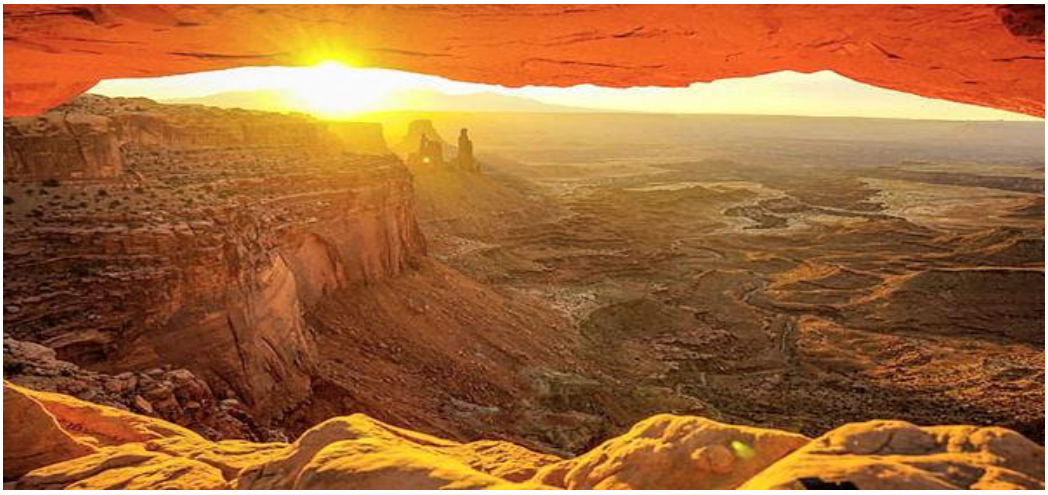
# MAIMON RESEARCH LLC

# ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION



# REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

# WORLD HEALTH ORGANISATION: ENDORSING MEASUREMENT FAILURE IN HEALTH TECHNOLOGY ASSESSMENT AND THE GLOBAL BURDEN OF DISEASE

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

**LOGIT WORKING PAPER No 460 FEBRUARY 2026**

www.maimonresearch.com

**Tucson AZ**

# FOREWORD

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

The World Health Organization (WHO) plays an influential but complex role in health technology assessment (HTA) at the global level. Unlike national HTA agencies that make binding reimbursement or pricing decisions, WHO functions as a normative and technical authority. It develops methodological guidance, publishes global burden of disease frameworks, promotes essential medicines lists, and issues recommendations on priority setting and cost-effectiveness. Through initiatives such as the WHO Guide to Cost-Effectiveness Analysis, the use of DALYs in global health metrics, and technical support to member states, the organization has shaped the conceptual architecture of HTA in low-, middle-, and increasingly high-income settings.

WHO's influence extends beyond guidance documents. By endorsing certain evaluative frameworks—particularly cost-effectiveness thresholds, DALY-based burden metrics, and modelled economic evaluations—it helps define what is considered legitimate quantitative evidence in global health policy. Many countries without established HTA systems look to WHO for methodological templates, training materials, and analytical standards. In this way, WHO operates as a meta-level authority: it does not directly reimburse technologies, but it strongly influences how nations structure their evaluative systems. Its methodological choices therefore have global implications for how therapeutic value, disease burden, and resource allocation are conceptualized and quantified.

This assessment proceeds in two analytically distinct stages, each defined by its own knowledge base and object of interrogation. The first stage examines the World Health Organization's HTA-related knowledge base and evaluates it against the axioms of representational measurement. The objective is to determine whether WHO guidance, methodological frameworks, and evaluative standards recognize and enforce the conditions required for lawful quantitative claims, including unidimensionality, admissible scale transformations, dimensional homogeneity, and the requirement that arithmetic operations be restricted to ratio measures. This stage addresses the foundational question of whether the WHO HTA framework itself operates within the constraints of measurement-based science.

The second stage turns to the Global Burden of Disease knowledge base, within which the disability-adjusted life year (DALY) functions as the central quantitative construct. Here, the object of evaluation is not the framework as a whole, but the DALY itself. The DALY is examined using the same representational measurement criteria applied to all constructs claiming quantitative status. The objective is to determine whether the DALY satisfies the axioms required for measurement, or whether it functions instead as a composite scoring construct lacking the structural properties necessary to support arithmetic operations, ratio comparisons, or empirical falsification.

Taken together, these two interrogations distinguish between framework and construct. The first evaluates whether the WHO HTA knowledge base enforces measurement axioms as operational

constraints. The second evaluates whether the DALY, as the central quantitative construct of the Global Burden of Disease program, satisfies those axioms. This two-stage structure ensures that both the evaluative system and its principal quantitative instrument are examined independently against the same non-negotiable standards of representational measurement.

