

**MAIMON RESEARCH LLC**

**ARTIFICIAL INTELLIGENCE LARGE LANGUAGE  
MODEL INTERROGATION**



**REPRESENTATIONAL MEASUREMENT FAILURE IN  
HEALTH TECHNOLOGY ASSESSMENT**

**SPAIN: THE REDETS NATIONAL NETWORK  
ENDORSES FALSE MEASUREMENT FOR HTA**

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

**LOGIT WORKING PAPER No 1282 APRIL 2026**

[www.maimonresearch.com](http://www.maimonresearch.com)

**Tucson AZ**

# FOREWORD

## HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

The RedETS (Red Española de Agencias de Evaluación de Tecnologías Sanitarias y Prestaciones del Sistema Nacional de Salud) is Spain's national network for coordinating health technology assessment (HTA) activities across the public healthcare system. Established to support evidence-based decision-making, RedETS brings together regional HTA agencies and units under a common framework to evaluate healthcare technologies, services, and interventions.

Its primary role is to produce systematic assessments of pharmaceuticals, medical devices, diagnostic tools, and healthcare programs to inform decisions on coverage, reimbursement, and clinical use within Spain's National Health System. These assessments typically include reviews of clinical effectiveness, safety, and economic value, often incorporating cost-effectiveness and cost-utility analyses. RedETS operates through a distributed model, where participating agencies contribute to jointly commissioned reports following standardized methodologies and coordinated work plans.

In addition to producing HTA reports, RedETS contributes to the development of methodological guidelines and promotes consistency in evaluation practices across regions. It collaborates with national health authorities, regional governments, and international HTA networks, including European initiatives, to align assessment processes and share evidence. Through these activities, RedETS functions as a central mechanism for integrating scientific, economic, and policy considerations into healthcare decision-making in Spain, ensuring that technology adoption and resource allocation are informed by structured evaluation frameworks.

The objective of this study is to undertake a structured interrogation of the HTA-related knowledge base associated with the RedETS using a standardized 24-item canonical diagnostic instrument grounded in the axioms of representational measurement theory. The purpose is not to assess individual reports or agency outputs, but to determine whether the conceptual framework that informs RedETS satisfies the minimum requirements for scientific measurement. In particular, the study evaluates whether the knowledge base recognizes the conditions necessary for admissible arithmetic operations, including unidimensionality, invariance, the existence of a true zero, and the requirement that latent constructs be measured through Rasch-conformant models. By assigning categorical probabilities to each statement and transforming these into normalized logits, the analysis provides a structured profile of conceptual endorsement, allowing an explicit assessment of whether the knowledge base supports evaluable, replicable, and falsifiable value claims.

The findings demonstrate a clear and consistent pattern of measurement inversion. Statements that are true within representational measurement theory are weakly endorsed or rejected, while statements that are false—particularly those underpinning cost-utility analysis—are strongly endorsed, frequently at maximum logit levels. Constructs such as utilities and QALYs are treated as if they possess interval or ratio scale properties, supporting multiplication, aggregation, and

modeled projection, despite lacking the required measurement foundations. At the same time, key principles such as the requirement that multiplication demands a ratio scale and that measurement must precede arithmetic are not recognized. Rasch measurement is entirely absent, indicating that the only scientifically defensible pathway for constructing measures of latent constructs is not present. The resulting profile confirms that the knowledge base does not meet the axioms of representational measurement and therefore supports the generation of non-evaluable claims.

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

The modern articulation of this principle can be traced to Stevens' seminal 1946 paper, which introduced the typology of nominal, ordinal, interval, and ratio scales <sup>1</sup>. 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) <sup>2</sup>. 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<sup>3</sup>. 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<sup>4</sup>.

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](mailto: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 not 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 RedETS NETWORK HTA KNOWLEDGE BASE**

The HTA-related knowledge base associated with RedETS is best understood as a coordinated, network-based framework that integrates multiple regional HTA agencies into a unified national system for evaluating healthcare technologies. Its primary function is to support decision-making within Spain’s National Health System by producing structured assessments of pharmaceuticals, medical devices, diagnostic tools, and healthcare programs. These assessments typically combine evidence on clinical effectiveness and safety with economic evaluation, most commonly expressed through cost-effectiveness and cost-utility analysis. Within this framework, preference-based measures of health outcomes—particularly quality-adjusted life years—serve as the central metric for comparing interventions and informing resource allocation decisions.

