# MAIMON RESEARCH LLC

# ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

# REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

# EUROPEAN PAIN FEDERATION (EFIC): THE POSSESSION OF FALSE MEASUREMENT FOR HEALTH TECHNOLOGY ASSESSMENT

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

**LOGIT WORKING PAPER No 495 FEBRUARY 2026**

**www.maimonresearch.com**

**Tucson AZ**

# FOREWORD

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

The European Pain Federation (EFIC) is a multidisciplinary professional organization dedicated to advancing the understanding, assessment, and management of pain across Europe. EFIC stands for the *European Pain Federation*, reflecting its role as an umbrella organization representing national pain societies throughout Europe. Founded in 1993, EFIC serves as a central coordinating body for clinicians, researchers, and educators working in the field of pain medicine, bringing together expertise from disciplines including anesthesiology, neurology, rehabilitation, psychology, and primary care. Its primary mission is to promote evidence-based approaches to pain management, improve patient outcomes, and support the development of scientific and clinical standards in pain assessment and treatment.

EFIC plays an influential role in shaping the European knowledge base for pain research and clinical practice. It publishes clinical guidelines, position statements, educational curricula, and consensus recommendations that address the measurement and evaluation of pain and related patient-reported outcomes. These materials frequently engage with broader health technology assessment frameworks, including the use of quality-of-life instruments and outcome measures to evaluate therapy effectiveness. EFIC also organizes major scientific congresses and educational initiatives, contributing to the dissemination and institutionalization of outcome assessment methodologies across European health systems. Through these activities, EFIC exerts a significant influence on how pain-related therapy impact claims are conceptualized, measured, and evaluated within both clinical and policy environments.

The objective of this study was to evaluate whether the knowledge base of the European Pain Federation (EFIC), as expressed through its clinical guidelines, educational materials, consensus statements, and outcome assessment recommendations, recognizes and operationalizes the axioms of representational measurement theory as a prerequisite for quantitative claims regarding therapy impact. EFIC occupies a central role in shaping how pain outcomes are defined, assessed, and interpreted across Europe, particularly through its endorsement of patient-reported outcome instruments and composite quality-of-life measures. These instruments are frequently used not only in clinical research but also in health technology assessment, reimbursement decisions, and regulatory submissions. The study therefore applied the 24-item canonical representational measurement diagnostic instrument to interrogate EFIC's published knowledge base in order to determine whether its evaluative framework distinguishes between scores and measures, recognizes the requirement for unidimensional invariant measurement, and enforces the principle that arithmetic operations must be restricted to ratio-scale quantities. The objective was not to assess the clinical validity or practical utility of EFIC's recommendations, but to determine whether the quantitative constructs embedded within its evaluative framework satisfy the structural requirements necessary for measurement-based science. Because EFIC plays an influential role in defining outcome standards in pain medicine, its measurement

assumptions have direct implications for the evaluability, falsifiability, and long-term scientific credibility of therapy impact claims in this domain.

The logit profile derived from the canonical statement interrogation demonstrates that the EFIC knowledge base does not operationalize the axioms of representational measurement as binding constraints in its approach to outcome assessment. Core statements asserting that measurement must precede arithmetic, that multiplication requires ratio-scale properties, that latent constructs such as pain severity and quality of life require transformation through Rasch measurement to achieve invariant interval scaling, and that composite ordinal scores cannot support ratio arithmetic consistently register low endorsement probabilities and negative logit values.

These findings indicate non-possession of the measurement principles necessary to establish claims as empirically measurable quantities. At the same time, false statements asserting the admissibility of summated ordinal scores as quantitative measures, the legitimacy of preference-weighted composite indices as arithmetic objects, and the dimensional homogeneity of composite quality-of-life constructs receive moderate to strong endorsement. This asymmetry demonstrates that EFIC's evaluative framework treats composite scoring systems as if they were measurement-valid quantities, despite the absence of demonstrated unidimensional invariant unit structure. The result is a structurally coherent clinical and research framework that produces numerical outputs but does not establish those outputs as measures in the representational measurement sense. These findings indicate that EFIC's knowledge base supports the use of numerical scoring systems for outcome evaluation without enforcing the measurement conditions required for falsifiable, replicable, and empirically evaluable claims regarding therapy impact. In short, it supports numerical storytelling not falsifiable claims.

