comparing interventions across disease areas. These evaluations are typically supported by decision-analytic models that synthesize data from clinical trials, observational studies, and other sources to estimate long-term costs and outcomes.

The CHE also undertakes methodological research aimed at refining the tools used in economic evaluation. This includes work on handling uncertainty, incorporating heterogeneity across patient populations, and improving the design and interpretation of models. Techniques such as probabilistic sensitivity analysis and value of information analysis have been developed and advanced within this research environment to support decision-making under uncertainty.

Another important area of activity is the analysis of healthcare policy and system performance. The CHE examines how healthcare services are organized, financed, and delivered, and evaluates the impact of policy interventions on access, quality, and outcomes. This includes research on priority setting, resource allocation frameworks, and the development of decision rules for healthcare funding.

The Centre maintains strong links with policy-making bodies, most notably its long-standing association with the National Institute for Health and Care Excellence (NICE). Through this relationship, its research has contributed to the development of guidelines and methods used in national decision-making processes, particularly in the assessment of new technologies and interventions.

In addition to research, the CHE is involved in teaching and training, offering postgraduate programs and contributing to the education of health economists and policy analysts. Through its publications, collaborations, and advisory roles, the CHE has had a significant and enduring influence on the development and global dissemination of health technology assessment practices.

The purpose of this study is to present the first deconstruction of the CHE, knowledge base through the lens of representational measurement theory, a dimension that is typically neglected, if not entirely absent, in the health technology assessment (HTA) literature across its various knowledge base domains. The issue is not that nothing has been written about the CHE. Quite the opposite. There is an extensive and influential body of work that has developed over more than four decades. Given the global status attached to the CHE as a formative contributor to the development and diffusion of HTA methodology, it is important to return to this foundational knowledge base and examine how measurement is treated within it.

This is the critical point. Across more than 110 interrogations of HTA knowledge bases in some 30 countries, there is no evidence that the question of measurement, specifically, whether the numerical constructs employed satisfy the axioms of representational measurement, has ever been considered. The objective here is not to anticipate conclusions, but to pose a specific and bounded question: what role, if any, do the axioms of representational measurement play in the concepts, methods, and analytical structures that define the CHE knowledge base?

The findings from the interrogation are clear cut and, not surprisingly, in complete accord with all previous knowledge base deconstructions. The pattern of responses is consistent, stable, and internally coherent. Statements that are true under the axioms of representational measurement are weakly endorsed, with probabilities concentrated in the lower range and corresponding negative logits. In contrast, statements that are false but central to HTA practice are strongly endorsed, with probabilities in the upper range and positive logits.

This pattern is not partial or ambiguous. It is systematic. The same configuration is observed and reported in this database with 110 HTA knowledge bases across the globe. There is virtually no variation in probability–logit assignments. There are no outliers or countervailing positions that would suggest internal recognition of measurement constraints. The results indicate the absence of the axioms required for measurement. The interrogation of the CHE knowledge base does not identify isolated inconsistencies but a coherent axiom-free framework. The absence of engagement with the axioms of representational measurement is not incidental. It is the defining characteristic of the CHE knowledge base; it puts arithmetic before measurement yet it cannot support even arithmetic.

The modern recognition of the role of the axioms of representational measurement can be traced back 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 CHE KNOWLEDGE BASE: STRUCTURE AND CONTENT

The knowledge base associated with CHE can be traced over more than four decades. Over this period, the CHE has produced a substantial and continuous body of work, including methodological texts, peer-reviewed publications, working papers, commissioned reports, and contributions to national and international policy frameworks. This corpus is sufficiently extensive and internally consistent to be treated as a coherent domain, allowing a systematic reconstruction of its concepts, methods, and analytical commitments.

For the purposes of interrogation, the CHE knowledge base can be understood as comprising a number of interrelated layers, each of which contributes to the structure of HTA as it is currently practiced. .

