MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

UNITED STATES: INAHTA AND THE ABSENCE OF MEASUREMENT

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

LOGIT WORKING PAPER No 7 JANUARY 2026

www.maimonresearch.com

Tucson AZ

FOREWORD

HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

This Logit Working Paper series documents a finding as extraordinary as it is uncomfortable: health technology assessment (HTA), across nations, agencies, journals, and decades, has developed as a global system of non-measurement. It speaks the language of numbers, models, utilities, QALYs, "value for money," thresholds, discounting, incremental ratios, extrapolations, and simulations. It demands arithmetic at every turn, multiplication, division, summation, aggregation, discounting, yet it never once established that the quantities to which these operations are applied are measurable. HTA has built a vast evaluative machinery on foundations that do not exist. The probabilities and normalized logits in the reports that follow provide the empirical confirmation of this claim. They show, with unsettling consistency, that the global HTA knowledge base neither possesses nor applies the principles of scientific measurement.

The objective of this study is to interrogate the epistemic foundations of the International Network of Agencies for Health Technology Assessment (INAHTA) as a global coordinating body for HTA practice. Rather than treating INAHTA as a neutral facilitator of information exchange, the analysis examines whether the network functions as a conduit for a coherent belief system about measurement, evidence, and quantitative legitimacy. Using the 24-item diagnostic grounded in representational measurement theory, the study evaluates the extent to which the concepts, methods, and evaluative norms propagated through INAHTA member agencies satisfy the axioms required for admissible arithmetic, falsifiable claims, and the cumulative evolution of objective knowledge. The focus is not on the performance of individual agencies in isolation, but on whether INAHTA, as a networked knowledge infrastructure, stabilizes and disseminates measurement-valid standards or, alternatively, institutionalizes forms of false measurement across jurisdictions.

The findings are unambiguous. INAHTA exhibits a pervasive and systematic rejection of the axioms of representational measurement while simultaneously endorsing the arithmetic consequences that those axioms alone can license. Core principles such as the precedence of measurement over arithmetic, the requirement of unidimensionality, the necessity of ratio scales for multiplication and aggregation, and the inadmissibility of composite constructs such as QALYs are weakly endorsed or rejected outright. In contrast, propositions that sustain conventional HTA practice, including the treatment of utilities as ratio measures, the aggregation of QALYs, the legitimacy of summated ordinal scores, and the authority of reference-case simulation models, are strongly endorsed, often at or near the ceiling of the normalized logit scale. Rasch measurement, the only framework capable of converting subjective observations into invariant quantitative measures of latent trait possession, is effectively absent from the INAHTA belief structure. The resulting profile is not one of methodological diversity or cautious pragmatism, but of structural inversion: arithmetic is treated as primary, while measurement is treated as optional. As a consequence, INAHTA functions not as a corrective mechanism for methodological error, but as a powerful amplifier of a global HTA memeplex that normalizes non-falsifiable claims and arrests the evolution of objective knowledge in therapy evaluation.

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

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 INAHTA KNOWLEDGE BASE

The knowledge base of INAHTA is best understood not as a single methodological doctrine, but as a coordinated, distributed infrastructure that stabilizes and disseminates a common evaluative worldview across national and regional HTA agencies. It is constituted by the recurring concepts, analytic conventions, report templates, and methodological expectations that member agencies share, reference, and reinforce through joint working groups, comparative assessments, horizon scanning outputs, and mutual citation practices. What gives this knowledge base its coherence is not formal consensus on measurement theory, but functional alignment around what counts as acceptable evidence in HTA practice.

At its core, the INAHTA knowledge base privileges comparative evaluation framed through cost-effectiveness, budget impact, and modeled projections of long-term outcomes. The dominant numerical artifacts circulating within this ecosystem are utilities, QALYs, ICERs, and reference-case simulation outputs. These are treated as commensurable across jurisdictions, disease areas, and technologies, despite originating from heterogeneous instruments, populations, and preference elicitation methods. The assumption of commensurability is rarely examined; it is embedded structurally through shared reporting formats and cross-agency benchmarking rather than defended as a measurement property.

Patient-reported outcomes occupy an important but methodologically constrained position within the INAHTA knowledge base. Subjective instruments are widely invoked to support claims about quality of life, functioning, or symptom burden, yet they are typically operationalized through summed scores, index values, or preference-weighted composites. The knowledge base treats psychometric reliability, responsiveness, and construct validity as sufficient indicators of quantitative legitimacy, while remaining largely silent on scale type, unidimensionality as a strict measurement requirement, or invariance across populations. As a result, latent attributes are referenced continuously but are not measured in the representational sense; they are scored and then incorporated into downstream arithmetic.

