MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

AUSTRIA: INSTITUTE FOR HEALTH TECHNOLOGY ASSESSMENT (AIHTA) AND THE OPERATIONALIZING OF FALSE MEASUREMENT

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

LOGIT WORKING PAPER No 244 FEBRUARY 2026

www.maimonresearch.com

Tucson AZ

FOREWORD

HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

The Austrian Institute for Health Technology Assessment (AIHTA) is Austria's national authority for the scientific evaluation of medical technologies, pharmaceuticals, and clinical interventions. Established in 2020 as the successor to the Ludwig Boltzmann Institute for Health Technology Assessment, AIHTA operates as an independent research institute providing evidence assessments to support reimbursement, coverage, and policy decisions within Austria's publicly funded healthcare system. Its primary operational role is to evaluate the clinical effectiveness, safety, and economic implications of new and existing health technologies, including hospital interventions, pharmaceuticals, medical devices, and screening programs. These assessments inform decision-making by national and regional health authorities, particularly the Federal Ministry of Social Affairs, Health, Care and Consumer Protection, and the Austrian social health insurance system.

AIHTA produces structured assessment reports, horizon scanning evaluations, and methodological analyses that guide coverage decisions, reimbursement eligibility, and clinical adoption. Its work contributes directly to determining whether therapies are funded within Austria's social insurance framework and whether hospitals adopt new interventions. Although AIHTA does not make binding reimbursement decisions itself, its assessments serve as the scientific basis for operational decisions made by national payers and healthcare authorities. Through this role, AIHTA functions as the analytic core of Austria's HTA system, translating clinical and economic evidence into formal recommendations that influence therapy availability, pricing negotiations, and national healthcare resource allocation.

The objective of this study was to determine whether the Austrian Institute for Health Technology Assessment (AIHTA), as Austria's principal scientific body supporting health technology assessment and reimbursement-related evaluation, operates within the axioms of representational measurement when advancing quantitative claims concerning therapy effectiveness, comparative value, and resource implications. While AIHTA presents itself as an independent scientific institute tasked with providing rigorous evidence synthesis and economic evaluation, the critical question addressed here is more fundamental: whether the quantitative constructs embedded in its methodological guidance, assessment reports, and evaluative frameworks satisfy the structural requirements necessary for lawful arithmetic. Using the 24-item canonical representational measurement diagnostic, the study interrogated the AIHTA knowledge base to determine whether it enforces the prerequisites for admissible multiplication, division, and aggregation, including unidimensionality, dimensional homogeneity, invariant unit structure, and the requirement that latent constructs be transformed through Rasch measurement to establish interval and ratio properties. The objective was not to evaluate procedural sophistication or policy influence, but to determine whether the institute's evaluative architecture is grounded in measurement-valid constructs capable of supporting empirically evaluable, replicable, and falsifiable claims concerning therapy impact.

The logit profile demonstrates systematic exclusion of representational measurement axioms as operational constraints within the AIHTA knowledge base. Core propositions asserting that measurement must precede arithmetic, that multiplication requires ratio measurement, and that latent traits must be transformed through Rasch measurement to establish invariant interval structure collapse to floor or near-floor logit values, indicating effective non-possession of these principles within the evaluative framework. Conversely, false propositions required to sustain composite utility arithmetic—including the treatment of EQ-5D-derived utility scores and QALYs as ratio measures and the admissibility of summated ordinal preference scores as arithmetic objects—are strongly endorsed, with positive logit values indicating structural reinforcement within the knowledge base. This asymmetry demonstrates that AIHTA recognizes ratio measurement when evaluating manifest attributes such as time but does not apply the same discipline to latent constructs such as quality of life. The resulting evaluative framework performs arithmetic operations on composite indices that do not satisfy representational measurement axioms. Quantitative outputs therefore function as administratively coherent scoring constructs rather than empirically grounded measurement outcomes capable of supporting falsification, replication, or cumulative scientific knowledge.