The first layer is outcome specification, which defines how health effects are described and compared. This includes the development and use of preference-based measures of health-related quality of life, typically derived from multiattribute instruments. These instruments generate numerical representations of health states, which are used as the basis for comparative evaluation across interventions and disease areas. This layer establishes the foundational descriptors of outcome that are carried forward into subsequent analyses.

The second layer is aggregation and metric construction, where outcome measures are combined with time to produce summary metrics, most notably quality-adjusted life years (QALYs). This layer is central to the CHE framework, as it provides a single numerical index intended to capture both the quantity and quality of life. The resulting metric serves as the primary outcome in cost-utility analysis and enables comparison across heterogeneous interventions.

The third layer is economic evaluation, where costs are related to outcomes through cost-effectiveness or cost-utility frameworks. This includes the calculation of incremental cost-effectiveness ratios and the comparison of interventions in terms of cost per unit of outcome. This layer operationalizes the relationship between resource use and health outcomes and provides the basis for decision rules and thresholds.

The fourth layer is decision-analytic modeling, which integrates data from multiple sources to estimate outcomes over extended time horizons. Models are used to simulate disease progression, intervention effects, and patient pathways, often incorporating assumptions about behavior, treatment adherence, and system dynamics. Techniques such as probabilistic sensitivity analysis and value of information analysis are used to characterize uncertainty and support decision-making under conditions of incomplete evidence.

The fifth layer is policy translation and institutionalization, where analytical outputs are incorporated into decision frameworks. This includes contributions to reference case specifications, the development of decision rules, and engagement with policy bodies such as

NICE. At this level, the methods and metrics developed within the knowledge base are translated into operational criteria for resource allocation and priority setting.

These layers are not independent. They form a cumulative structure in which each layer depends on the assumptions and outputs of the preceding one. The CHE knowledge base is therefore not simply a collection of studies, but an integrated framework in which outcome specification, aggregation, evaluation, modeling, and policy application are aligned. The continuity of this structure over more than 40 years allows it to be interrogated as a stable and well-defined domain, with clearly identifiable components and relationships.

Practice does not make perfect, nor can it substitute for what this interrogation makes clear: the complete absence of the constraints imposed on quantitative analysis by the axioms of representational measurement. Although by 1983 the typology proposed by Stevens and the formalization of the axioms by Krantz et al in 1971 were well established, there is no evidence for, or defense of, setting representational measurement aside. Instead, the knowledge base proceeds in a context where these constraints are not engaged, as if the contributions of measurement theory were absent from consideration. There is no attempt to address, adapt, or reject the axioms of representational measurement; they are simply not part of the conceptual framework. The issue is not that measurement theory is debated and set aside, but that it does not appear as a reference point for arithmetic and the construction of quantitative claims. In effect, the framework develops without acknowledging the existence of a body of work that had already established the conditions under which numerical operations are meaningful. As a result, the constraints that would normally govern the transition from observation to measurement are not applied. The omission is therefore not argumentative but structural: quantitative constructs are introduced as given facts.

What is absent is not a particular method, but a set of necessary conditions. There is no requirement that attributes be demonstrated as unidimensional, no recognition that arithmetic operations demand ratio scales with a true zero, and no application of dimensional analysis to ensure homogeneity of constructed measures. Preference-based scores are treated as if they possess interval or ratio properties without demonstration, and composite metrics are formed and manipulated without reference to admissible transformations. The question of whether these constructs meet the axioms of measurement is not addressed; there is no thought of addressing it.

This absence extends to latent constructs. There is no incorporation of a measurement model that would justify transforming ordinal observations into measurable quantities. Rasch had been in place since 1960 with the Rasch rules shown to be identical to the axioms of representational measurement, with its focus on manifest attributes, since the mid-1970s. There is no recognition of measurement and latent attributes within the knowledge base framework. Instead, ordinal responses are directly aggregated or weighted and treated as if they were measures. The result is a knowledge base in which quantitative outputs are generated and applied without the prior establishment of measurement, and without the constraints that would give those outputs scientific meaning. The system was constructed without representational measurement from the outset,

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: CENTER FOR HEALTH ECONOMICS, UNIVERSITY OF YORK

Table 1 presents 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.