Modeling occupies a central organizing role. Reference-case simulations are treated as the integrative mechanism that reconciles diverse data sources into a single evaluative output. Within the INAHTA knowledge base, sensitivity analysis and scenario testing are routinely presented as indicators of robustness, effectively substituting internal model variation for empirical falsification. This practice reinforces the view that decision relevance arises from coherence within a modeling framework rather than from the existence of measured quantities that can be independently replicated or refuted.

Equally defining are the boundaries of what the knowledge base excludes. Representational measurement theory is not part of the shared conceptual toolkit. Explicit discussion of scale admissibility, the conditions under which arithmetic is meaningful, or the distinction between ordinal ordering and quantitative measurement is largely absent. Rasch measurement, despite its direct relevance to latent trait claims, does not function as a governing standard within INAHTA outputs. When it appears, it does so as a technical option rather than as a non-negotiable requirement for quantifying subjective outcomes.

The INAHTA knowledge base is therefore best characterized as structurally conservative and behaviorally self-reinforcing. It transmits a stable set of evaluative practices across agencies by emphasizing procedural alignment and mutual recognition rather than methodological interrogation. What evolves within this system is not objective knowledge in the strong scientific sense, but the consistency of application of a shared HTA memeplex one in which arithmetic is routine, models are authoritative, and measurement assumptions remain largely unexamined.

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

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: INAHTA

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

It is essential to understand how to interpret these results.

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

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

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

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

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

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	0.25	-1.10
MULTIPLICATION REQUIRES A RATIO MEASURE	1	0.15	-1.75

TIME TRADE-OFF PREFERENCES ARE UNIDIMENSIONAL	0	0.85	+1.75
RATIO MEASURES CAN HAVE NEGATIVE VALUES	0	0.85	+1.75
EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES	0	0.90	+2.20
THE QALY IS A RATIO MEASURE	0	0.90	+2.20
TIME IS A RATIO MEASURE	1	0.95	+2.50
MEASUREMENT PRECEDES ARITHMETIC	1	0.15	-1.75
SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES	0	0.85	+1.75
MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC	1	0.15	-1.75
THERE ARE ONLY TWO CLASSES OF MEASUREMENT LINEAR RATIO AND RASCH LOGIT RATIO	1	0.10	-2.20
TRANSFORMING SUBJECTIVE RESPONSES TO INTERVAL MEASUREMENT IS ONLY POSSIBLE WITH RASH RULES	1	0.10	-2.20
SUMMATION OF LIKERT QUESTION SCORES CREATES A RATIO MEASURE	0	0.90	+2.20
THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE	0	0.80	+1.40
CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.20	-1.40
QALYS CAN BE AGGREGATED	0	0.95	+2.50
NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.60	+0.40
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.85	+1.75
THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.65	+0.60
THE RASCH LOGIT RATIO SCALE IS THE ONLY BASIS FOR ASSESSING THERAPY IMPACT FOR LATENT TRAITS	1	0.10	-2.20
A LINEAR RATIO SCALE FOR MANIFEST CLAIMS CAN ALWAYS BE COMBINED WITH A LOGIT SCALE	0	0.60	+0.40

THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT	1	0.25	-1.10
THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.10	+2.20

INAHTA AND THE GLOBALIZATION OF MEASUREMENT ILLITERACY

The International Network of Agencies for Health Technology Assessment (INAHTA) presents itself as a coordinating body: a facilitator of knowledge exchange, best practice, and methodological learning across national HTA agencies. It claims no single methodological allegiance, positioning itself instead as a neutral convener. Yet the 24-statement assessment reveals something far more consequential. INAHTA is not neutral with respect to measurement. It is neutral only in the sense that it has no concept of measurement to be neutral about.

What emerges from the endorsement profile is not a contested understanding of representational measurement theory, but its near-total absence. INAHTA's documents, guidance, and linked agency outputs behave as if the foundational rules that govern when numbers mean anything at all simply do not exist. This is not methodological pluralism. It is epistemic vacancy.

Across the statements that encode elementary truths of measurement, unidimensionality, scale typology, the requirement that measurement precede arithmetic, endorsement collapses toward the floor. These are not advanced psychometric propositions. They are preconditions for arithmetic taught implicitly at the secondary-school level. Their absence indicates that INAHTA does not merely reject these principles; it does not recognize them as constraints. Numbers are treated as inherently quantitative, and arithmetic is treated as inherently permissible.

This matters because INAHTA's role is not to innovate, but to disseminate. When a national HTA agency lacks measurement competence, the damage is local. When INAHTA lacks it, the damage is global.

