MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

CANADA: HEALTH TECHNOLOGY ASSESSMENT INTERNATIONAL (HTAi) AND THE COMMITMENT TO FALSE MEASUREMENT

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

LOGIT WORKING PAPER No 59 FEBRUARY 2026

www.maimonresearch.com

Tucson AZ

FOREWORD

HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

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

The objective of this study is to evaluate the epistemic foundations of Health Technology Assessment International (HTAi) using the canonical 24-item diagnostic grounded in representational measurement theory. The purpose is not to assess HTAi's organizational effectiveness, governance structure, or international reach, but to determine whether the knowledge environment created and sustained by the organization recognizes the axioms required for legitimate quantitative inference. These axioms include the logical precedence of measurement over arithmetic, the requirement of unidimensionality, the distinction between ordinal, interval, and ratio scales, the necessity of invariance, and the conditions under which latent constructs may be transformed into quantities.

The analysis treats HTAi as an epistemic authority rather than an implementing body. Through its annual meetings, policy forums, working groups, training activities, and close association with its flagship journal, HTAi defines and disseminates what constitutes acceptable HTA methodology worldwide. By applying a standardized probability–logit interrogation framework to this institutional corpus, the study seeks to determine whether HTAi's global leadership role is grounded in principles consistent with quantitative science or whether it perpetuates numerical conventions that cannot support empirically evaluable value claims.

The canonical assessment demonstrates that the HTAi knowledge base exhibits a fully consolidated epistemic structure in which the axioms of representational measurement do not function as governing constraints. Propositions defining the necessary conditions for quantitative science consistently collapse toward the lower bound of endorsement, while propositions that violate those conditions receive strong reinforcement. Measurement does not precede arithmetic within HTAi discourse; instead, arithmetic is treated as methodologically necessary and institutionally protected.

At the same time, the assumptions required to sustain cost-utility analysis are strongly reinforced. Preference-based utilities are treated as interval or ratio measures, negative values are accepted on purported ratio scales, summated subjective responses are treated as quantitative magnitudes, and

QALYs are assumed to support aggregation and multiplication. Rasch measurement principles, which provide the only lawful basis for transforming ordinal responses into quantitative latent-trait measures, are almost entirely absent. The resulting probability–logit profile is internally coherent and structurally invariant with those observed across national agencies and academic HTA research centers. The findings indicate not misunderstanding or inconsistency, but systematic non-possession of measurement theory at the level where HTA standards are internationally defined.

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 HTAi KNOWLEDGE BASE

The knowledge base of Health Technology Assessment International is best understood as a global epistemic system rather than as a collection of discrete organizational outputs. HTAi functions as the primary international forum through which HTA concepts, methods, and norms are articulated, circulated, and legitimized. Its authority derives not from regulatory power, but from its role in convening experts, framing methodological discourse, and defining what constitutes acceptable practice within the HTA community.

This epistemic system is constructed through multiple channels. Annual scientific meetings serve as sites where methodological consensus is reinforced through repetition rather than adjudication. Policy forums and special interest groups develop position statements and guidance documents that codify prevailing assumptions. Training workshops and educational initiatives transmit these assumptions to new cohorts of analysts. Through these mechanisms, HTAi shapes the intellectual environment within which national agencies, academic centers, and practitioners operate.

Central to this knowledge base is the normalization of cost-utility analysis as the dominant evaluative framework. HTAi materials routinely treat incremental cost per QALY as the benchmark for value assessment. This orientation establishes arithmetic as the organizing principle of evaluation. Numbers must be produced, compared, and aggregated to enable decision making. The question of whether those numbers satisfy the conditions required for measurement is displaced by the operational need for comparability.

Preference-based instruments occupy a foundational position within this system. Utilities derived from instruments such as the EQ-5D are presented as quantitative measures of health suitable for arithmetic operations. The transformation of ordinal descriptive responses into numerical utilities is treated as unproblematic. Valuation is assumed to confer measurement properties, despite the absence of any explicit transformation model demonstrating invariance or unit structure.