CENTRE FOR HEALTH ECONOMICS: THE ABSENCE OF REPRESENTATIONAL MEASUREMENT AND THE ENDORSEMENT OF FALSE MEASUREMENT

The interrogation of the knowledge base associated with the CHE is of particular importance because this is not merely another academic unit contributing to health technology assessment. It is one of the principal intellectual sources from which modern HTA false measurement practice has been shaped, codified, and exported. If the current HTA architecture has a formative academic center in the United Kingdom, CHE is an obvious candidate. This makes the present interrogation results especially significant. It is not simply an assessment of one knowledge base among many; it is an assessment of a knowledge base that has helped define the standards, expectations, and analytical conventions of the field itself; a global memeplex of false measurement and impossible arithmetic. A unique non-science achievement.

TABLE 1: ITEM STATEMENT, RESPONSE, ENDORSEMENT AND NORMALIZED LOGITS CENTRE FOR HEALTH ECONOMICS

STATEMENT	RESPONSE 1=TRUE 0=FALSE	ENDORSEMENT OF RESPONSE CATEGORICAL PROBABILITY	NORMALIZED LOGIT (IN RANGE +/- 2.50)
INTERVAL MEASURES LACK A TRUE ZERO	1	O,20	-1.40
MEASURES MUST BE UNIDIMENSIONAL	1	0.15	-1.60
MULTIPLICATION REQUIRES A RATIO MEASURE	1	0.10	-2.20
TIME TRADE-OFF PREFERENCES ARE UNIDIMENSIONAL	0	0.90	+2.20
RATIO MEASURES CAN HAVE NEGATIVE VALUES	0	0.95	+2.50
EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES	0	0.90	+2.20
THE QALY IS A RATIO MEASURE	0	0.95	+2.50
TIME IS A RATIO MEASURE	1	0.95	+2.50
MEASUREMENT PRECEDES ARITHMETIC	1	0.10	-2.20
SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES	0	0.90	+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	0.95	+2.50
CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.15	-1.75
QALYS CAN BE AGGREGATED	0	0.95	+2.50

NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.65	+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	+0,85
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	0.35	-1.25
THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT	1	0.25	-1.87
THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.05	-2.50

The results in Table 1 are clear and highly structured. As in all previous knowledge base interrogations, statements that are true under the axioms of representational measurement are weakly endorsed, while statements that are false but embedded within HTA practice are strongly endorsed. In the CHE case, however, this pattern is particularly forceful. Core HTA propositions are not merely accepted; they are maximally reinforced, often at the upper end of the probability range. At the same time, the axioms required to justify these propositions are weakly endorsed or absent. The result is a strongly polarized profile of measurement inversion.

Consider first the statements that are true. “MEASURES MUST BE UNIDIMENSIONAL” is assigned a probability of 0.15 with a normalized logit of -1.60. “MULTIPLICATION REQUIRES A RATIO MEASURE” is assigned 0.10 and -2.20. “MEASUREMENT PRECEDES ARITHMETIC” is also assigned 0.10 and -2.20. “MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC” is similarly weakly endorsed. These are not peripheral propositions. They are the basic conditions under which numerical operations can be interpreted as meaningful. Yet in the York knowledge base they are not recognized.

The same is true for the Rasch-based statements. “THERE ARE ONLY TWO CLASSES OF MEASUREMENT LINEAR RATIO AND RASCH LOGIT RATIO,” “TRANSFORMING SUBJECTIVE RESPONSES TO INTERVAL MEASUREMENT IS ONLY POSSIBLE WITH RASCH RULES,” “THE RASCH LOGIT RATIO SCALE IS THE ONLY BASIS FOR ASSESSING THERAPY IMPACT FOR LATENT TRAITS,” and “THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT” are all assigned a probability of 0.05 with a normalized logit of -2.50. This is the lowest possible category. It indicates complete or near-complete non-possession within the knowledge base. There is no meaningful engagement with Rasch measurement as a basis for the evaluation of latent traits. There is no recognition, even after some 20 years, that Rasch is the necessary and sufficient rules for transforming ordinal subjective responses to interval measures defined in logit terms.