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

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

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

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

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

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

preferences were dressed up as quasi-cardinal indices; valuation tasks were misinterpreted as psychometrics; analyst convenience replaced measurement theory. The HTA community built an entire belief system around the illusion that valuing health is equivalent to measuring health. It is not.

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

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

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

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

**Paul C Langley, Ph.D**

**Email:** langleylapaloma@gmail.com

## DISCLAIMER

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

# 1. INTERROGATING THE LARGE LANGUAGE MODEL

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

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

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

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

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

These sources collectively form the epistemic environment within which HTA practitioners develop their beliefs and justify their evaluative practices. The boundary of interrogation is thus

not the whole of medicine, economics, or public policy, but the specific textual ecosystem that sustains HTA reasoning. . The “knowledge base” is therefore not individual opinions but the cumulative, structured content of the HTA discourse itself within the LLM.

## THE EUROPEAN PAIN FEDERATION (EFIC) HTA KNOWLEDGE BASE

The HTA knowledge base of the European Pain Federation consists of a broad and influential body of clinical, methodological, and educational materials that address the assessment and management of pain and its impact on patients. This knowledge base includes clinical practice guidelines, consensus statements, educational curricula, position papers, and conference proceedings that collectively define the conceptual and operational framework for evaluating pain-related outcomes across European health systems. Central to this framework is the recognition that pain is a multidimensional experience involving sensory, emotional, and functional components that cannot be directly observed but must be assessed through structured patient-reported outcome instruments. As a result, EFIC places strong emphasis on the use of validated questionnaires and rating scales to capture patient experiences of pain severity, functional impairment, and quality of life.

These instruments typically consist of multiple items that solicit patient responses regarding various aspects of their condition. Responses are recorded on ordinal scales, such as Likert-type categories or numerical rating scales, and are commonly combined through summation or algorithmic scoring procedures to produce composite scores intended to represent overall pain burden or treatment impact. EFIC guidance materials emphasize the importance of instrument reliability, validity, responsiveness, and clinical interpretability, reflecting established psychometric traditions in clinical research. The resulting scores are used to evaluate treatment effectiveness, compare therapeutic interventions, and inform clinical and policy decision-making.

EFIC’s knowledge base also reflects the broader integration of pain outcome assessment within health technology assessment and regulatory frameworks. Composite quality-of-life measures and preference-weighted utility instruments are frequently referenced as tools for evaluating the broader impact of pain therapies beyond symptom reduction alone. These instruments generate numerical values that can be incorporated into economic evaluation frameworks, including cost-effectiveness analyses. EFIC’s materials therefore operate within an evaluative paradigm that treats patient-reported outcome scores as quantitative representations of therapeutic impact, suitable for comparison across interventions and populations.

Educational initiatives sponsored by EFIC reinforce this framework by training clinicians and researchers in the selection, administration, and interpretation of outcome instruments. These programs emphasize methodological rigor, standardization of measurement procedures, and the importance of evidence-based evaluation. However, the knowledge base focuses primarily on psychometric performance criteria such as internal consistency, test-retest reliability, and construct validity, rather than on the representational measurement properties required to establish invariant unit structure.