Methodological pluralism is often invoked within HTAi discourse, yet this pluralism operates within narrow epistemic boundaries. Debate occurs over model structure, discount rates, or perspective, but not over whether utilities constitute quantities or whether QALYs are mathematically coherent. Measurement theory does not function as a reference point for methodological disagreement. As a result, fundamental assumptions remain insulated from challenge.

The close institutional relationship between HTAi and its flagship journal further reinforces this structure. Published research overwhelmingly reflects the same numerical conventions promoted through HTAi activities. Peer review evaluates submissions based on conformity to accepted practice rather than coherence with measurement axioms. This reciprocal reinforcement between organization and journal stabilizes the epistemic environment.

Importantly, the HTAi knowledge base is not unified by explicit theoretical claims about measurement. The organization does not assert that utilities satisfy representational axioms; it proceeds as if they do. Legitimacy arises through consensus and repetition. Over time, repeated use substitutes for justification.

This process produces epistemic closure. Concepts such as unidimensionality, scale type, invariance, and arithmetic permission fall outside the boundaries of legitimate inquiry. They are not debated because they are not recognized as relevant. Within such a system, numerical practices appear self-evident even when they lack empirical grounding.

The result is a global HTA knowledge environment in which numbers circulate with authority while remaining detached from the conditions that make quantitative science possible. HTAi's influence ensures remarkable international consistency in HTA practice, but that consistency reflects shared assumptions rather than shared measurement principles.

Understanding HTAi as an epistemic authority is therefore essential to interpreting the canonical diagnostic results. The absence of measurement theory at this level explains its absence elsewhere. When axioms are not present where standards are defined, they cannot emerge where standards are implemented. The HTAi knowledge base thus plays a decisive role in sustaining numerical storytelling as the dominant mode of HTA reasoning worldwide.

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: QUEBEC INESSS

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.

TABLE 1: ITEM STATEMENT, RESPONSE, ENDORSEMENT AND NORMALIZED LOGITS HTAi

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	0.15	-1.75
TIME TRADE-OFF PREFERENCES ARE UNIDIMENSIONAL	0	0.80	+1.40
RATIO MEASURES CAN HAVE NEGATIVE VALUES	0	0.85	+1.75

EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES	0	0.80	+1.40
THE QALY IS A RATIO MEASURE	0	0.85	+1.75
TIME IS A RATIO MEASURE	1	0.75	+1.15
MEASUREMENT PRECEDES ARITHMETIC	1	0.15	-1.75
SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES	0	0.85	+1.75
MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC	1	0.15	-1.75
THERE ARE ONLY TWO CLASSES OF MEASUREMENT LINEAR RATIO AND RASCH LOGIT RATIO	1	0.10	-2.20
TRANSFORMING SUBJECTIVE RESPONSES TO INTERVAL MEASUREMENT IS ONLY POSSIBLE WITH RASH RULES	1	0.10	-2.20
SUMMATION OF LIKERT QUESTION SCORES CREATES A RATIO MEASURE	0	0.85	+1.75
THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE	0	0.80	+1.40
CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.10	-2.20
QALYS CAN BE AGGREGATED	0	0.85	+1.75
NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.25	-1.15
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.85	+1.75
THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.35	-0.65
THE RASCH LOGIT RATIO SCALE IS THE ONLY BASIS FOR ASSESSING THERAPY IMPACT FOR LATENT TRAITS	1	0.10	-2.20
A LINEAR RATIO SCALE FOR MANIFEST CLAIMS CAN ALWAYS BE COMBINED WITH A LOGIT SCALE	0	0.75	+1.15
THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT	1	0.10	-2.20
THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL	1	0.05	-2.50

TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT			

HEALTH TECHNOLOGY ASSESSMENT INTERNATIONAL (HTAi): BETWEEN A ROCK AND A HARD PLACE

The canonical diagnostic assessment of Health Technology Assessment International (HTAi) reveals with striking clarity that the organization functions as a central epistemic authority within global HTA while simultaneously lacking possession of the axioms required for quantitative science. The probability–logit profile does not suggest institutional ambiguity, methodological pluralism, or theoretical transition. Instead, it reveals a fully consolidated belief system in which numerical reasoning is privileged and protected, while the principles that determine whether numbers represent quantities are systematically absent.