This absence is not confined to one or two propositions. It defines the structure of the true statements as a group. Even “THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT” is assigned only 0.25 and -1.87. “CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT” is assigned 0.15 and -1.75. In other words, even when the diagnostic directly challenges the central numerical claims of HTA, the York knowledge base does not move toward recognition. It remains within the negative logit domain.

In direct contrast, the statements that are false under the axioms of representational measurement but central to HTA are strongly endorsed. “TIME TRADE-OFF PREFERENCES ARE UNIDIMENSIONAL” is assigned 0.90 and +2.20. “EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES” is assigned 0.90 and +2.20. “THE QALY IS A RATIO MEASURE” is assigned 0.95 and +2.50. “THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE” is assigned 0.95 and +2.50. “QALYS CAN BE AGGREGATED” is assigned 0.95 and +2.50. These are not moderate endorsements. They represent the highest possible confidence in propositions that are, under representational measurement, false.

The contradiction is not subtle. “MULTIPLICATION REQUIRES A RATIO MEASURE” is weakly endorsed, while “THE QALY IS A RATIO MEASURE” is maximally endorsed. “MEASUREMENT PRECEDES ARITHMETIC” is weakly endorsed, while arithmetic operations on utilities and QALYs are fully accepted. “MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC” is weakly endorsed, while “SUMMATION OF LIKERT QUESTION SCORES CREATES A RATIO MEASURE” and “SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES” are strongly endorsed. The CHE knowledge base therefore does not merely overlook measurement axioms; it endorses propositions that directly violate them, denying meaningful arithmetic operations.

This is what makes the CHE result so important. If this pattern had emerged only in peripheral or derivative knowledge bases, one might argue that the problem lay in transmission, simplification, or methodological drift. But the CHE has long been one of the principal sources from which the modern HTA framework has drawn legitimacy. The strong positive endorsement of the QALY, utilities, and cost-effectiveness constructs, combined with the strong negative endorsement of measurement axioms, suggests that the inversion is not a corruption of an originally sound framework. It is present as the foundation commitment to false measurement.

There are two statements that show a different pattern: “NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED” and “THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO,” both assigned 0.65 and +0.85. These are revealing. They indicate that the York knowledge base can recognize certain features of scientific and formal reasoning at an abstract level. Falsifiability is acknowledged in principle. The formal definition of the logit is acknowledged in principle. But this recognition does not extend to the actual numerical constructs that define HTA. That is, scientific language is present, but it is not applied where it would challenge the framework itself.

This selective recognition is a hallmark of a closed epistemic system. The system does not reject the language of science; it appropriates it. It accepts falsifiability in general, but not where it would threaten QALYs, utilities, or reference case simulations. It accepts formal mathematical definition, but not where it would require the replacement of ordinal preference algorithms with a valid measurement model. This is why the statement "REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS" is assigned 0.95 and +2.50. The reference case is treated as fully legitimate and fully scientific, despite the absence of measurable attributes.

The implications are considerable. If the CHE knowledge base is, as many would claim, one of the originating centers of modern HTA methodology, then the measurement debacle is not a later development. It is built into the architecture of the field. The constructs that have come to dominate HTA such as time trade-off preferences, EQ-5D-based utility algorithms, QALYs, and cost-effectiveness ratios were not later corrupted by misuse. They were adopted and normalized in the absence of the axioms required for measurement from the outset.

This conclusion also explains why the same pattern appears globally. The consistency observed internationally is not surprising if the originating knowledge bases already exhibit this inversion. What is reproduced in later research centers is a framework that was already closed to measurement. The persistence of the pattern is therefore not accidental. It is the expected consequence of a knowledge system that reproduces itself without falsification or awareness of required axioms for quantitative analysis.