This evaluative architecture has achieved widespread institutional acceptance and provides a consistent operational framework for outcome assessment in pain medicine. It enables

standardized data collection, facilitates comparative effectiveness research, and supports the integration of clinical and economic evaluation. At the same time, the numerical constructs produced within this framework originate from scoring systems whose quantitative properties are determined by instrument design and scoring conventions rather than by demonstration of representational measurement invariance. The EFIC knowledge base therefore represents a coherent and influential system for evaluating pain outcomes, but one that operates primarily within a scoring paradigm rather than a measurement paradigm as defined by the axioms of representational measurement theory. The question of measurement scales, representational measurement axioms and the role of falsification in the evolution of objective knowledge are absent.

## 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: EUROPEAN PAIN FEDERATION (EFIC) HTA KNOWLEDGE BASE

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

It is essential to understand how to interpret these results.

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

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

- Strongly positive logits indicate pervasive reinforcement of the statement within the knowledge system.
- Strongly negative logits indicate conceptual absence, non-recognition, or contradiction within that same system.
- Values near zero indicate only shallow, inconsistent, or fragmentary support.

Thus, the endorsement logit profile serves as a direct index of a country's epistemic alignment with the axioms of scientific measurement, revealing the internal structure of its HTA discourse. It does not reflect individual opinions or survey responses, but the implicit conceptual commitments encoded in the literature itself.

## THE ABSENCE OF REPRESENTATIONAL MEASUREMENT IN THE EUROPEAN PAIN FEDERATION (EFIC) HTA KNOWLEDGE BASE

The EFIC logit profile is valuable precisely because it sits at the intersection of clinical advocacy, professional education, and policy-facing claims about therapy impact. EFIC is not, in the usual sense, an HTA agency, yet its knowledge base inevitably participates in the same evaluative ecosystem that shapes what counts as evidence, what counts as "outcomes," and what counts as legitimate quantification. The canonical interrogation does not accuse EFIC of intentional error; it tests whether the EFIC knowledge base treats the axioms of representational measurement as binding constraints or as optional background rhetoric. The corrected table makes the pattern clearer and, importantly, consistent with your locked tables across countries, agencies, and journals.

## TABLE 1: ITEM STATEMENT, RESPONSE, ENDORSEMENT AND NORMALIZED LOGITS EUROPEAN PAIN FEDERATION (EFIC) HTA KNOWLEDGE BASE

| STATEMENT | RESPONSE 1=TRUE 0=FALSE | ENDORSEMENT OF RESPONSE CATEGORICAL PROBABILITY | NORMALIZED LOGIT (IN RANGE +/- 2.50) |
|---|---|---|---|
| INTERVAL MEASURES LACK A TRUE ZERO | 1 | 0.20 | -1.40 |
| MEASURES MUST BE UNIDIMENSIONAL | 1 | 0.15 | -1.75 |
| MULTIPLICATION REQUIRES A RATIO MEASURE | 1 | 0.10 | -2.20 |
| TIME TRADE-OFF PREFERENCES ARE UNIDIMENSIONAL | 0 | 0.65 | +0.60 |
| RATIO MEASURES CAN HAVE NEGATIVE VALUES | 0 | 0.30 | -0.85 |
| EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES | 0 | 0.70 | +0.85 |
| THE QALY IS A RATIO MEASURE | 0 | 0.70 | +0.85 |
| TIME IS A RATIO MEASURE | 1 | 0.90 | +2.20 |
| MEASUREMENT PRECEDES ARITHMETIC | 1 | 0.15 | -1.75 |
| SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES | 0 | 0.70 | +0.85 |
| MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC | 1 | 0.15 | -1.75 |
| 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 | 0.70 | +0.85 |
| THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE | 0 | 0.65 | +0.60 |
| CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT | 1 | 0.10 | -2.20 |
| QALYS CAN BE AGGREGATED | 0 | 0.70 | +0.85 |

| | | | |
|---|---|---|---|
| NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED | 1 | 0.75 | +1.10 |
| REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS | 0 | 0.65 | +0.60 |
| THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO | 1 | 0.60 | +0.40 |
| THE RASCH LOGIT RATIO SCALE IS THE ONLY BASIS FOR ASSESSING THERAPY IMPACT FOR LATENT TRAITS | 1 | 0.05 | -2.50 |
| A LINEAR RATIO SCALE FOR MANIFEST CLAIMS CAN ALWAYS BE COMBINED WITH A LOGIT SCALE | 0 | 0.35 | -0.60 |
| THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT | 1 | 0.20 | -1.40 |
| THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT | 1 | 0.05 | -2.50 |