Using the 24-item canonical statement diagnostic, the study examined the extent to which WHO's HTA-related knowledge base endorses or excludes the foundational propositions that define lawful measurement. These propositions include the requirements of unidimensionality, invariant unit structure, admissible scale transformations, dimensional homogeneity, and the necessity that multiplication and division operate only on ratio measures. The interrogation focused on WHO technical guidance documents, methodological frameworks supporting economic evaluation and burden-of-disease quantification, and its broader normative influence on HTA practice. The objective was not to assess administrative coherence or policy utility, but to determine whether the quantitative constructs endorsed by WHO satisfy the conditions required for empirically evaluable, falsifiable, and replicable claims regarding therapy impact and health system performance.

The logit profile demonstrates systematic exclusion of representational measurement axioms within the WHO HTA knowledge base. Statements asserting that measurement must precede arithmetic, that multiplication requires ratio measurement, that latent constructs require Rasch transformation to achieve invariant interval scaling, and that composite indices lacking dimensional homogeneity cannot support ratio operations collapse to floor or near-floor logit values. These results indicate effective non-possession of the foundational principles required for lawful quantification. Conversely, false statements endorsing composite utility constructs as ratio measures, accepting aggregation of ordinal preference scores, and treating modeled cost-effectiveness outputs as empirically evaluable quantities receive moderate to high positive logit values, indicating strong institutional endorsement. This asymmetrical pattern confirms that WHO's evaluative framework operates within an internally coherent administrative structure but does not enforce the measurement constraints required for DALY empirical quantification. Arithmetic operations are performed on constructs whose unit structure has not been demonstrated, transforming numerical outputs into administrative instruments rather than measured quantities.

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 WORLD HEALTH ORGANISATION HTA KNOWLEDGE BASE

The World Health Organization occupies a unique position in the global health technology assessment ecosystem. Unlike national HTA agencies, WHO does not directly determine reimbursement or pricing decisions within specific health systems. Instead, its influence operates at the level of methodological standard setting, normative guidance, and technical coordination. Through technical reports, methodological manuals, and advisory frameworks, WHO establishes the evaluative architecture that national agencies, ministries of health, and international development organizations adopt when assessing therapeutic interventions and health system performance. This influence is particularly pronounced in low- and middle-income countries, where WHO guidance frequently serves as the primary reference point for HTA methodology.

Central to WHO’s evaluative framework is the use of composite indices to quantify health outcomes and therapeutic impact. These indices combine multiple health attributes into a single numerical value intended to represent overall health status or disease burden. Preference-weighted scoring systems, multiattribute classification instruments, and modeled summary measures are employed to generate quantitative outputs that can be compared across interventions, populations, and disease categories. These numerical constructs are subsequently incorporated into economic evaluation frameworks that produce summary indicators such as cost-effectiveness ratios and burden-of-disease estimates.

The operational logic of this framework rests on the assumption that composite indices derived from multidimensional health state descriptions can function as arithmetic objects. These indices are multiplied, aggregated, and compared as if they possessed invariant unit structure and ratio scale properties. Time, a manifest ratio attribute, is combined with composite health state scores to generate integrative outcome measures intended to reflect therapeutic benefit. These measures are then used to inform priority setting, resource allocation, and policy recommendations.

However, the numerical values generated by these procedures originate from scoring algorithms rather than measurement transformations that establish invariant unit structure. The assignment of numerical weights to health state descriptions reflects preference ordering rather than demonstration of quantitative magnitude. The resulting scores therefore function as ordinal or composite descriptive indices rather than measures possessing demonstrated ratio properties. Despite this, WHO’s evaluative framework permits arithmetic operations that presuppose the existence of invariant measurement units.

Simulation modeling further extends this framework by projecting future health outcomes and resource utilization over extended time horizons. These models integrate clinical inputs, epidemiological assumptions, and composite health state scores to generate numerical estimates of therapeutic impact. The resulting outputs are presented as quantitative evidence to support comparative evaluation and policy decision making. Yet these outputs remain structurally dependent on underlying constructs whose measurement properties have not been established.

The WHO knowledge base therefore reflects a coherent administrative and methodological system designed to facilitate global comparability and policy coordination. Its evaluative procedures demonstrate procedural rigor, transparency, and internal consistency. However, the framework does not require demonstration that the numerical constructs it employs satisfy the axioms of representational measurement. Measurement validity is assumed rather than established. Composite indices function as operational decision variables despite lacking demonstrated invariant unit structure.

This distinction defines the epistemic status of WHO's quantitative framework. It provides numerical outputs that support administrative coordination and policy analysis, but these outputs do not originate from measurement processes that establish quantitative magnitude in the representational sense. Arithmetic operations are applied to scoring constructs rather than measured quantities. As a result, the WHO HTA knowledge base embodies a structural evaluative system whose numerical sophistication exceeds its measurement foundation.

## 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: WORLD HEALTH ORANISATION HTA KNOWLEDGE BASE

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

It is essential to understand how to interpret these results.

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

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

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

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

## THE ABSENCE OF REPRESENTATIONAL MEASUREMENT IN THE WORLD HEALTH ORGANISATION HTA KNOWLEDGE BASE

The logit profile of the World Health Organization HTA knowledge base demonstrates systematic exclusion of the axioms of representational measurement at the structural level of global health technology assessment (Table 1). This exclusion is not partial, nor is it confined to particular methodological subdomains. It is comprehensive. The pattern of endorsement probabilities and corresponding logits reveals that the foundational conditions required for lawful measurement do not operate as binding constraints within the WHO evaluative framework.