HTAi occupies a unique position within the HTA ecosystem. It is not a reimbursement authority, a regulator, or a funding agency. Its authority is epistemic rather than administrative. Through annual meetings, policy forums, working groups, methodological task forces, training programs, and its close institutional relationship with the *International Journal of Technology Assessment in Health Care*, HTAi defines what HTA is understood to be. It does not merely disseminate methods. It legitimizes them.

For that reason, the canonical diagnostic is particularly revealing when applied to HTAi. If measurement axioms were to exist anywhere within HTA, they would be expected to appear here. The absence of such axioms within an organization dedicated to "advancing HTA worldwide" therefore carries profound implications. The diagnostic profile demonstrates that foundational propositions governing quantitative measurement collapse toward the lower bound of endorsement. The proposition that measurement must precede arithmetic registers at p = 0.15 (−1.75). This indicates that within HTAi's knowledge environment, numerical operations are not conditioned on prior establishment of measurement validity. Arithmetic is treated as a methodological tool rather than as a logically constrained operation.

Closely related is the equally weak endorsement of the proposition that arithmetic must satisfy the axioms of representational measurement. At p = 0.15 (−1.75), scale-type admissibility does not operate as a gatekeeping principle within HTAi discourse. There is no epistemic mechanism within HTAi outputs that prohibits multiplication, aggregation, or ratio construction when scale properties are undefined. The rejection of the proposition that multiplication requires a ratio measure is particularly revealing. This item collapses to p = 0.15 (−1.75), despite multiplication serving as the mathematical foundation of cost-utility analysis. Utilities are multiplied by time, QALYs are aggregated across individuals, and ratios are constructed as central decision metrics. The diagnostic exposes the necessary inversion that sustains this framework. Multiplication is not permitted because it is lawful; it is assumed because it is required.

In contrast, propositions that support QALY arithmetic receive strong endorsement. The claim that the QALY is a ratio measure registers at p = 0.85 (+1.75). The proposition that QALYs can be aggregated receives the same level of reinforcement. These values do not reflect empirical demonstration. They reflect doctrinal necessity. The QALY must be treated as a ratio measure or the HTA enterprise collapses.

The endorsement of negative values on purported ratio scales further reveals the depth of epistemic inversion. The proposition that ratio measures can have negative values is endorsed at p = 0.85 (+1.75). This directly contradicts the defining property of ratio measurement: a true zero representing absence of the attribute. Yet within HTAi discourse, "states worse than dead" are treated as routine analytic constructs. The contradiction generates no tension because the axioms that would render it incoherent are not recognized.

Preference-based instruments play a central role in sustaining this structure. The proposition that EQ-5D preference algorithms create interval measures is endorsed at p = 0.80 (+1.40). This reflects HTAi's consistent promotion of preference-based outcomes as quantitative measures suitable for arithmetic. Valuation is treated as transformation. Preference is treated as magnitude. Similarly, the strong endorsement of the proposition that summated subjective responses create ratio measures confirms the absence of distinction between scoring and measurement. Numbers are treated as quantities by virtue of their numerical form alone.

Against this background, the collapse of Rasch-related propositions is decisive. The claim that transforming subjective responses into interval measurement is only possible under Rasch rules registers at p = 0.10 (−2.20). The proposition that the Rasch logit ratio scale provides the only defensible basis for latent-trait measurement collapses to the same level. The equivalence between Rasch axioms and representational measurement theory reaches the absolute floor at p = 0.05 (−2.50). These results do not indicate rejection following evaluation. They indicate epistemic absence. Rasch measurement does not function as a conceptual reference point within HTAi activities. It is not debated, challenged, or contrasted. It is not part of the discourse.