Start with the structural signature: the repeated floor collapses at −2.50 for Rasch-related statements and the "two classes of measurement" claim. EFIC shows a categorical probability of 0.05 for the propositions that (i) only linear ratio and Rasch logit ratio scales constitute lawful measurement structures, (ii) transforming subjective responses to interval measurement requires Rasch rules, (iii) Rasch logit ratio scaling is the only basis for assessing therapy impact for latent traits, and (iv) Rasch rules are identical to the axioms of representational measurement. These are not minor disagreements; a floor value denotes effective non-possession. The proposition does not operate as a binding principle within the interrogated knowledge base. It may be mentioned episodically in narrative form, but it does not function as an operational constraint on what is accepted as an "outcome," what is deemed a meaningful difference, or what is treated as a legitimate object of arithmetic. That is the key: the floor is not a measure of ignorance in the casual sense; it is a measure of exclusion from operational reasoning.

The next cluster is equally decisive: the classical axioms that define admissible quantification. "Multiplication requires a ratio measure" sits at $p = 0.10$ (−2.20). "Measurement precedes arithmetic" and "meeting representational measurement axioms is required for arithmetic" both sit at $p = 0.15$ (−1.75). "Measures must be unidimensional" is also $p = 0.15$ (−1.75). These are not specialist Rasch propositions; they are the minimal conditions for treating numbers as quantities rather than labels or scores. If a knowledge base does not endorse these propositions, then it cannot credibly claim that its quantitative outputs are scientific measurements rather than administrative coding. In this profile EFIC does not merely fall short. It shows an inverted structure: arithmetic is treated as prior, and scale-type constraints are treated as negotiable. That inversion is the defining property of the false HTA global memeplex.

The profile also shows the familiar asymmetry between manifest quantities and composite or subjective constructs. Time is correctly identified as a ratio measure with p = 0.90 (+2.20). That matters because it demonstrates that the knowledge base is not uniformly hostile to measurement concepts. It can recognize a ratio scale when it is socially uncontested and physically grounded. The problem is that this recognition is not generalized into a discipline of quantification. Time remains an island of lawful measurement inside a wider architecture that treats non-measures as if they were measures. The moment the analysis moves from manifest duration to latent traits or multiattribute composites, the axioms collapse and scoring replaces measurement.

That replacement is visible in the "false measurement endorsement" cluster. EFIC shows p = 0.70 (+0.85) for the false statements that EQ-5D-3L preference algorithms create interval measures, that QALYs are ratio measures, that summations of subjective responses are ratio measures, that summating Likert scores creates a ratio measure, and that QALYs can be aggregated. These are the working assumptions of standard HTA practice. They are the numerical permissions that allow cost-per-QALY arithmetic to proceed without confronting admissibility. Yet in measurement science they represent category errors; they are in measurement and mathematical terms, meaningless. A preference algorithm that maps ordinal choices over multiattribute health states into a single index does not demonstrate interval structure, let alone ratio structure. It produces a scoring rule. It can be repeatable and convenient without being a measure. The EFIC profile indicates that within this knowledge base, the difference between a scoring rule and a measurement structure is not operationally enforced.