A defining characteristic of this knowledge base is the absence of any formal requirement to establish the measurement properties of the constructs it employs. Preference-based measures derived from instruments such as EQ-5D are treated as if they possess interval or ratio scale characteristics, enabling their use in arithmetic operations such as multiplication by time and aggregation across individuals. However, there is no requirement to demonstrate unidimensionality, invariance, or the existence of a true zero, nor is there systematic use of Rasch modeling to transform ordinal observations into valid measures of latent constructs. Instead, the framework proceeds on the implicit assumption that numerical representation is sufficient to constitute measurement. This assumption allows the construction of QALYs and the application of economic models without addressing whether the underlying quantities meet the axioms of representational measurement.

The networked structure of RedETS reinforces this approach. By coordinating multiple regional agencies under a shared methodological framework, RedETS standardizes evaluation practices and disseminates common assumptions across the Spanish healthcare system. This creates a high degree of internal consistency, as participating agencies apply similar methods and produce comparable outputs. However, this consistency is achieved without reference to measurement validity. The shared framework ensures that numerical outputs are comparable within the system, but it does not establish that these outputs represent measurable quantities. As a result, the knowledge base functions as a system of coordinated numerical representation rather than a framework grounded in measurement.

The methodological sophistication of RedETS further contributes to the resilience of this structure. Advanced statistical techniques, decision-analytic models, and simulation approaches are employed to evaluate interventions, project outcomes, and explore uncertainty. These methods provide a high degree of analytical rigor in terms of internal consistency and technical execution. However, they operate within a framework that assumes the validity of the measures used rather than establishing it. The focus is on improving estimates, refining models, and enhancing policy

relevance, rather than on the prior question of whether the quantities being analyzed are measurable. This results in a situation where methodological complexity coexists with a lack of measurement validity.

The integration of RedETS into national and regional decision-making processes further reinforces the knowledge base. HTA reports produced within the network are used to inform coverage, reimbursement, and clinical practice decisions, embedding the framework within the operational structure of the healthcare system. This creates a feedback loop in which methodological assumptions are validated through their application in policy, rather than through adherence to measurement principles. The knowledge base is therefore both influential and self-reinforcing, with limited capacity for internal critique or methodological correction.

In this context, the absence of representational measurement is not an incidental oversight but a structural feature of the framework. The reliance on QALYs and related constructs as central metrics for evaluation reflects a commitment to a form of numerical reasoning that does not meet the conditions required for measurement. The resulting value claims cannot be considered evaluable in a scientific sense, as they are not grounded in quantities that support lawful arithmetic. Despite their widespread use and institutional acceptance, these claims remain products of a framework that substitutes numerical representation for measurement, limiting their capacity to support replication, falsification, and the accumulation of objective knowledge within the standards of normal science.

## **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 a 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  $\pm 2.50$  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  $\pm 4.0$  logits to avoid extreme distortions, and normalized to  $\pm 2.50$  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: THE RedETS NETWORK HTA KNOWLEDGE BASE**

Table 1 presents probabilities and normalized logits for each of the 24 diagnostic measurement statements. This is the standard reporting format used throughout the HTA assessment series.

It is essential to understand how to interpret these results.

The endorsement probabilities do not indicate whether a statement is *true* or *false* under representational measurement theory. Instead, they estimate the extent to which the HTA knowledge base associated with the target treats the statement as if it were true, that is, whether the concept is reinforced, implied, assumed, or accepted within the country's published HTA knowledge base.

The logits provide a continuous, symmetric scale, ranging from +2.50 to -2.50, that quantifies the degree of this endorsement. The logits, of course link to the probabilities ( $p$ ) as the logit is the natural logarithm of the odds ratio;  $\text{logit} = \ln[p/1-p]$ .

- 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.

## **THE MEASUREMENT FAILURE OF THE RedETS NETWORK HTA KNOWLEDGE BASE**

The interrogation of the HTA-related knowledge base associated with RedETS provides a definitive assessment of measurement standards within a coordinated national HTA system. Unlike single academic centers, RedETS represents a structured network integrating multiple regional HTA agencies operating under a shared methodological framework. The knowledge base considered in this assessment is therefore not inferred from general HTA practice, but defined explicitly in terms of RedETS outputs, including its reports, methodological guidance, and participation in European HTA initiatives. (Table 1)