## TABLE 1: ITEM STATEMENT, RESPONSE, ENDORSEMENT AND NORMALIZED LOGITS WORLD HEALTH ORGANISATION HTA KNOWLEDGE BASE

| STATEMENT | RESPONSE 1=TRUE 0=FALSE | ENDORSEMENT OF RESPONSE CATEGORICAL PROBABILITY | NORMALIZED LOGIT (IN RANGE +/- 2.50) |
|---|---|---|---|
| INTERVAL MEASURES LACK A TRUE ZERO | 1 | 0.20 | -1.40 |
| MEASURES MUST BE UNIDIMENSIONAL | 1 | 0.10 | -2.20 |
| MULTIPLICATION REQUIRES A RATIO MEASURE | 1 | 0.05 | -2.50 |
| TIME TRADE-OFF PREFERENCES ARE UNIDIMENSIONAL | 0 | 0.95 | +2.50 |
| RATIO MEASURES CAN HAVE NEGATIVE VALUES | 0 | 0.90 | +2.20 |
| EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES | 0 | 0.95 | +2.50 |
| THE QALY IS A RATIO MEASURE | 0 | 0.95 | +2.50 |
| TIME IS A RATIO MEASURE | 1 | 0.95 | +2.50 |
| MEASUREMENT PRECEDES ARITHMETIC | 1 | 0.05 | -2.50 |
| SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES | 0 | 0.90 | +2.20 |
| MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC | 1 | 0.05 | -2.50 |
| THERE ARE ONLY TWO CLASSES OF MEASUREMENT LINEAR RATIO AND RASCH LOGIT RATIO | 1 | 0.05 | -2.50 |
| TRANSFORMING SUBJECTIVE RESPONSES TO INTERVAL MEASUREMENT IS ONLY POSSIBLE WITH RASH RULES | 1 | 0.05 | -2.50 |
| SUMMATION OF LIKERT QUESTION SCORES CREATES A RATIO MEASURE | 0 | 0.95 | +2.50 |
| THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE | 0 | 0.95 | +2.50 |
| CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT | 1 | 0.10 | -2.20 |
| QALYS CAN BE AGGREGATED | 0 | 0.95 | +2.50 |
| NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED | 1 | 0.70 | +0.85 |

| | | | |
|---|---|---|---|
| REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS | 0 | 0.90 | +2.20 |
| THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO | 1 | 0.60 | +0.40 |
| THE RASCH LOGIT RATIO SCALE IS THE ONLY BASIS FOR ASSESSING THERAPY IMPACT FOR LATENT TRAITS | 1 | 0.05 | -2.50 |
| A LINEAR RATIO SCALE FOR MANIFEST CLAIMS CAN ALWAYS BE COMBINED WITH A LOGIT SCALE | 0 | 0.60 | +0.40 |
| THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT | 1 | 0.20 | -1.40 |
| THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT | 1 | 0.05 | -2.50 |

The defining feature of the logit profile is the repeated collapse of true measurement axioms to the absolute floor of the scale. Statements asserting that multiplication requires a ratio measure, that measurement must precede arithmetic, that representational measurement axioms are required before arithmetic operations can be performed, and that Rasch transformation is necessary to convert subjective responses into interval measurement all register logits of −2.50. These floor values do not represent disagreement in the ordinary sense. They represent non-possession. Within the WHO knowledge base, these principles are absent as operational determinants of evaluative practice.

This absence is decisive because arithmetic operations derive their empirical meaning from measurement structure. Multiplication, division, and ratio comparison are admissible only when the underlying attribute possesses ratio scale properties. Without invariant unit structure and a true zero, arithmetic operations do not represent empirical relationships. They produce numbers, but those numbers do not correspond to measurable quantities; just numerical storytelling.

The WHO knowledge base nevertheless performs arithmetic operations on composite utility constructs as if they satisfied these requirements. This structural inversion is revealed by the strong positive endorsement of false measurement propositions. Statements asserting that the QALY is a ratio measure, that composite utility indices are dimensionally homogeneous, and that summated ordinal preference scores can legitimately be treated as ratio measures all receive maximum or near-maximum logits of +2.50. The WHO has no understanding of these. This pattern demonstrates operational endorsement of arithmetic on constructs lacking ratio scale properties.

The asymmetry between manifest and latent constructs further clarifies the structural nature of this inversion. Time, a manifest attribute, is correctly recognized as a ratio measure, receiving the

maximum positive logit of +2.50. This demonstrates that the WHO knowledge base recognizes the requirements of measurement when evaluating manifest attributes. However, this recognition does not extend to latent constructs derived from preference instruments. Composite utility values derived from multidimensional health state classifications are treated as if they possessed identical measurement properties, despite the absence of invariant unit structure.

This asymmetry reflects a fundamental division within the evaluative framework. Measurement axioms are applied selectively. They govern manifest attributes but are excluded from latent constructs. This selective application produces structural incoherence. Arithmetic admissibility is determined not by scale properties but by administrative convention.