The QALY homogeneity issue appears in a slightly lower but still positive endorsement: "the QALY is dimensionally homogeneous" is p = 0.65 (+0.60). This is a critical tell. A knowledge base that treats the QALY as homogeneous is effectively endorsing a mixed-dimension arithmetic fiction: time, a ratio measure, is multiplied by a utility index whose scale properties are not established as ratio, and whose substantive meaning is not a single attribute but an amalgam of attributes plus valuation conventions. Even if one were to concede, for argument's sake, that the utility index was interval, multiplication by time would still be inadmissible. EFIC's pattern indicates that the knowledge base does not treat dimensional homogeneity as a necessary gatekeeper for arithmetic. It treats the product as meaningful because the institution needs it to be meaningful.

The item on "claims for cost-effectiveness fail the axioms of representational measurement" is endorsed at p = 0.10 (−2.20). This is an unusually clear diagnostic: the knowledge base is not merely silent on the failure; it rejects the proposition that the failure exists. This is exactly how a memeplex stabilizes itself. If the critical proposition cannot be endorsed, then the belief system does not have to defend itself on measurement grounds; it can treat critiques as out-of-scope. It can continue to produce cost-effectiveness claims with procedural confidence, because the very criterion by which those claims would be deemed invalid is excluded from the system's admissibility rules.

EFIC's profile also displays an important nuance regarding falsification. "Non-falsifiable claims should be rejected" is endorsed at p = 0.75 (+1.10). At first glance, this looks like a foothold: a commitment to scientific discipline. But its co-existence with endorsement of QALY arithmetic and simulation legitimacy reveals the deeper issue: the knowledge base can endorse falsification

while operationally relying on constructs that cannot be falsified as empirical quantities. In this ecosystem, falsification becomes rhetorical rather than structural. It is expressed as a general value while the quantitative machinery is built from assumptions, scoring algorithms, and composite indices that have no measurement referent against which they could be proven wrong. The profile makes that split explicit: a nominal endorsement of falsification alongside the exclusion of measurement prerequisites that would make falsification possible.

The simulation statement reinforces the same pattern. “Reference case simulations generate falsifiable claims” is rejected at p = 0.65 (+0.60) because the response is 0 (false). Reference case and associated simulations do not generate falsifiable claims because they are functions of assumptions. But the problem is that EFIC’s knowledge base, like the wider HTA environment, still permits simulation outputs to acquire evaluative authority through procedural stabilization. If they are not falsifiable, then they are not scientific claims about therapy impact; they are conditional projections. The EFIC table indicates partial recognition of this point, but without the measurement discipline that would replace simulation closure with empirically evaluable protocols.

The item “time trade-off preferences are unidimensional” is rejected (response 0) with p = 0.65 (+0.60). That is directionally correct: TTO preferences are not measures of a single attribute; they are preference judgments over bundled states. Yet again, this correct recognition does not propagate to a rejection of QALY arithmetic, because the evaluative system does not require unidimensional measurement as a binding constraint before arithmetic is allowed. The knowledge base can hold two incompatible positions simultaneously: it can deny unidimensionality of TTO while still treating the output as arithmetic-ready.

The “ratio measures can have negative values” item is rejected (response 0) with p = 0.30 (−0.85). This is weak endorsement of the correct position (that ratio measures do not take negative values if the zero is a true absence). The weakness matters because negative utilities are one of the operational escape hatches by which QALY frameworks reveal their lack of ratio structure. If the evaluative system were disciplined by ratio axioms, negative utilities would be an immediate stop sign: the construct cannot be ratio and therefore cannot be multiplied and aggregated as if it were. EFIC’s weak endorsement indicates that this stop does not function robustly inside the knowledge base.

The item “a linear ratio scale for manifest claims can always be combined with a logit scale” is rejected at p = 0.35 (−0.60). This is another structural marker of non-possession. Therapy impact requires parallel claim types: linear ratio claims for manifest events and Rasch logit ratio claims for latent traits, with protocols that respect each scale’s properties. If the knowledge base does not accept that these can co-exist coherently in a single evaluative platform, then it remains trapped in the false integrative ideal: one composite number to rule them all. That is precisely the administrative convenience logic EFIC inherits from the HTA environment. Importantly, it is only by accepting that these two claims type can coexist that HTA therapy impact analysis meets the required standards to support falsification and the evolution of objective knowledge. The existing HTA memeplex belief must be rejected completely.