**TABLE 1: ITEM STATEMENT, RESPONSE, ENDORSEMENT AND NORMALIZED LOGITS THE RedETS NETWORK KNOWLEDGE BASE**

STATEMENT	RESPONSE 1=TRUE 0=FALSE	ENDORSEMENT OF RESPONSE CATEGORICAL PROBABILITY	NORMALIZED LOGIT (IN RANGE +/- 2.50)
INTERVAL MEASURES LACK A TRUE ZERO	1	0.25	-1.10
MEASURES MUST BE UNIDIMENSIONAL	1	0.20	-1.40
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	0.95	+2.50
THE QALY IS A RATIO MEASURE	0	0.95	+2,50
TIME IS A RATIO MEASURE	1	0.85	+1.75
MEASUREMENT PRECEDES ARITHMETIC	1	0.20	-1.40
SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES	0	0.95	+2.50
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	0.95	+2.50
THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE	0	0.95	+2.50
CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.10	-2.20
QALYS CAN BE AGGREGATED	0	0.95	+2.50

NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.50	0.00
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.95	+2.50
THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.80	+1.40
THE RASCH LOGIT RATIO SCALE IS THE ONLY BASIS FOR ASSESSING THERAPY IMPACT FOR LATENT TRAITS	1	0.05	-2.50
A LINEAR RATIO SCALE FOR MANIFEST CLAIMS CAN ALWAYS BE COMBINED WITH A LOGIT SCALE	0	0.90	+2.20
THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT	1	0.05	-2,50
THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.05	-2.50

The interrogation of the HTA-related knowledge base associated with RedETS provides a definitive assessment of measurement standards within a coordinated national HTA system. Unlike single academic centers, RedETS represents a structured network integrating multiple regional HTA agencies operating under a shared methodological framework. The knowledge base considered in this assessment is therefore not inferred from general HTA practice, but defined explicitly in terms of RedETS outputs, including its reports, methodological guidance, and participation in European HTA initiatives.

The results demonstrate a clear and unambiguous pattern of measurement inversion. Statements that are true within representational measurement theory are weakly endorsed or rejected, while statements that are false are strongly endorsed, frequently at the highest logit level. This pattern is not incidental. It reflects the methodological commitments embedded in the RedETS corpus, where economic evaluation is structured around cost-effectiveness and cost-utility analysis, with QALYs as the central outcome metric.

The strongest endorsements are associated with the propositions that sustain cost-utility analysis. The claim that the QALY is a ratio measure is effectively treated as true, with a probability of 0.95 and a normalized logit of +2.50. This is not stated explicitly in RedETS documentation, but it is operationally embedded in the use of QALYs in arithmetic operations. QALYs are multiplied, aggregated, and compared across populations, which requires that they be treated as ratio-scale quantities. Similarly, the propositions that QALYs are dimensionally homogeneous and aggregable are endorsed at the same level. These are not theoretical positions but implicit assumptions required for the framework to function.

This is where the fundamental problem arises. A ratio scale requires a true zero and invariant unit structure. Preference-based utility measures derived from instruments such as EQ-5D do not meet these conditions. They are constructed from ordinal responses and transformed through valuation algorithms that do not establish interval or ratio properties. The multiplication of such values by time to generate QALYs is therefore not an admissible operation within the axioms of representational measurement. It is not a matter of approximation or modeling convenience. It is a categorical violation of the conditions required for arithmetic.

The endorsement of the proposition that EQ-5D algorithms create interval measures, again at the maximum logit level, reinforces this conclusion. Within the RedETS knowledge base, these measures are treated as if they possess interval properties, enabling arithmetic operations. However, there is no requirement within the framework to demonstrate unidimensionality, invariance, or the existence of a true zero. The transformation from ordinal responses to numerical values is assumed to produce measurement, without establishing the properties required for measurement.

This pattern extends to the treatment of subjective instruments more generally. The use of summed responses from patient-reported outcome measures implies endorsement of the false claim that such summations produce ratio measures. These values are used as inputs to economic evaluation without validation of their scale properties. This is the defining feature of measurement inversion: numerical outputs are treated as measures, and arithmetic operations are applied without establishing the conditions required for those operations.

In contrast, the propositions that define the axioms of representational measurement are systematically rejected. The requirement that multiplication demands a ratio scale is assigned a probability of 0.10, indicating effective non-possession. The principle that measurement must precede arithmetic is weakly endorsed, while the necessity that arithmetic operations be grounded in measurement axioms is rejected. These results demonstrate that the RedETS knowledge base does not recognize the foundational requirements of measurement. Arithmetic is applied directly to numerical representations, bypassing the question of whether those representations are measures.