The proposition that the outcome of interest for latent traits is possession of that trait likewise collapses to p = 0.10 (−2.20). This reveals a fundamental displacement of the object of inquiry. HTAi discourse does not conceptualize outcomes as attributes possessed by individuals. Instead, it treats outcomes as values assigned to hypothetical states through social preference elicitation. Measurement of persons is replaced by valuation of descriptions. This distinction is crucial. Measurement concerns empirical attributes of individuals. Valuation concerns judgments of observers. HTAi systematically conflates the two, treating preference scores as if they were measures of health.

The endorsement of reference-case simulation as a source of falsifiable claims further stabilizes this system. The proposition that reference-case simulations generate falsifiable claims is endorsed at p = 0.85 (+1.75). Modeled futures are treated as evidence despite lacking observable referents. Falsifiability is redefined as internal sensitivity analysis rather than confrontation with empirical data. Although the proposition that non-falsifiable claims should be rejected registers modestly higher at p = 0.25 (−1.15), this value reflects rhetorical commitment rather than operational enforcement. HTAi actively promotes long-term modeling as best practice while lacking any mechanism for empirical verification of its outputs.

Taken together, the diagnostic reveals a fully coherent epistemic system. Measurement axioms are excluded. Valuation is elevated. Arithmetic is protected. Simulation replaces observation. Each component reinforces the others, producing a stable global framework that appears methodologically rigorous while remaining epistemically ungrounded.

The most striking feature of the HTAi profile is its invariance. The probability–logit pattern is effectively identical to those observed for CADTH/CDA-AMC, INESSS, Canadian academic research centers, and international HTA agencies. This invariance is not coincidental. HTAi functions as the institutional mechanism through which this epistemic structure is standardized and disseminated globally.

HTAi does not merely reflect HTA practice. It curates it. Through its policy forums, working groups, and educational initiatives, it establishes what counts as acceptable evidence. The absence of representational measurement theory within these activities therefore has cascading consequences. When axioms are absent at the global level, they cannot appear at the national level. This explains the remarkable global uniformity of HTA practice despite diverse health systems. The epistemic core is shared. Countries do not independently invent QALYs, utilities, and reference cases. They inherit them.

The diagnostic therefore reveals HTAi not as a passive convener but as an epistemic gatekeeper. Its authority does not derive from coercion but from consensus. Yet consensus without measurement is not science. It is convention. The implication is profound. Reform cannot originate within agencies alone. As long as HTAi continues to legitimate non-measurement as quantitative evidence, national reforms will remain constrained by international norms. Change requires epistemic leadership at the level where standards are defined.

The canonical assessment does not accuse HTAi of methodological negligence. It identifies structural non-possession. The organization operates coherently within a paradigm it did not invent but now sustains. The question it faces is not whether its practices are widely accepted, but whether acceptance can substitute for measurement. Until HTAi confronts the axioms of representational measurement, HTA will remain a discipline that manipulates numbers without quantities. The diagnostic makes that condition explicit. It exposes the epistemic foundation upon which forty years of HTA practice have been built. In doing so, it reframes the challenge facing HTAi. The issue is not how to improve models, refine value sets, or harmonize methods. The issue is whether HTA wishes to remain a numerical belief system or become a quantitative science. That choice has not yet been made.

NEXT STEPS: THE FUTURE OPTIONS FOR HTAi

The canonical diagnostic assessment places Health Technology Assessment International at a crossroads. The findings do not identify a set of methodological deficiencies that can be addressed through incremental refinement. They reveal a structural condition: the absence of representational measurement as a governing authority within the global HTA knowledge system. This condition cannot be resolved by better modeling guidance, expanded training programs, or further harmonization of existing practice. It requires a choice about the future identity of HTAi itself.

One option is institutional continuity. HTAi may continue to function as the steward of established HTA conventions, refining reference cases, improving transparency standards, and supporting methodological convergence across jurisdictions. This path preserves stability and protects institutional legitimacy in the short term. Yet it also implies acceptance of the diagnostic conclusion. If numerical constructs such as utilities and QALYs do not satisfy the axioms of measurement, then continued promotion of these constructs transforms HTAi from a scientific convener into a curator of numerical belief systems. The organization would remain influential, but its authority would rest on convention rather than empirical accountability.