Finally, consider the item on the "outcome of interest for latent traits is the possession of that trait," endorsed at p = 0.20 (−1.40). This is a subtle but decisive point. If therapy impact for latent traits is not framed in terms of Rasch trait possession on an invariant scale, then outcome assessment remains a matter of score change, category movement, or average differences on ordinal instruments. Those are descriptive artifacts, not measured therapy impacts. EFIC's low endorsement indicates that the possession framing is not structurally present as a guiding principle.

Taken together, the corrected logit profile places EFIC squarely within the global pattern that the global memeplex describes: strong operational endorsement of false measurement permissions, systematic exclusion of Rasch and representational prerequisites, and an evaluative posture oriented toward closure of cost-effectiveness imaginary claims rather than cumulative empirical learning. This matters for EFIC because pain medicine is precisely the domain in which patient-reported outcomes and latent constructs dominate. If any clinical area needs invariant measurement discipline, it is pain. Without Rasch transformation and the enforcement of unidimensionality and admissible arithmetic, claims about comparative therapy impact in pain are inherently unstable: differences cannot be interpreted as quantities, replication cannot accumulate into objective knowledge, and long-term reassessment cannot proceed as a scientific program.

That is where duty of care becomes unavoidable. An organization that influences standards, disseminates guidance, and shapes what clinicians and systems treat as credible evidence has an implicit obligation to ensure that its quantitative claims are more than numerically decorated narratives. If EFIC's knowledge base accepts scoring systems as if they were measures, it legitimizes decisions and recommendations that cannot be empirically adjudicated. This is not a philosophical quibble; it is a structural defect in the chain from patient experience to quantitative claim to clinical and policy decision. The consequence is predictable: disagreement is managed procedurally rather than resolved empirically; models and indices become immune to refutation; and the evolution of objective knowledge is halted because the system cannot generate claims that can be proven wrong.

The EFIC profile therefore does not call for incremental improvement or better reporting. It calls for reconstruction. If EFIC wishes to remain scientifically accountable in therapy evaluation, it must treat measurement as prior to arithmetic, reject composite utilities as measures, and adopt a two-track measurement platform: linear ratio measures for manifest attributes and Rasch logit ratio measures for latent traits. Anything less preserves the current inversion. It preserves the memeplex. And it preserves a future in which numerical authority is maintained without measurement validity.

EFIC is not alone. A parallel assessment of the HTA knowledge base of the US Association for the Study of Pain (IASP-US) yields an almost identical logit profile for the diagnostic statements; a systematic inversion of the axioms of representational measurement at the level of pain assessment, patient-reported outcomes, and health technology evaluation.

## ARE THERE OPTIONS FOR THE EUROPEAN PAIN FEDERATION (EFIC)?

The European Pain Federation is not unique in confronting the implications of representational measurement. It is part of a wider international knowledge environment in which composite scores,

ordinal summations, and preference-weighted indices have become institutionalized as if they were quantitative measures. This false condition did not arise from malice or incompetence, but from historical inheritance. Psychometric scoring systems were adopted because they provided practical tools for clinical research and appeared to offer numerical summaries of patient experience. Over time, these scoring systems acquired authority through repetition, institutional endorsement, and integration into regulatory and HTA frameworks. The resulting evaluative architecture is internally coherent, administratively convenient, and widely accepted. Yet coherence and acceptance do not confer measurement validity. The logit interrogation demonstrates that EFIC's knowledge base, like that of many other professional organizations, operates within a scoring paradigm rather than a measurement paradigm.