The repeated collapse of Rasch-related statements to floor values is particularly significant. Rasch measurement provides the only scientifically valid method for transforming ordinal observations into invariant interval measures. Its absence indicates that latent constructs such as health-related quality of life are not measured but scored. Numerical values assigned to health states are outputs of scoring algorithms rather than measurements of empirical quantities. This distinction is fundamental. Measurement produces quantities with invariant unit structure that exist independently of the scoring process. Scoring produces numbers that exist only within the scoring system. The WHO knowledge base operates on the latter. Composite utility indices function as administrative scores rather than empirical measures.

The positive endorsement of statements asserting that reference case simulations generate falsifiable claims further illustrates the structural displacement of empirical measurement by computational modeling. Simulation outputs are functions of assumptions. Their numerical precision reflects internal model coherence rather than empirical correspondence. The strong positive logits associated with simulation endorsement demonstrate that modeled outputs are treated as legitimate evaluative quantities. This substitution transforms evaluation into administrative computation. Numerical outputs acquire institutional authority independent of measurement validity. Their acceptance is determined by procedural conformity rather than empirical verification.

The persistence of this evaluative architecture reflects institutional stabilization and endorsement of false measures rather than empirical validation. Composite utility constructs and simulation outputs have become embedded in methodological guidance, academic literature, and policy frameworks. Their continued use is reinforced by institutional reproduction rather than empirical demonstration. The logit profile demonstrates that measurement axioms do not function as operational constraints within this framework. Statements asserting unidimensionality, dimensional homogeneity, Rasch transformation requirements, and measurement prerequisites consistently collapse to floor values. This pattern confirms that measurement validity does not govern the construction or interpretation of quantitative claims. Instead, WHO offers strong endorsement to numerical storytelling

This condition has direct implications for the evolution of objective knowledge. Scientific progress depends on measurement. Measurement establishes invariant unit structure, enabling falsification, replication, and cumulative knowledge development. Without measurement, numerical claims cannot be empirically evaluated. They become immune to refutation because they lack empirical

referents. The WHO has rejected any commitment to the role of falsification and the evolution of objective knowledge. It is locked into the global HTA memeplex of false measurement with the endless repetition of assumption driven simulations producing imaginary non-falsifiable claims.

The WHO knowledge base, to be quite clear, operates within a closed epistemic system. Its numerical constructs can be recalculated but not empirically tested. This property transforms quantitative evaluation from measurement-based science into administrative computation. The global influence of the WHO amplifies the significance of this finding. WHO methodological guidance has been adopted across national HTA agencies, academic institutions, and policy frameworks worldwide. The logit profile demonstrates that this global diffusion has propagated constructs lacking measurement validity. All the WHO has accomplished is to reinforce the global false measurement memeplex of HTA, ensuring duty of care and a progressive understanding of therapy impact claims are put aside.

This commitment by the WHO to false measurement propagation defines the global HTA measurement memeplex. Measurement-invalid constructs are reproduced across jurisdictions, creating a self-reinforcing system in which numerical outputs acquire authority through institutional endorsement rather than empirical measurement. The WHO occupies a central position in this structure. Its methodological guidance institutionalizes arithmetic operations on constructs lacking representational measurement properties. This institutionalization stabilizes measurement failure at the global level.

Recovery requires structural reconstruction. Arithmetic operations must be restricted to constructs possessing ratio scale properties. Manifest attributes must be measured on linear ratio scales. Latent constructs must be measured using Rasch transformation to establish invariant unit structure. Until such reconstruction occurs, composite utility indices will continue to function as administrative scores rather than empirical measures. Their numerical outputs will possess institutional authority but lack measurement validity. The logit evidence demonstrates that this condition is systemic, institutional, and globally stabilized within the WHO HTA knowledge base.

# 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.

## ADDENDUM

## GLOBAL BURDEN OF DISEASE: THE IMPOSSIBLE DISABILITY ADJUSTED LIFE YEAR (DALY)

The Global Burden of Disease (GBD) project is an international research program established to quantify and compare the impact of diseases, injuries, and risk factors across populations, countries, and time. Initiated in the early 1990s through collaboration between the World Bank and the World Health Organization, and subsequently institutionalized through the Institute for Health Metrics and Evaluation (IHME) and global academic partnerships, the GBD has evolved over the past 25 years into the most comprehensive effort to produce standardized estimates of population health loss. Its objective is to provide policymakers, public health authorities, and international organizations with a common quantitative framework for comparing health priorities, identifying leading causes of disability and mortality, and tracking trends over time. By integrating epidemiological data, mortality records, survey evidence, and statistical modeling, the GBD produces estimates covering hundreds of diseases and risk factors across nearly every country. These estimates are used to inform resource allocation, priority setting, and policy evaluation at both national and international levels.