A second option is managed pluralism. HTAi could acknowledge the limits of current quantitative frameworks while permitting parallel methodological approaches to coexist. Under this model, preference-based evaluation, deliberative assessment, and emerging measurement-based approaches would be treated as complementary perspectives. While this appears inclusive, it avoids the central issue. Measurement axioms are not methodological preferences. They define when arithmetic is permissible. Pluralism cannot reconcile mutually exclusive logical claims. Treating non-measurement and measurement as interchangeable positions preserves ambiguity but does not resolve incoherence.

A third option is epistemic leadership. HTAi could explicitly recognize representational measurement theory as a foundational constraint on quantitative inference and reposition itself as the global forum for rebuilding HTA on scientifically defensible grounds. This would not require abandoning evaluation or decision support. It would require restoring the logical sequence by which claims become numbers and numbers become evidence.

Under such a transition, HTAi's role would change fundamentally. Rather than promoting specific modeling architectures, the organization would establish admissibility standards for quantitative claims. Numerical outputs would be accepted only where scale properties are demonstrated. Manifest outcomes expressed on linear ratio scales would be distinguished from latent outcomes requiring Rasch transformation. Preference-based valuation would be clearly separated from measurement and repositioned as deliberative input rather than quantitative evidence.

This shift would not weaken HTAi's influence. It would redefine it. The organization would become the custodian of epistemic integrity rather than the harmonizer of convention. Its conferences would become venues for adjudicating measurement claims rather than rehearsing modeling practices. Its working groups would focus on defining protocols for evaluable claims rather than refining hypothetical projections. Its educational programs would train analysts not merely in technique, but in the logical conditions that make technique meaningful.

Such a transition would necessarily be gradual. HTAi cannot invalidate four decades of practice overnight. Nor should it attempt to do so. What it can do is establish a clear forward boundary. New methodological guidance can be framed around evaluable claims rather than reference-case simulations. Training programs can introduce measurement theory as foundational rather than optional. Journals associated with HTAi can begin requiring explicit articulation of scale properties for quantitative outcomes. Over time, the evidentiary superiority of measurement-based claims would assert itself.

The risks of this path are institutional rather than scientific. Epistemic leadership requires confronting deeply entrenched beliefs. It may provoke resistance from stakeholders whose professional identities are tied to existing frameworks. Yet scientific progress has never occurred without such disruption. The credibility of HTA depends not on preserving consensus, but on ensuring that consensus rests on defensible foundations.

The alternative is stagnation. As external scrutiny of HTA intensifies, reliance on non-falsifiable models and imaginary quantities will become increasingly difficult to defend. Agencies will continue to produce precise numbers that cannot be tested. Public confidence will erode. HTAi's role as a global authority will weaken as its methods lose persuasive power outside the HTA community itself.

The diagnostic assessment therefore presents HTAi with an opportunity rather than a threat. Few organizations are positioned to redefine an entire field's epistemic foundations. HTAi is uniquely situated to do so. Its international reach, convening power, and institutional legitimacy provide the platform required for reform. The future of HTAi depends on whether it chooses to remain the steward of numerical tradition or to become the architect of a measurement-based HTA. The former offers comfort. The latter offers coherence. Only one offers scientific legitimacy. The question confronting HTAi is therefore not how HTA should be improved, but what HTA should be. The answer will determine whether the field evolves into a discipline grounded in evaluable claims or remains an elaborate system of numerical storytelling.

GIVEN ITS LEGACY: DOES HTAi HAVE A FUTURE?

Health Technology Assessment International stands today as one of the most influential institutions in global health policy. Over four decades, it has helped shape the language, methods, and professional identity of HTA across jurisdictions. Its conferences, networks, and affiliated publications have provided a shared forum through which national agencies, academic centers, and practitioners have come to understand what HTA is and how it should be practiced. That legacy is substantial and cannot be dismissed. Yet it is precisely this legacy that now places HTAi in question.