The absence of Rasch measurement is complete. There is no evidence within the RedETS corpus of the use of Rasch modeling to construct invariant measures of latent constructs. This absence is decisive. Without Rasch transformation, there is no scientifically defensible pathway from ordinal observations to measurement. Preference-based valuation does not substitute for this process. It produces a numerical ordering, not a measure. The subsequent use of these values in arithmetic operations is therefore unjustified.

The networked structure of RedETS amplifies the significance of these findings. The knowledge base is not confined to a single institution but shared across multiple agencies operating within a coordinated framework. This means that the measurement assumptions identified here are not localized. They are reproduced across the network and embedded in the outputs of multiple agencies. The consistency of the logit profile reflects this shared framework. It is not the result of a single methodological choice but of a coordinated system that standardizes HTA practice across Spain.

This network structure also reinforces the resilience of the framework. RedETS operates as a mechanism for the dissemination and normalization of HTA methods, aligning regional agencies with national and European standards. This creates a high degree of internal consistency, as agencies apply similar methods and produce comparable outputs. However, this consistency is achieved without reference to measurement validity. The framework ensures comparability of numerical outputs but does not establish that those outputs represent measurable quantities.

The integration of RedETS into European HTA initiatives further extends this structure. Participation in EUnetHTA and alignment with EU HTA regulation embed the same methodological assumptions within a broader transnational framework. This indicates that the problem is not confined to Spain but reflects a shared European approach to HTA. The RedETS knowledge base therefore serves as a representative case of network-based HTA systems operating within this broader context.

The treatment of falsifiability illustrates an additional dimension of the problem. Simulation models are treated as sources of evidence, despite their lack of falsifiability in the Popperian sense. Model outputs depend on assumptions and parameter choices that can be adjusted without exposing the model to decisive empirical test. This does not prevent their use within the knowledge base as a basis for decision-making. The distinction between internal consistency and empirical testability is not recognized, allowing modeled projections to function as evidence.

The result is a closed and internally coherent system. The knowledge base supports a set of methods that produce precise numerical outputs, supported by advanced statistical and modeling techniques. These outputs are treated as measures and used to inform policy decisions. However, the validity of the underlying quantities as measures is not established. The system operates on the assumption that numerical representation is sufficient, bypassing the requirements of representational measurement.

This is not a failure of technical competence. The RedETS network demonstrates a high level of methodological sophistication in economic evaluation and HTA practice. The problem lies in the absence of a requirement to validate measurement. The focus is on the application and refinement of methods, rather than on the prior question of whether the constructs being analyzed are measurable. This allows the framework to function effectively within its own terms, but it undermines its scientific legitimacy.

The implications are direct. If the quantities used in HTA do not meet the axioms of representational measurement, then the claims derived from them cannot be considered evidence. They are not empirically evaluable in a meaningful sense. They cannot be replicated as measures, and they cannot be falsified within the standards of normal science. They are products of a framework that substitutes numerical representation for measurement.

The RedETS case therefore provides a definitive demonstration of measurement inversion within a coordinated HTA network. It shows that the problem is not confined to individual institutions but embedded in the structure of the system. The consistency of the logit profile across statements reflects the operation of a shared knowledge base that reproduces the same methodological assumptions across agencies.

In conclusion, the interrogation of the RedETS knowledge base demonstrates that the Spanish HTA framework does not satisfy the axioms of representational measurement. The strong endorsement of false propositions and the rejection of true axioms indicate that the framework operates independently of the principles that govern measurement. This is not a sustainable position. Without adherence to the axioms of representational measurement, HTA cannot generate evaluable, falsifiable claims. It remains an exercise in numerical representation without measurement, regardless of its institutional structure or methodological sophistication.

## **THE FUTURE OF HEALTH TECHNOLOGY ASSESSMENT IN SPAIN**

The interrogations reported for Spain point to a clear and consistent conclusion: the HTA-related knowledge base is fully committed to measurement inversion. Across both the RedETS and AQuAS knowledge bases, the same structure is observed. Statements that are false within the axioms of representational measurement are strongly endorsed, while those that define the conditions for valid measurement are rejected or absent. This is not a marginal inconsistency or a localized weakness. It is a systematic and internally coherent framework in which numerical representation substitutes for measurement, and where arithmetic is applied to quantities that do not exist.