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 AUSTRIAN INSTITUTE FOR HEALTH TECHNOLOGY ASSESSMENT KNOWLEDGE BASE

The Austrian Institute for Health Technology Assessment occupies a central role within Austria’s decentralized but scientifically structured HTA environment. Established as the successor to the Ludwig Boltzmann Institute for Health Technology Assessment, AIHTA provides systematic assessments, methodological guidance, and evidence syntheses intended to inform reimbursement decisions, clinical policy development, and resource allocation within the Austrian healthcare system. Although Austria does not operate a single centralized reimbursement authority analogous to NICE in England or TLV in Sweden, AIHTA functions as the primary scientific institution responsible for evaluating therapeutic interventions and supporting national and regional decision-making bodies. Its reports address clinical effectiveness, comparative therapeutic value, and economic implications, with economic evaluation frequently expressed through cost-effectiveness frameworks aligned with international HTA standards.

The institute’s methodological framework reflects extensive integration with European and international HTA practice, particularly through collaboration with EUnetHTA and adherence to internationally recognized economic evaluation conventions. These conventions include the use of preference-weighted composite instruments such as the EQ-5D and Health Utilities Index to derive utility scores representing health-related quality of life. These utility scores are then combined with survival time to produce quality-adjusted life years (QALYs), which function as integrative outcome measures intended to summarize therapy impact. Economic models incorporate these composite utility measures alongside projected resource utilization to generate cost-effectiveness ratios intended to inform comparative value assessment.

Simulation modeling plays a central operational role in the knowledge base. Clinical trial results, epidemiological data, and observational evidence are integrated into decision models designed to project long-term therapy impact beyond observed follow-up periods. These models generate estimates of incremental costs and incremental QALYs, allowing comparison across therapies and disease areas. Sensitivity analyses are used to explore uncertainty and variability in model assumptions. The resulting quantitative outputs provide decision makers with numerical summaries of anticipated therapeutic value and resource implications.

Despite procedural rigor and methodological sophistication, the quantitative constructs embedded within this knowledge base originate from composite preference indices whose measurement properties are not established through representational measurement transformation. Utility scores derived from multiattribute classification systems reflect aggregated preference judgments rather than invariant measurements of a unidimensional attribute. Their numerical values depend on scoring algorithms and valuation conventions rather than demonstration of invariant unit structure. Nonetheless, these scores are treated as if they support arithmetic operations, including multiplication, aggregation, and ratio comparison.

This framework reflects institutional adoption of internationally standardized HTA methods rather than independent reconstruction of measurement foundations. AIHTA's knowledge base demonstrates consistency with European HTA methodological norms and facilitates cross-jurisdictional comparability of economic evaluation outputs. However, it does not enforce the measurement prerequisites necessary to ensure that arithmetic operations performed on latent constructs satisfy representational measurement axioms. As a result, composite utility scores function operationally as decision variables despite lacking demonstrated ratio scale properties.

The institute therefore operates within a quantitatively structured but measurement-indifferent evaluative architecture. It demonstrates procedural transparency, analytical consistency, and institutional rigor in applying established HTA conventions. Yet its quantitative framework relies on composite indices that function as scoring constructs rather than invariant measurement outcomes. The AIHTA knowledge base thus exemplifies a structurally coherent but epistemically constrained HTA system in which numerical sophistication coexists with the systematic exclusion of the axioms required for empirically valid measurement.

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: AUSTRIAN INSTITUTE FOR HRALTH TECHNOLOGY ASSESSMENT (AIHTA)

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.

AUSTRIA: AIHTA INVERTS REPRESENTATIONAL MEASUREMENT

The AIHTA logit profile is not the portrait of an institution that has "partially" misunderstood the foundations of measurement; it is the signature of an HTA knowledge base in which measurement has been logically inverted and then stabilized as if this inversion were methodological sophistication (Table 1). AIHTA presents itself as an analytic body supporting evidence-based health policy and decision making. It is the successor to the Ludwig Boltzmann Institute for HTA and has positioned itself as a national reference point for assessment reports and methodological work. Yet the question the canonical interrogation asks is more basic than institutional mission statements: do the numbers used in evaluation actually satisfy the axioms that make quantification possible? The answer delivered by the profile is that they do not, and worse, that the knowledge base behaves as if the axioms are not binding constraints but optional philosophical commentary.