The most telling feature of the assessment is INAHTA's implicit endorsement of false propositions that enable routine HTA arithmetic. The belief that EQ-5D preference algorithms generate interval measures is accepted. The belief that QALYs function as ratio measures is accepted. The belief that summated subjective responses can support arithmetic comparison is accepted. None of these propositions is defensible. Their persistence indicates that INAHTA treats preference scoring, utility estimation, and composite indices as measurement by default. There is no requirement to demonstrate scale properties, invariance, or dimensional coherence. Numerical output is sufficient.

This default assumption explains why Rasch measurement is functionally absent. Rasch is not excluded explicitly; it is rendered irrelevant. Statements asserting that Rasch logit ratio scales are the only lawful basis for latent trait impact fall to the bottom of the endorsement distribution. This does not reflect disagreement with Rasch. It reflects the fact that INAHTA does not operate in a conceptual space where “lawful basis” has meaning. Rasch measurement imposes constraints. INAHTA’s ecosystem depends on the absence of constraints.

The handling of Likert-type instruments is particularly revealing. INAHTA-linked assessments routinely summarize patient-reported outcomes using summed or averaged scores and then interpret differences as quantities. The false statement that summation of Likert scores creates a ratio measure receives strong functional endorsement. This is not a technical error; it is mathematical illiteracy. Ordinal category labels do not acquire equal units or a true zero through addition. That this error persists across agencies coordinated by INAHTA demonstrates that the organization does not possess even a rudimentary framework for distinguishing scores from measures.

The same pattern governs INAHTA’s treatment of utilities and QALYs. Utilities are treated as cardinal measures despite lacking unidimensionality, invariance, and a true zero. Negative utilities are tolerated while ratio arithmetic is performed as if those negatives were inconsequential. QALYs are aggregated across individuals, discounted over time, and divided into costs to produce ICERs. Each of these operations violates a different measurement axiom. None is flagged. None is discussed. None is treated as a problem.

This silence is not accidental. INAHTA’s function is to harmonize practice across agencies, not to interrogate foundations. But harmonization without measurement literacy becomes harmonization of error. The 24-statement profile shows that INAHTA does not provide corrective oversight; it provides amplification. It ensures that the same false arithmetic is reproduced consistently across jurisdictions under the banner of “best practice.”

INAHTA’s endorsement of abstract scientific virtues exposes the depth of the contradiction. The statement that non-falsifiable claims should be rejected receives relatively high endorsement. Yet the belief that reference-case simulations generate falsifiable claims persists. This is incoherent. Simulation does not create falsifiability when the outputs cannot be observed and the inputs are not measures. Sensitivity analysis explores internal consistency, not empirical testability. INAHTA’s documents routinely treat modeled lifetime QALYs as if they were claims about the world rather than numerical artifacts of assumption. This reflects not methodological naivety, but a failure to grasp what falsifiability requires.

Crucially, INAHTA does not distinguish between manifest outcomes and latent constructs. Hospital days, survival time, and counts of events, genuine ratio measures, are placed in the same analytic frames as composite scores and preference indices. By presenting them together as commensurable “outcomes,” INAHTA dissolves the categorical boundary between counting and measuring. This is one of the most damaging effects of its coordination role. Agencies are encouraged to believe that all numbers are alike, differing only in units or labels, not in logical status.

The consequence is that INAHTA’s methodological materials do not educate; they normalize. They teach agencies how to apply established tools without ever asking whether the tools are capable of producing measures. The 24-statement assessment makes this explicit. Where measurement axioms should function as gatekeepers, there is indifference. Where Rasch should function as a requirement, there is optionality. Where arithmetic should be constrained, there is assumption.

It is important to be precise about what this means. INAHTA is not “behind the curve.” It is not awaiting the diffusion of new theory. Representational measurement theory predates HTA by decades. Stevens’ scale typology was articulated in 1946. The axioms governing ratio and interval scales are elementary. Rasch measurement has been available since 1960. The failure here is not temporal. It is conceptual.

This is why the endorsement profile looks indistinguishable from that of HTAi, WHO, OECD, and other international bodies. These organizations share a common belief system: arithmetic is legitimate if it is standardized, published, and policy-relevant. Measurement is treated as a technical detail rather than a logical prerequisite. INAHTA’s role is to stabilize that belief system across agencies, not to challenge it.

The absence of any corrective signal is therefore decisive. INAHTA does not identify measurement failure as a problem. It does not warn agencies about the limitations of ordinal scoring. It does not caution against ratio arithmetic on utilities. It does not require explicit justification of scale type. It does not recognize Rasch measurement as a boundary condition for latent trait claims. An organization that does not recognize these distinctions cannot plausibly claim to support evidence-based assessment.