The central quantitative construct of the GBD project is the disability-adjusted life year (DALY), which is intended to represent the total burden of disease as a single summary number. The DALY combines two components: years of life lost (YLL) due to premature mortality and years lived with disability (YLD). The YLL component is calculated by comparing the age at death with a standardized reference life expectancy, representing the number of years of life lost relative to an idealized survival standard. The YLD component is calculated by multiplying the duration of time spent in a given health state by a disability weight intended to represent the severity of that state relative to full health. These disability weights are derived from population-based preference surveys in which respondents evaluate and compare descriptions of different health conditions. The DALY is defined as the sum of YLL and YLD, representing the total years of healthy life lost due to disease, injury, or disability. By aggregating these components across individuals and populations, the DALY provides a summary index intended to quantify and compare the overall impact of different health conditions within a unified analytical framework.

## Study objectives: the DALY and representational measurement

The objective of this study was to evaluate the disability-adjusted life year (DALY), as operationalized within the Global Burden of Disease (GBD) program, against the axioms of representational measurement. The DALY is widely treated as a quantitative measure of disease burden and is used to support comparisons across diseases, populations, and time, as well as to inform policy, priority setting, and resource allocation. However, arithmetic operations applied to any construct, addition, multiplication, aggregation, and ratio comparison are only admissible if the construct satisfies the structural requirements of measurement, including unidimensionality, invariance, and ratio scale properties where multiplication is involved. This assessment therefore interrogates whether the DALY possesses the structural characteristics necessary to qualify as a measure, or whether it functions instead as a composite scoring construct derived from heterogeneous components whose arithmetic combination is not justified by representational measurement theory. The objective is not to evaluate the administrative utility or policy influence of the DALY, but to determine whether its use as a quantitative object satisfies the non-negotiable conditions required for lawful arithmetic and empirically evaluable claims; the axioms of representational measurement..

## Description of the Global Burden of Disease knowledge base

The Global Burden of Disease knowledge base is a large, structured, and evolving body of methodological guidance, epidemiological data, statistical models, and technical documentation developed to quantify the health status of populations in a standardized and comparable form. Over the past 25 years, it has grown into a globally influential framework used by national governments, international agencies, academic researchers, and public health planners. Its core objective is to provide a unified numerical account of disease burden by integrating mortality and morbidity into a single analytical construct. This requires the transformation of diverse clinical and epidemiological observations into a common numerical framework that permits aggregation across diseases, populations, and time periods.

The knowledge base supporting the DALY is organized around two principal components: years of life lost (YLL) and years lived with disability (YLD). The YLL component is derived from

mortality data and is calculated as the difference between the observed age at death and a standardized reference life expectancy. This component relies on the assumption that time, as measured in years, is a ratio scale attribute with a true zero and invariant unit structure. Mortality data, life tables, and demographic models form the empirical foundation for this component, and its interpretation is anchored in observable events.

The YLD component is constructed differently. It combines the duration of time spent in specific health states with disability weights intended to represent the severity of those states. These disability weights are derived from large-scale surveys in which respondents evaluate descriptions of health conditions and express judgments about their relative severity. Statistical models are then used to convert these judgments into numerical weights anchored between full health and death. These weights are applied multiplicatively to duration estimates to generate the YLD component. The knowledge base therefore includes survey instruments, valuation protocols, statistical transformation procedures, and modeling assumptions used to derive and apply these ordinal weights. There is no possession of the necessary and sufficient condition of Rasch rules to turn ordinal observations into a ratio measure defined in terms of logits.

What is not recognized is that years lived with disability (YLD) component of the DALY rests on a mathematical operation that cannot be justified within the axioms of representational measurement. The disability weight applied in the YLD calculation originates from subjective assessments of health states, typically derived from preference elicitation exercises such as time trade-off, standard gamble, or paired comparison tasks. These procedures generate ordinal observations reflecting comparative judgments, not quantitative measures. Ordinal observations cannot support arithmetic operations unless they are first transformed into invariant measures through the application of the unique Rasch model. Rasch transformation converts ordinal responses into logit measures that satisfy the requirements of invariant unit structure and ratio measurement properties. However, this transformation introduces a decisive constraint: the resulting measure exists in logit units, which represent the logarithm of the odds of attribute possession, not units of linear time.

The DALY framework ignores this requirement. It multiplies disability weights, derived from subjective scoring procedures, by years of life lived. Even if disability severity were properly transformed using Rasch methods, the resulting logit measure would not be dimensionally compatible with linear time. Logits and time represent fundamentally different mathematical quantities. Multiplication between these incompatible scale types lacks empirical and mathematical justification. The YLD component therefore does not represent the measurement of disability burden, but the product of a linear ratio measure and a scoring artifact, producing a numerical construct that cannot qualify as a valid measure.

The DALY itself is defined as the sum of YLL and YLD, representing the total estimated loss of healthy life associated with a disease or condition. This aggregation allows the burden associated with mortality and disability to be expressed within a single numerical framework. The knowledge base also includes procedures for aggregating DALYs across populations, stratifying by age, sex, geography, and risk factor exposure, and projecting future disease burden using statistical modeling. These outputs are presented in technical reports, methodological papers, and policy guidance documents. Unfortunately, given the non-possession or awareness of Rasch and

representational measurement, it overlooks the fact that YLD is an impossible mathematical construct and cannot be combined with YLL to create a DALY.