TABLE 1: ITEM STATEMENT, RESPONSE, ENDORSEMENT AND NORMALIZED LOGITS AUSTRIAN INSTITUTE FOR HEALTH TECHNOLOGY ASSESSMENT (AIHTA)

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

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

Begin with the most revealing structural fact: the items that define measurement as a prerequisite for arithmetic do not merely score low; they collapse toward the floor. “Meeting the axioms of representational measurement is required for arithmetic” sits at an endorsement probability of 0.10 (−1.87). “Measurement precedes arithmetic” sits at 0.15 (−1.47). In combination these results state something devastating: the AIHTA knowledge base does not operationalize the idea that numerical operations are conditional on scale properties. Arithmetic is treated as a tool that can be applied to whatever numbers are convenient, rather than as an operation whose meaningfulness depends on admissible transformations and scale type. This is not a minor technical deficiency. It is the definitional line between measurement-based science and numerical storytelling. In normal science, one does not begin by multiplying and aggregating and then later wonder what the numbers meant. One begins by establishing the measurement structure that legitimizes the arithmetic. The logit profile shows that, within AIHTA’s evaluative framework, that order is reversed.

The second structural feature is the striking asymmetry between what the knowledge base recognizes about manifest quantities and what it pretends about latent constructs. “Time is a ratio measure” is correctly endorsed at 0.95 (+2.50). Time, with a true zero and invariant unit, is a canonical example of a ratio measure. The AIHTA knowledge base can therefore recognize lawful ratio measurement when it is obvious and uncontroversial. But immediately the asymmetry appears: the same discipline is not extended to the latent traits that dominate HTA rhetoric, such as “health-related quality of life,” “quality-adjusted survival,” or “utility.” Instead, the profile shows systematic endorsement of false propositions required to make the QALY machine run. The items “EQ-5D-3L preference algorithms create interval measures” (FALSE) and “the QALY is a

ratio measure" (FALSE) both sit at high endorsement (0.85–0.90, +1.47 to +1.87). These positive logits represent the institutional normalization of a category error: treating preference-weighted composite scores as measurement outcomes.

This is where the duty-of-care implication becomes unavoidable. If an agency's operational focus is to support decisions about therapy adoption, reimbursement, or displacement, then "effectiveness" must be anchored in empirically evaluable claims. That requires outcomes measured on linear ratio scales for manifest attributes (e.g., hospital days avoided, exacerbations prevented, time to relapse, survival time) and Rasch logit ratio scales for latent traits (where subjective observation is unavoidable). The AIHTA profile shows that this is not the framework AIHTA operationalizes. It operationalizes a framework where "effectiveness" is collapsed into an invented composite—QALYs—built by multiplying a manifest ratio measure (time) by a preference score whose scale properties are neither established nor treated as binding constraints. This multiplication is not a technical detail. It is the core arithmetic step that claims to convert "health state descriptions" into "units of health." Yet the profile indicates that the knowledge base treats this conversion as legitimate by default.

Consider now the Rasch block, which is the cleanest discriminator between a measurement culture and a scoring culture. "There are only two classes of measurement: linear ratio and Rasch logit ratio" sits at 0.05 (−2.50). "Transforming subjective responses to interval measurement is only possible with Rasch rules" sits at 0.05 (−2.50). "The Rasch logit ratio scale is the only basis for assessing therapy impact for latent traits" sits at 0.05 (−2.50). "The Rasch rules for measurement are identical to the axioms of representational measurement" sits at 0.05 (−2.50). These four floor collapses are not an accident of emphasis; they reveal the operational identity of the knowledge base. A floor value denotes effective non-possession: the propositions do not function as constraints on what counts as admissible evidence or admissible quantification. They may appear in occasional narrative acknowledgments, but they do not govern practice. In practical terms, AIHTA behaves like a scoring culture: it prefers to treat ordinal survey responses, preference rankings, or multiattribute classifications as if they were already measures, so that the system can proceed to the preferred endpoint—cost-effectiveness ratios and threshold discussion—without the discipline of measurement construction.

The profile reinforces this conclusion through the cluster of high endorsements for false arithmetic propositions. "Summation of Likert question scores creates a ratio measure" (FALSE) sits at 0.90 (+1.87). "Summations of subjective instrument responses are ratio measures" (FALSE) sits at 0.85 (+1.47). These are not marginal errors; they are the enabling myths of HTA arithmetic. If you accept these propositions, you can do anything: add, average, multiply, discount, aggregate across patients, and then declare you have "quantified quality of life." If you reject them, the entire cost-per-QALY edifice collapses, and you are forced back to the discipline of measurement—unidimensionality, invariance, and admissible transformation. The AIHTA profile indicates the knowledge base chooses the myths.