EFIC therefore faces a clear choice. It can continue to operate within its inherited framework, accepting composite scores as if they were measures and continuing to support evaluative claims that cannot be empirically falsified in the strict measurement sense. This path offers short-term stability. It avoids institutional disruption, preserves continuity with existing guidelines, and aligns with prevailing international HTA practices. However, it also commits EFIC to a future in which therapy impact claims remain structurally detached from measurement foundations. Numerical outputs will continue to function as administrative artifacts rather than empirical quantities. Over time this will diminish the scientific credibility of outcome claims while constraining the evolution of objective knowledge in pain medicine. Practitioners will be aware that critics and reviewers can always point to the absence of representational measurement in therapy impact claims.

The alternative is to commit explicitly to representational measurement as the foundation for therapy evaluation. This requires recognition of a principle that is both simple and non-negotiable: arithmetic operations are admissible only when applied to quantities that satisfy ratio scale requirements. Within this framework, there are only two lawful measurement structures for evaluating therapy impact. Manifest attributes, such as survival time, hospital days, or objectively observable resource utilization, must be measured on linear ratio scales possessing a true zero and invariant unit structure. Latent attributes, such as pain severity, functional impairment, or quality of life, must be measured through Rasch transformation, which produces invariant logit ratio scales derived from conjoint simultaneous measurement of persons and items. These logit scales satisfy the axioms of representational measurement and provide the only scientific basis for quantifying latent traits.

This transition does not require abandoning EFIC's commitment to rigorous outcome assessment. On the contrary, it strengthens that commitment by aligning outcome evaluation with the requirements of normal science. Existing instruments can be reconstructed within the Rasch framework, transforming ordinal responses into invariant measures while preserving clinical relevance. The focus shifts from scoring to measurement, from numerical convenience to empirical validity. Therapy impact claims become falsifiable, replicable, and capable of supporting cumulative scientific knowledge.

The choice EFIC faces is therefore not between practicality and idealism. It is between continued reliance on administratively convenient scoring systems and alignment with the scientific principles that govern all quantitative disciplines. Other fields confronted similar transitions in the past. The establishment of invariant measurement structures in physics, chemistry, and engineering

did not occur spontaneously; it required explicit recognition that measurement precedes arithmetic. Pain medicine now confronts the same moment of decision.

EFIC has the institutional authority, intellectual leadership, and educational reach to initiate this transition. By explicitly endorsing representational measurement and Rasch transformation as the standard for latent trait evaluation, EFIC could reposition pain medicine at the forefront of measurement-based clinical science. The alternative is to remain within a legacy framework whose numerical outputs possess administrative utility but lack measurement validity. The logit evidence demonstrates the present condition. The future depends on whether EFIC chooses to preserve that condition or to reconstruct its evaluative foundation on lawful measurement principles.

# III. THE TRANSITION TO MEASUREMENT IN HEALTH TECHNOLOGY ASSESSMENT

## THE IMPERATIVE OF CHANGE

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

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

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

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

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

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

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

## MEANINGFUL THERAPY IMPACT CLAIMS

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

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

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

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

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

## THE PATH TO MEANINGFUL MEASUREMENT

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

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

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

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

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

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

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

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

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

## TRANSITION REQUIRES TRAINING

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

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

Training must therefore begin with scale types and their permissible operations. Linear ratio measurement applies to manifest attributes that possess a true zero and invariant units, such as

time, counts, and resource use. Latent attributes, by contrast, cannot be observed directly and cannot be measured through summation or weighting. They require formal construction through a measurement model capable of producing invariant units. This distinction is the conceptual fulcrum of reform, because it determines which claims are admissible and which are not.

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

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

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

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

**A NEW START IN MEASUREMENT FOR HEALTH TECHNOLOGY ASSESSMENT**

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

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

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

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

## DESIGNED FOR CLOSURE

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

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

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

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

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

them to be posed. The absence of measurement standards was not an oversight; it was functionally essential.

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

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

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

**ACKNOWLEDGEMENT**

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

## REFERENCES

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

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

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

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