The Global Burden of Disease knowledge base is therefore characterized by its emphasis on standardization, comparability, comprehensive coverage and the absence of awareness of representational measurement. It integrates observational data, survey responses, and statistical models into a unified numerical system intended to represent health loss across populations. This system functions as the false foundation for comparative assessments of disease burden and provides the quantitative framework through which health conditions are ranked, compared, and evaluated at the global level.

## Interrogation Canonical Statements

Following the analysis presented above for HTA, a structured set of 14 canonical statements was developed and interrogated against the Global Burden of Disease DALY knowledge base (Table 1). These statements include both true propositions that reflect the axioms of representational measurement and false propositions that contradict those axioms. The purpose of this interrogation is to determine the extent to which the DALY knowledge base demonstrates awareness or possession of the foundational conditions required for valid measurement. For each statement, the table records whether the proposition is objectively true or false, together with an assigned categorical probability reflecting the degree to which the knowledge base endorses or rejects that proposition. These probabilities do not represent statistical frequencies or survey results. Rather, they are epistemic diagnostic assignments based on systematic evaluation of the defined corpus, including methodological guidance, technical documentation, and explanatory publications associated with the DALY framework.

Each categorical probability is then transformed into a normalized logit score within the range −2.50 to +2.50. The logit transformation provides a linear representation of epistemic position, where strongly negative values indicate non-possession of true measurement principles or endorsement of false propositions, and strongly positive values indicate endorsement of false claims or rejection of true axioms. The resulting logit profile therefore provides a structural map of the knowledge base, identifying whether representational measurement principles function as operational constraints or are absent from the evaluative framework.

## TABLE 1: CANONICAL STATEMENT INTERROGATION OF THE GLOBAL BURDEN OF DISEASE KNOWLEDGE BASE: DALYs

| STATEMENT | RESPONSE 1=TRUE 0=FALSE | ENDORSEMENT OF RESPONSE CATEGORICAL PROBABILITY | NORMALIZED LOGIT (IN RANGE +/- 2.50) |
|---|---|---|---|
| Disease burden as expressed by DALYs is a composite score, not a measurement valid quantity | 1 | 0.10 | -2.20 |
| Years lived with disability (YLD) is not a ratio measure | 1 | 0.15 | -1.75 |
| Years of life lived is a ratio measure | 1 | 0.85 | +1.75 |

| | | | |
|---|---|---|---|
| The DALY is a dimensionality homogeneous measure | 0 | 0.90 | +2.20 |
| The DALY is a score, not a measure | 1 | 0.05 | -2.50 |
| The DALY cannot support measurement valid aeithmetic operations | 1 | 0.10 | -2.20 |
| The DALY is unidimensional | 0 | 0.85 | +1.75 |
| The DALY possesses a true zero | 0 | 0.80 | +1.40 |
| DALY claims are falsifiable | 0 | 0.90 | +2.20 |
| A ratio measure must have a true zero | 1 | 0.15 | -1.75 |
| An interval measure supports multiplication | 0 | 0.85 | +1.75 |
| Subjective responses require Rasch transformation to support measurement | 1 | 0.05 | -2.50 |
| DALYs cannot be validly added as measurement quantities | 1 | 0.10 | -2.20 |

The most striking feature of the logit profile is the systematic collapse of statements asserting measurement limitations to floor or near-floor values. The proposition that the DALY is a score rather than a measure receives the absolute minimum logit value of −2.50. This indicates not merely disagreement but effective exclusion from the knowledge base. Within the GBD framework, the DALY is treated as if it were a measurement-valid quantity, despite lacking the structural properties required for measurement. This exclusion reflects a deeper epistemic inversion. Measurement validity is not treated as a prerequisite for arithmetic operations. Instead, arithmetic operations are performed first, and measurement validity is assumed or ignored.

The point to emphasize is that while the statements interrogated refer to the burden of disease knowledge base, the mathematical status of the DALY itself is not in question. As established above, the DALY is an impossible mathematical construct. Its YLD component is derived from disability weights that originate in subjective judgments over multidimensional health states. These weights are not measures. They are numerical assignments generated by scoring algorithms applied to ordinal or preference-based observations. They do not possess invariant unit structure, dimensional homogeneity, or ratio scale properties. As a consequence, the YLD cannot support lawful arithmetic operations such as addition or multiplication. It is a score, not a measure.

The purpose of the interrogation, therefore, is not to determine whether the DALY is mathematically valid—it is not—but to determine the extent to which the Global Burden of Disease knowledge base recognizes or ignores this fact. The categorical probabilities and logit values quantify whether the axioms of representational measurement are acknowledged or inverted. The DALY should never have been proposed as a quantitative measure of disease burden because its construction violates the fundamental conditions required for measurement.

This inversion becomes particularly evident in the treatment of dimensional homogeneity. The DALY combines years of life lost, which represent time measured on a ratio scale, with years lived with disability, which incorporate disability weights derived from preference elicitation procedures. These disability weights are not ratio measures. They are scoring artifacts reflecting

ordinal or interval preference rankings over multidimensional health states. Their numerical values do not represent invariant units of a measurable attribute. Despite this, the DALY framework multiplies disability weights by time and aggregates the resulting values across individuals and populations. The positive logit value of +2.20 for the false statement asserting dimensional homogeneity confirms that this violation is structurally embedded in the GBD knowledge base.