The unidimensionality items provide another structural diagnosis. "Measures must be unidimensional" is endorsed at 0.15 (−1.47). "Time trade-off preferences are unidimensional" is endorsed strongly in the FALSE direction (0.85, +1.47). These are mutually reinforcing. HTA is saturated with multiattribute instruments precisely because unidimensional measurement was

never enforced. The EQ-5D is explicitly multiattribute. The HUI is explicitly multiattribute. AQoL is explicitly multiattribute. They do not measure a single attribute; they classify states across multiple domains and then apply preference weights. That is why they can generate a single index only by imposing an external scoring rule. Yet the AIHTA profile shows that the knowledge base does not treat this as disqualifying. Instead, it treats the resulting index as a quasi-physical quantity, capable of multiplication, aggregation, and ratio comparison.

This is why the dimensional homogeneity problem matters, and the profile makes clear that the knowledge base does not recognize it as binding. "The QALY is a dimensionally homogeneous measure" (FALSE) sits at 0.85 (+1.47). This tells you the institution is comfortable with a quantity that is not a quantity in the measurement-theoretic sense. A homogeneous quantity is one where units are commensurate and the arithmetic respects the structure of the attribute. The QALY is not that. It is "time × preference score," where the preference score is not shown to be ratio and is not derived from a unidimensional measurement model. Even if you granted an interval interpretation—which itself is unsupported without transformation—you would not thereby rescue ratio multiplication. The profile signals that this conceptual barrier is not internalized.

Now bring in the cost-effectiveness machinery. AIHTA has produced work engaging explicitly with thresholds and the ICER logic. That is not a criticism in itself; it simply locates AIHTA within the international HTA mainstream. But your diagnostic asks a sharper question: do cost-effectiveness claims meet the axioms of representational measurement? The statement "claims for cost-effectiveness fail the axioms of representational measurement" is endorsed only weakly at 0.15 (−1.47). That weak endorsement indicates that the knowledge base does not treat measurement invalidity as decisive. Instead, it treats cost-effectiveness ratios as legitimate objects of comparison even when their components are composites. If costs are treated as monetary totals, they are context-dependent and non-invariant; if they are treated as resource units, they may be ratio measures but then require a different evaluative architecture—one based on resource-specific claims, not a single composite cost numerator. Either way, coupling that numerator to a QALY denominator that lacks measurement validity yields a ratio that is, at best, a policy score, not a scientific measure.

This is where falsifiability becomes the dividing line between science and closure. "Non-falsifiable claims should be rejected" sits at 0.60 (+0.34). That moderate endorsement is revealing: the knowledge base gestures toward Popperian language—reject unfalsifiable claims—but it does not allow that principle to bite. The very next item shows why: "reference case simulations generate falsifiable claims" (FALSE) is endorsed at 0.85 (+1.47). This means that within the AIHTA evaluative architecture, simulation outputs are treated as if they were empirically testable claims about the world, rather than functions of assumptions. A model can be recalibrated; it cannot be falsified in the sense that an empirical claim is falsified. When the model output differs from observed outcomes, the standard response is not rejection but adjustment: change a parameter, revise an assumption set, broaden sensitivity analysis. The system absorbs error procedurally rather than learning empirically. This is the hallmark of the HTA memeplex: closure through refinement, not progress through falsification.

The profile also shows how the knowledge base handles the one concept that could anchor latent-trait quantification—possession. "The outcome of interest for latent traits is the possession of that

trait" is endorsed at only 0.25 (−0.93). That is not a floor collapse, but it is clearly not a binding principle. Yet possession is the entire point of measurement for latent constructs. If you cannot specify what it means to "possess" more or less of the attribute—on an invariant scale—then you are not measuring the attribute. You are scoring responses. The AIHTA profile indicates that latent traits remain framed in the language of preferences, utilities, and indices, not in the language of measurement as possession.