What makes this especially serious is that INAHTA’s coordination function gives its silence normative force. When an agency adopts QALYs or composite PRO scores, it can reasonably infer that such practices are internationally accepted. INAHTA’s materials reinforce that inference. The result is a self-sealing global system in which measurement failure is reproduced not through coercion, but through reassurance.

The 24-statement assessment therefore leads to a stark conclusion. INAHTA does not misunderstand measurement. It does not disagree with representational measurement theory. It operates as if measurement does not exist as a governing concept. Numbers are treated as manipulable tokens rather than representations of attributes. Arithmetic is treated as a procedural entitlement rather than a conditional operation.

In this sense, INAHTA is not merely complicit in the HTA measurement failure. It is one of its principal vectors. By coordinating agencies without measurement literacy, it ensures that false arithmetic is not an isolated defect, but a global norm.

If INAHTA were to take measurement seriously, its entire role would change. It would be required to distinguish scores from measures, to separate manifest outcomes from latent constructs, to prohibit arithmetic where scale properties are absent, and to recognize Rasch measurement as a necessary and sufficient, not optional, condition for latent trait quantification. That would disrupt comparability, slow production, and invalidate a large share of existing HTA outputs. It would also restore scientific coherence.

The fact that none of this occurs is not an oversight. It is the defining feature of INAHTA’s contribution to HTA. The organization exists to harmonize practice, and practice has been built on false measurement. Harmonization, under these conditions, is not progress. It is the globalization of error.

Once this is acknowledged, INAHTA can no longer be defended as a neutral convener. It must be understood as an institutional mechanism through which measurement illiteracy has been stabilized and disseminated worldwide. The 24-statement assessment does not merely critique INAHTA’s methods. It demonstrates that INAHTA has no working concept of measurement at all.

3. NEXT STEPS: TRANSITION TO SINGLE-CLAIM MEASUREMENT

The results of LLM interrogation leave no middle path. The measurement cat is out of the bag, and any system that continues using QALYs, utilities, DALYs, or simulation modelling invites scientific ridicule. INAHTA must recognize and respond to this fact.

DISOWN THE PRESENT BELIEF SYSTEM

The first step toward scientific rehabilitation is an unambiguous renunciation of the non-measurement architecture that has underpinned HTA decision-making for decades. The logic is not rhetorical but structural: if the axioms of representational measurement are violated at the foundation, then no amount of statistical sophistication, modelling embellishment, or "best practice guidelines" can rescue the outputs from incoherence. QALYs, ordinal utilities, DALYs, and reference-case simulations are not merely suboptimal, they are incompatible with any conception of measurement. They lack a legitimate scale type, violate the requirements for meaningful arithmetic, and cannot be integrated into a numerically coherent comparison across interventions. A belief system built on such constructs cannot be amended or partially retained; it must be disowned.

The QALY is the clearest illustration of this impossibility. It is constructed by multiplying ordinal preferences by time, a procedure that lacks dimensional justification and produces outputs that cannot be interpreted as measures of anything. Yet this fiction has persisted because it supplies administrators with a single number, something they can rank, apply a threshold, or negotiate against. The same is true for DALYs, whose lineage in burden-of-disease accounting does nothing to endow them with legitimate measurement properties. Reference-case simulation modelling compounds the error: it takes non-measures as inputs, adds speculation about future clinical and economic pathways, and then outputs a figure that is treated as if it were evidence. The entire apparatus survives only because reviewers, policymakers, and faculty have never been trained in measurement, and thus have lacked the conceptual tools to recognize that these constructs are scientifically impossible.

Disowning the belief system is therefore not an admission of past failure but an unavoidable act of disciplinary self-correction. A field cannot progress while clinging to artefacts that cannot, even in principle, support falsifiable claims. NICE as the exemplar must say so explicitly, not as a symbolic gesture but as the precondition for rebuilding a scientifically credible evaluative architecture.

RECONSTRUCT HTA FROM MEASUREMENT UP

Once the non-measurement framework has been dismantled, reconstruction must begin from the only defensible starting point: measurement theory. There is no shortcut, no incremental reform, and no "middle way" in which QALYs or utilities are patched, modified, or reweighted. The fundamental lesson of representational measurement theory is simple: numbers have meaning only when the empirical structure of the attribute supports a specific scale type. If NICE, assuming it still exists, wants to produce claims that can be evaluated, replicated, and falsified, then it must adopt scale types capable of sustaining the arithmetic it wishes to perform.