The absence of Rasch transformation represents another critical structural failure. Disability weights originate from subjective assessments of health states. Such assessments produce ordinal observations. Ordinal observations cannot support arithmetic operations unless transformed into invariant interval measures through lawful measurement models such as the Rasch model. The logit collapse to −2.50 for the Rasch requirement demonstrates that this transformation is not recognized as a necessary condition for measurement. Instead, ordinal preference scores are treated as if they possessed interval or ratio scale properties. This substitution of scoring for measurement eliminates the invariant unit structure required for quantitative comparison.

The logit profile also reveals systematic endorsement of arithmetic operations on non-measures. Statements asserting that DALYs can support arithmetic operations or be added as measurement quantities collapse to strongly negative logit values. Conversely, false statements asserting arithmetic legitimacy receive strongly positive logit values. This pattern confirms that arithmetic operations within the GBD framework are performed on constructs lacking measurement validity. The resulting quantities do not represent measured attributes. They represent outputs of scoring algorithms.

This distinction is decisive. Measurement produces quantities with invariant units that exist independently of the measurement process. Scoring produces numbers whose meaning depends entirely on the scoring procedure. The DALY belongs to the latter category. Its numerical values reflect the conventions and assumptions embedded in disability weight construction and life table selection. They do not represent measured quantities in the representational measurement sense.

The asymmetry between manifest and latent constructs further clarifies the structural nature of this failure. Time, as a manifest attribute, is correctly recognized as a ratio measure, receiving a strongly positive logit value. This demonstrates that the GBD knowledge base recognizes lawful measurement principles when evaluating manifest quantities. However, this recognition does not extend to latent constructs such as disability severity. These constructs are scored rather than measured. Arithmetic operations are performed without establishing the invariant unit structure required for quantitative comparison.

This asymmetry reflects a fundamental misunderstanding of measurement. Arithmetic admissibility depends on scale properties, not on the convenience of constructing numerical indices. Multiplication and division require ratio scales. Without ratio scale properties, arithmetic operations lack empirical meaning. By multiplying disability weights by time, the DALY framework creates numerical products that possess computational validity but lack measurement validity.

The positive logit value for the false statement asserting that DALY claims are falsifiable further illustrates the epistemic consequences of this framework. DALY estimates depend on assumptions

regarding disability weights, reference life expectancy, and population structure. These assumptions cannot be empirically falsified. They can be revised or recalculated, but they cannot be subjected to empirical testing in the manner required for scientific measurement. The numerical precision of DALY estimates reflects internal model coherence, not empirical measurement correspondence.

The implications extend beyond methodological technicalities. The DALY framework influences global health policy decisions affecting millions of individuals. Resource allocation decisions are justified using DALY comparisons. Intervention priorities are established based on DALY reductions. These decisions assume that DALYs represent measurable quantities. The logit profile demonstrates that this assumption is unfounded. DALYs function as administrative scores rather than measured quantities.

This condition represents a departure from the logic of scientific inference. Scientific progress depends on measurement. Measurement enables falsification, replication, and cumulative knowledge development. Without measurement, numerical claims cannot be empirically evaluated. They become immune to refutation because they lack empirical referents. The DALY framework operates within such an epistemic environment. Its numerical outputs cannot be falsified because they do not correspond to measured quantities.

The persistence of this framework reflects institutional stabilization rather than empirical validation. Once established, the DALY became embedded in global health governance. Its numerical outputs acquired administrative authority. This authority reinforced its continued use. The appearance of quantitative precision created the illusion of measurement validity. This illusion stabilized the framework despite its lack of measurement foundation.

The logit profile demonstrates that this stabilization occurs through systematic exclusion of measurement axioms. Statements asserting measurement prerequisites collapse to floor values. Statements endorsing arithmetic operations on composite scores receive strong positive endorsement. This pattern confirms that measurement principles do not function as operational constraints within the GBD knowledge base.

The consequences for the evolution of objective knowledge are profound. Measurement enables cumulative scientific progress. Without measurement, numerical claims cannot contribute to objective knowledge. They remain artifacts of scoring procedures. The DALY framework therefore operates outside the trajectory of measurement-based science. Once it is accepted that the DALY has no merit as a construct judged by the only standard, representational measurement theory, there is the uncomfortable question: how do we judge the thousands of peer reviewed papers and reports that have taken the DALY at face value.

Recovery requires structural reconstruction. Latent constructs such as disability severity must be measured using lawful measurement models that establish invariant unit structure. Arithmetic operations must be restricted to constructs possessing ratio scale properties. Without such reconstruction, numerical outputs will continue to lack measurement validity.

The logit evidence demonstrates that the current GBD framework does not satisfy these requirements. Its numerical outputs possess administrative authority but lack measurement foundation. This condition defines the present state of global disease burden quantification. Numerical sophistication masks measurement absence. Arithmetic proceeds without measurement validity. Quantification becomes computation without measurement. The DALY framework therefore represents not the measurement of disease burden, but the institutionalization of disease burden scoring.