Put bluntly, the AIHTA knowledge base appears to satisfy the administrative needs of HTA more than the scientific needs of therapy evaluation. It is structured to deliver a single summary number (the ICER, typically cost-per-QALY) and to close appraisal through threshold discourse, sensitivity analysis, and committee process. That is exactly what international HTA systems reward: comparability, standardization, and decision closure. AIHTA sits within that ecosystem and contributes to it. The problem is that these conveniences are purchased by abandoning the measurement requirements that make quantitative claims empirically meaningful.

What makes this particularly consequential is AIHTA's institutional posture as a scientific assessment body. It is not a lobbying organization; it is not merely a policy secretariat. It is an institute that supplies "support for health care decision-making" through reports and methods. When such an institute operationalizes composite preference scores as if they were measures, it does more than "use common methods." It legitimizes them. It trains analysts. It signals to decision makers that arithmetic on non-measures is acceptable. It thereby reinforces a global epistemic failure: the inversion of measurement and arithmetic.

The duty-of-care implication follows directly. Health systems, clinicians, and patients need claims that are credible, evaluable, and replicable. They need to know what a therapy does in measurable terms: what manifest resource units are changed, what clinical events are prevented, what measurable patient-centered outcomes improve, and—if subjective outcomes are used—whether those outcomes are measured on invariant scales through lawful transformation. A system that relies on simulated cost-per-QALY claims does not deliver that. It delivers an administrative score that may facilitate budgeting, but it does not produce empirically evaluable knowledge about therapy impact. It is therefore not merely scientifically weak; it risks being ethically incoherent. If the decision framework cannot be wrong in the way scientific claims can be wrong—because it lacks measurement referents—then it cannot learn in the way science learns. And if it cannot learn, it cannot meet a duty of care to improve decision quality over time.

This is why the Rasch floor collapses matter so much. They mark the absence of the only route by which subjective observations can become measurement outcomes with invariant units. Without that route, the "quality of life" component of HTA remains a preference-based scoring exercise. Once that scoring is multiplied by time and embedded in simulation models, the entire evaluative apparatus becomes a closed computational system. It can be refined indefinitely without ever confronting the prior question: are we measuring anything at all?

AIHTA's own positioning within Austrian HTA also makes the profile informative. One recent review notes that, within Austria, health economics has historically played a comparatively minor role in decision-maker-commissioned HTA reports, with economic studies representing a small fraction of outputs. That observation can tempt defenders to argue that Austria is therefore less

captured by cost-per-QALY modeling. But the logit profile cuts through that comfort: the issue is not how often economic evaluation is performed; it is what happens when quantification is invoked as the basis for comparative claims. The profile shows that when quantification is invoked, it is invoked within the global HTA memeplex—utilities, QALYs, multiattribute scores, and simulation claims treated as if they were measures.

The conclusion is therefore not that AIHTA is uniquely deficient, but that it is recognizably compliant with the same international belief system that is documented across agencies and journals. It accepts the administrative premise that a single composite index can stand in for effectiveness, and it accepts the methodological premise that simulation can stand in for falsifiable empirical claims. The logit evidence indicates that the axioms of representational measurement do not operate as constraints on admissible claims. That is the decisive indictment.

If AIHTA wished to re-enter the tradition of normal science, the route is conceptually simple even if institutionally disruptive. Claims must be restricted to defined unidimensional attributes. Manifest claims must be expressed in linear ratio units (resource units, event counts, time-to-event, survival). Latent claims must be expressed through Rasch logit ratio scales with invariance demonstrated. Each claim must have a protocol for replication and falsification in defined target populations and meaningful timeframes. Cost-per-QALY closure would be replaced by a portfolio of single-claim evaluations that are empirically contestable. This would not reduce rigor; it would restore it. It would also restore duty of care, because decision making would be anchored in claims that can be proven wrong and therefore improved.

Until such reconstruction occurs, AIHTA will remain a high-functioning node in a global HTA memeplex: administratively coherent, procedurally sophisticated, numerically productive—and epistemically unaccountable. The logit profile does not merely criticize. It diagnoses. It shows the systematic exclusion of the axioms that make quantitative reasoning possible, and it shows the corresponding endorsement of the myths required to keep the QALY and reference-case machinery operational. That is not “a different methodological tradition.” It is an abandonment of measurement, and therefore an abandonment of the only route by which HTA can contribute to the evolution of objective knowledge about therapy impact.

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