For manifest attributes, events that are directly observable, such as hospital days avoided, therapy switching, medication possession, or relapse counts, the appropriate structure is a linear ratio scale. Such scales have a true zero, constant unit intervals, and permit the full suite of permissible arithmetic operations. They allow NICE to make legitimate statements about proportional differences and resource utilization grounded in evidence rather than interpretation. Crucially, ratio scales for manifest outcomes are already ubiquitous in health system data; they require no modelling conjecture and no constructed preferences.

For latent attributes, experiential or subjective constructs such as symptom burden, need-fulfilment, or patient-reported outcomes, the only valid transformation model is the Rasch model. Rasch provides logit-based ratio scales generated through conjoint simultaneous measurement of person ability and item difficulty. Without Rasch, subjective outcomes collapse to ordinal scores that cannot be meaningfully compared or used alongside manifest ratio measures. With Rasch, we acquire disease specific instruments that satisfy unidimensionality, invariance, and interval structure, enabling legitimate claims about latent change.

Reconstruction means reinstating the basic rule that every claim must have the appropriate measurement architecture. This is not an aesthetic preference but the necessary foundation for a science of evaluation. HTA becomes coherent only when claims rest on instruments that conform to the axioms of measurement, not on the administrative desire for a "single number." The transition is radical only because the prior framework ignored measurement entirely.

MOVE TO PROTOCOL-BASED SINGLE CLAIMS

A measurement-valid HTA system cannot rely on summary constructs or composite evaluations. It must instead adopt a single-claim architecture in which each value claim stands alone, meeting the requirements of falsifiability, replication, and transparent reporting. This follows directly from the logic of science: a claim must be empirically testable, reproducible in the same target population, and supported by an agreed protocol that specifies exactly how evidence will be generated. Multi-outcome cost-effectiveness analysis cannot meet these standards because it integrates non-measures into speculative models and converts them into an imaginary "value for money" figure that cannot be falsified. Single claims, by contrast, are grounded in measurement.

Each claim begins with a precisely defined target population, typically patients initiated on a therapy within a defined window. This eliminates the ambiguity inherent in modelling lifetime populations or hypothetical cohorts. The endpoint must be measurement-valid; a linear ratio measure for manifest attributes or a Rasch logit ratio measure for latent ones. The protocol must articulate the evidence generation plan prospectively: how data will be collected, over what timeframe, using what analytic criteria, and under what conditions replication will be evaluated.

A single-claim architecture aligns HTA with the logic of clinical science. Claims are constructed in advance, not retrospectively assembled from model outputs. They are specific, narrow, and auditable. They permit comparability across therapies because each claim is defined in measurement terms rather than through the aggregation of unrelated dimensions. Importantly, single claims also eliminate the bureaucratic temptation to collapse multiple endpoints into an artificial summary. Instead, each outcome is assessed on its own merits, with its own ruler.

This shift does more than improve methodological defensibility; it transforms the institutional culture of evaluation. NICE, again as the exemplar, would no longer operate as a quasi-modelling agency but as a measurement-based adjudicator of empirically testable propositions. The result is a transparent, reproducible, and scientifically legitimate HTA system.

ADOPT THE MAIMON RESEARCH DISTANCE EDUCATION PROGRAM

Reconstruction requires education, and at present there is no conventional textbook, curriculum, or HTA training program that teaches measurement theory, Rasch, and protocol-based single-claim architecture in a scientifically coherent manner. The existing academic infrastructure remains trapped in the old belief system, recycling utilities, QALYs, and reference-case models as if these constructs were measures. Replacing that architecture therefore requires retraining. systematic, structured, and accessible to agencies, universities, and policy staff. The Maimon Research Distance Education Program is currently the only platform that provides this.

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/

The program builds HTA from measurement upward. It teaches representational measurement theory as the foundation for any evaluative claim. It trains participants in Rasch modelling, including item calibration, person–item maps, logit transformations, and the construction of valid, unidimensional latent-trait measures. It provides protocol templates that define how claims are constructed, evaluated, and replicated. It supplies checklists to ensure scale-type coherence, target population definition, and the exclusion of non-measures. It also addresses the institutional, pedagogical, and administrative barriers that have historically prevented HTA from adopting measurement standards.

Most importantly, the program replaces the HTA belief system with a scientific one. It does not attempt to “improve” QALYs or “modernize” utilities. It demonstrates why those constructs are impossible and shows how to build a new system from first principles that produces claims that can be defended in court, in peer review, and in public policy. The program equips faculty and decision-makers with the conceptual tools they were never given, tools that allow them to recognize the difference between a measure and a number masquerading as one. Adopting the program is therefore not supplementary; it is the enabling step. Without a trained workforce, we cannot transition to single-claim measurement.

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