This does not mean that the CHE knowledge base lacks internal coherence. On the contrary, it is internally coherent to those with no background in measurement theory. It has a stable set of constructs, accepted templates, and recognized outputs. It supports training, publication, and policy advice. But internal coherence is not the same as scientific validity. Scholastic systems can be internally coherent. What matters is whether the framework is constrained by measurement and open to falsification. The interrogation indicates that the CHE knowledge base is not.

THE PRIMACY OF MEASUREMENT

From its establishment in 1983, the CHE has played a central role in the construction of the modern HTA framework. The subsequent evolution of this framework can be traced through a series of identifiable phases; formation, codification, institutionalization, and global diffusion. At no point in this trajectory is there evidence that the axioms of representational measurement were incorporated. On the contrary, the framework that emerges is one in which ordinal preference data are systematically transformed into numerical constructs and treated as if they possessed interval and ratio properties. The result is not a later failure of application, but a measurement failure embedded in the architecture from the outset.

Judged against the standards of falsification and the axioms of representational measurement, HTA stands outside the requirements of normal science. It presents itself as a quantitative discipline, yet does so without meeting the conditions that make quantification meaningful. The result is a 40-year commitment to numerical storytelling: a body of work that is internally coherent, plausible, and operationally useful, but not grounded in measurement. The appearance of precision is maintained, but the foundation required to support it is absent.

Despite more than four decades of publications originating from the CHE there is no identifiable body of work that evaluates the HTA framework against the axioms of representational measurement. The literature documents the development, refinement, and global diffusion of cost-effectiveness analysis, utility-based constructs, and the QALY. It does not address their measurement legitimacy. This absence is not incidental. It reflects the construction of a knowledge base in which the conditions required for measurement were never treated as relevant constraints.

The consequences for practice, including legacy claims, are immediate. If the principal constructs of HTA are not measurable, then the policy decisions that rely on those constructs cannot claim a scientific quantitative foundation. This does not mean that decisions cannot be made. It means that the numerical claims used to support them are not measures in the representational sense. The appearance of precision is not the same as the existence of measurement.

At the same time, the path forward is straightforward. The alternative is not methodological disruption, but a return to established standards. For manifest attributes, this means simple linear ratio measures expressed in natural units. For latent constructs, it means Rasch logit ratio measurement. These are not innovations; they are standard scientific solutions. What is required is not invention, but recognition.

The interrogation of the CHE knowledge base makes one point with unusual clarity: the problem is not peripheral. It is foundational. The measurement inversion observed across global HTA knowledge bases is present in one of the field's principal intellectual centers. The issue is no longer whether measurement principles have occasionally been neglected. It is whether they were ever incorporated in the first place. On the evidence of this interrogation the answer is no.

The implications of this analysis are immediate and unavoidable. If the constructs that define HTA do not meet the axioms of representational measurement, then the framework cannot claim a quantitative scientific foundation. This is not a matter of refinement or methodological improvement, but of recognition. The issue is not how to adjust existing models or enhance their complexity, but whether the quantities they manipulate are measures in the first place. Without measurement, there is no basis for arithmetic, no basis for comparison, and no basis for falsifiable claims.

At the same time, the absence of measurement does not leave a void. The requirements for valid quantitative analysis are well established. What is at issue is not feasibility, but adoption. The question that follows is therefore straightforward: given the availability of measurement-based alternatives, on what grounds can the current framework be maintained?

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.

MEANINGFUL THERAPY IMPACT CLAIMS

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

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

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

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

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

THE PATH TO MEANINGFUL MEASUREMENT

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

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

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

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

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

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

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

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

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

TRANSITION REQUIRES TRAINING

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

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

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

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

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

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

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

A NEW START IN MEASUREMENT FOR HEALTH TECHNOLOGY ASSESSMENT

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

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

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

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

DESIGNED FOR CLOSURE

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

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

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

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

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

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

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

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

ACKNOWLEDGEMENT

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

REFERENCES

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

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

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

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