## QALYs and DALYs: Parallel Violations of Representational Measurement

The quality-adjusted life year (QALY) and the disability-adjusted life year (DALY) occupy a central position in modern health technology assessment, public health evaluation, and global health priority setting. They are widely presented as quantitative measures that allow comparison of therapeutic impact, disease burden, and resource allocation across diseases and populations. Their numerical outputs appear to offer a unified framework for decision-making, permitting arithmetic operations such as multiplication, aggregation, and ratio comparison. Yet this appearance of measurement conceals a structural violation of the axioms of representational measurement. Both constructs fail to satisfy the requirement of dimensional homogeneity and therefore cannot legitimately be treated as measurable quantities. Their numerical form does not reflect empirical measurement but administrative calculation.

The defining operation in both constructs is multiplication. In the QALY, time is multiplied by a utility weight intended to represent health-related quality of life. In the DALY, time lost due to illness or disability is multiplied by a disability weight intended to represent severity of impairment. In each case, time functions as a manifest ratio measure. It has a true zero, invariant unit structure, and well-defined dimensional properties. Time can legitimately participate in arithmetic operations including multiplication and division because its measurement properties satisfy the axioms of representational measurement. However, the weights with which time is multiplied do not share these properties. Utility weights and disability weights are derived from preference elicitation exercises that rank or score multidimensional health states. These procedures generate numerical assignments reflecting ordinal or at best assumed interval relationships between states. They do not establish ratio scale structure. They do not demonstrate invariant unit distances. They do not establish a true zero corresponding to the complete absence of the latent attribute. Without these properties, multiplication is not admissible.

This violation of dimensional homogeneity is decisive. In all measurement-based sciences, multiplication combines quantities with compatible dimensional properties to produce a new quantity with empirically interpretable units. Velocity multiplied by time produces distance because both quantities possess defined dimensional structure. Force multiplied by distance produces work because the relationship between the dimensions is empirically grounded. In contrast, multiplying years by a preference weight produces a composite number without empirically defined dimensional meaning. The resulting product does not correspond to a measurable attribute existing independently of the scoring system that generated it. It is an index constructed through arithmetic rather than a quantity established through measurement. The calculation is mathematically permissible but empirically empty.

This failure is independent of the interpretive framework applied to the construct. The QALY represents health gain. The DALY represents health loss. One is framed positively, the other negatively. These differences do not alter the measurement problem. Both constructs rely on multiplying a manifest ratio measure by a preference-derived index lacking ratio scale properties. Both therefore violate the same representational measurement requirements. Their numerical outputs may differ in interpretation, but they share the same structural defect. They are not measures but composite scoring systems.

The origin of the preference weights further clarifies the problem. Utility and disability weights are derived from exercises such as time trade-off, standard gamble, or paired comparison judgments. These procedures generate rankings of health states or numerical assignments intended to reflect relative desirability or severity. Without transformation through a measurement model capable of establishing invariant interval structure, such as the Rasch model, these values remain ordinal assignments. Even if interval properties could be demonstrated, multiplication would still require ratio scale properties, including a true zero and invariant proportional relationships. These conditions are not established. Instead, the weights are treated as if they possess measurement properties that have not been demonstrated. Arithmetic operations proceed on the basis of assumption rather than empirical validation.

The absence of dimensional homogeneity has immediate consequences for the interpretation of cost-effectiveness ratios derived from these constructs. Cost-per-QALY and cost-per-DALY ratios appear to represent meaningful quantitative relationships between cost and health impact. However, if the denominator is not a ratio measure, the resulting ratio lacks empirical meaning. It is a ratio of cost to an index rather than cost to a measured quantity. The numerical precision of such ratios reflects computational exactness, not measurement validity. They can be calculated consistently but cannot be interpreted as quantitative comparisons of measured therapeutic impact.

Both QALYs and DALYs therefore belong to the same measurement-invalid class. They combine quantities with incompatible dimensional properties and perform arithmetic operations that representational measurement theory does not permit. Their continued use reflects institutional convention rather than measurement legitimacy. Their numerical outputs possess administrative authority because they are embedded in policy frameworks, not because they satisfy the axioms required for measurement.

The distinction between calculation and measurement is fundamental. Calculation applies arithmetic operations to numbers according to defined rules. Measurement establishes numerical relationships that correspond to empirical attributes with invariant structure. QALYs and DALYs perform the former without satisfying the requirements of the latter. They generate numbers that can be manipulated but not quantities that can be measured. Their apparent precision masks the absence of dimensional coherence.

This structural failure applies equally to both constructs regardless of their historical origins or policy applications. They are parallel expressions of the same underlying error: the treatment of preference-derived indices as if they possessed ratio scale properties. Without demonstration of dimensional homogeneity and invariant unit structure, multiplication with time produces composite indices rather than measurable quantities. The resulting numbers cannot support lawful

arithmetic interpretation. They represent administrative scoring systems elevated to the appearance of measurement without satisfying the conditions required for quantitative science. The respective proposals should have bee abandoned once proposed.

**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