The canonical diagnostic assessment does not challenge HTAi's historical importance. It challenges the epistemic foundations upon which that importance was built. The organization emerged during a period when numerical modeling offered the promise of rational decision making in complex health systems. Cost-utility analysis, preference-based instruments, and long-term simulation appeared to provide a scientific language for allocating scarce resources. These tools were adopted not because they satisfied the axioms of measurement, but because no alternative framework for quantification was available at the time.

What began as pragmatic approximation gradually hardened into doctrine. Over time, the distinction between numerical convenience and quantitative legitimacy eroded. Utilities came to be treated as measures. QALYs were endowed with arithmetic properties they could not possess. Simulation outputs were interpreted as evidence rather than conjecture. HTAi did not invent these constructs, but it became the institution through which they were normalized and globalized.

That normalization now confronts a limit. The axioms of representational measurement have not changed. The logical conditions for quantitative inference remain what they have always been. What has changed is the visibility of their absence. When a field repeatedly produces numbers that cannot be empirically tested, replicated, or falsified, confidence eventually gives way to scrutiny. The diagnostic assessment renders explicit what routine practice has concealed: that much of HTA's numerical apparatus does not measure anything.

This recognition places HTAi in an unprecedented position. Institutions built around methodological consensus rarely face epistemic reckoning. Yet consensus cannot substitute for coherence. An organization whose authority rests on harmonizing practice must confront the possibility that what has been harmonized is not scientifically defensible.

Does this mean HTAi has no future? Not necessarily. But it does mean that its future cannot be a continuation of its past.

If HTAi remains committed to defending the existing HTA paradigm, its role will gradually diminish. Numerical outputs that cannot be evaluated will lose persuasive power beyond the HTA community itself. Policymakers, clinicians, and the public will increasingly question decisions justified by imaginary futures and unverifiable ratios. In such an environment, HTAi risks becoming an inward-facing professional society, sustaining belief rather than advancing knowledge.

Alternatively, HTAi may reinterpret its legacy not as a constraint but as a foundation for transformation. Its history provides legitimacy. Its networks provide reach. Its convening power provides the capacity to initiate reform. Few institutions are better positioned to acknowledge the limits of the current paradigm and to articulate a new one grounded in measurement, evaluation, and learning.

Such a transformation would require intellectual humility. It would require accepting that long-standing practices, widely taught and published, cannot be reconciled with the axioms of quantitative science. Yet science advances not by defending its approximations, but by replacing them when their limits become visible. HTAi's future depends on whether it is willing to facilitate that replacement.

The question is not whether HTAi should abandon HTA. It is whether it is willing to redefine what HTA means. A discipline centered on evaluable claims, measurable outcomes, and protocol-driven assessment would still require international coordination, methodological dialogue, and institutional leadership. In that future, HTAi could play a central role but not as the guardian of numerical tradition, but as the steward of epistemic integrity.

Legacy organizations often struggle at moments like this. Their authority is inseparable from the frameworks they helped institutionalize. Yet clinging to these frameworks risk transforming legacy into liability. The true test of institutional maturity is not longevity, but adaptability in the face of fundamental critique.

HTAi's future therefore turns on a single question. Will it continue to organize a field around numbers that cannot be measured, or will it help lead the transition toward claims that can be tested? One path preserves comfort. The other restores scientific meaning.

Whether HTAi has a future is not determined by its history. It is determined by how it responds to the limits of that history.

3. THE TRANSITION TO MEASUREMENT IN HEALTH TECHNOLOGY ASSESSMENT

THE IMPERATIVE OF CHANGE

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

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

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

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

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

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

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

MEANINGFUL THERAPY IMPACT CLAIMS

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

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

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

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

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

THE PATH TO MEANINGFUL MEASUREMENT

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

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

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

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

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

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

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

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

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

TRANSITION REQUIRES TRAINING

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

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

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

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

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

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

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

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