In both cases, the commitment to cost-utility analysis and QALY-based reasoning is explicit. Utilities are treated as if they possess interval or ratio scale properties, allowing multiplication by time, aggregation across individuals, and incorporation into modeled projections. These operations are foundational to HTA, yet they are not admissible. The requirement that multiplication demands a ratio scale is not recognized, nor is the prior condition that measurement must precede arithmetic. There is no recognition of Rasch-conformant measurement for latent constructs, no requirement for unidimensionality or invariance, and no attempt to establish lawful scale properties. False measurement is not an exception. It is the rule.

What distinguishes Spain is not the presence of an alternative framework, but the absence of any corrective mechanism. Academic centers do not challenge the evaluative structure. They reproduce it. There is no independent knowledge base that recognizes or applies the axioms of representational measurement. The absence is complete.

This raises an unavoidable question regarding the future of HTA in Spain. Can a framework that is demonstrably inconsistent with the principles of measurement be sustained once this inconsistency is made explicit? For decades, authority has rested on the presence of numbers including comparative effectiveness estimates, cost-effectiveness ratios, and model outputs combined with apparent methodological sophistication. The cat is now out of the bag. What are presented as measures are not measures at all. They do not meet the conditions required for measurement. Once this is recognized, the basis for that authority collapses.

The challenge is not technical but foundational. It is not a matter of refining models, improving data sources, or adjusting parameters. It is a question of whether the framework itself is admissible. A system that applies arithmetic to non-measurement cannot generate evaluable claims. It cannot support replication in a measurement-consistent sense, nor can it be falsified. The framework

collapses at the first step, because the quantities to which arithmetic is applied are not measures. Everything that follows is simply the institutionalized reproduction of invalid reasoning.

It produces outputs that are internally coherent but externally invalid. The logit profile makes this explicit: as a decision framework, HTA has nothing to offer. Its application should be discontinued.

There is no ambiguity in the evidence. Across more than 120 interrogations, the same pattern is observed: no recognition of measurement theory, and consistent endorsement of a framework that violates it. This is not ignorance alone. It is the institutionalization of a false belief system that has persisted for over 40 years.

Spanish research centers and agencies may choose to continue with the present framework. But that position is no longer defensible. The absence of measurement is now explicit. Continued reliance on cost-effectiveness claims that violate the axioms of measurement invites scrutiny and rejection.

There is only one path forward. If HTA is to survive, it must return to first principles. All claims must be grounded in valid measurement: linear ratio measures for manifest attributes and Rasch logit ratio measures for latent attributes. There are no alternatives. The fact that these requirements have been understood for more than half a century makes the persistence of the current framework all the more untenable.

### **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 three distance education programs to support the transition to a new paradigm in HTA. These comprise 12 module senior level program that details the standards for measurement, the failure of current HTA standards and the basis for protocol supported claims assessment for ratio measures of manifest attributes and Rasch logic ratio logit measures for latent attributes. The two other programs are only 5 modules but are designed to complement the 12-module program, for measurement axioms and Rasch attribute possession.

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.

## **MAIMON RESEARCH LLC**

### **DISTANCE EDUCATION PROGRAMS IN THE THEORY OF MEASUREMENT**

Three programs are available: two short 5-module programs and a 12-module program that is structured as a senior level course on the transition from the current HTA belief system to a new paradigm for HTA

The two short programs are (i) **NUMERICAL STORYTELLING: SYSTEMATIC MEASUREMENT FAILURE IN HEALTH TECHNOLOGY ASSESSMENT** and (ii) **A NEW START IN MEASUREMENT FOR HEALTH TECHNOLOGY ASSESSMENT**.

They are designed to complement the 12-module course program. They can be accessed through the **DISTANCE EDUCATION** section of the website with URL

<https://maimonresearch.com/distance-education-programs/>

The senior level course **HEALTH TECHNOLOGY ASSESSMENT REBUILT: EVIDENCE AND VALUE** is accessed through the **EVIDENCE AND VALUE** section of the website or

URL link <https://maimonresearch.com/evidence-and-value/>.

## **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**

---

<sup>1</sup> Stevens S. On the Theory of Scales of Measurement. *Science*. 1946;103(2684):677-80

<sup>2</sup> Krantz D, Luce R, Suppes P, Tversky A. Foundations of Measurement Vol 1: Additive and Polynomial Representations. New York: Academic Press, 1971

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

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

---