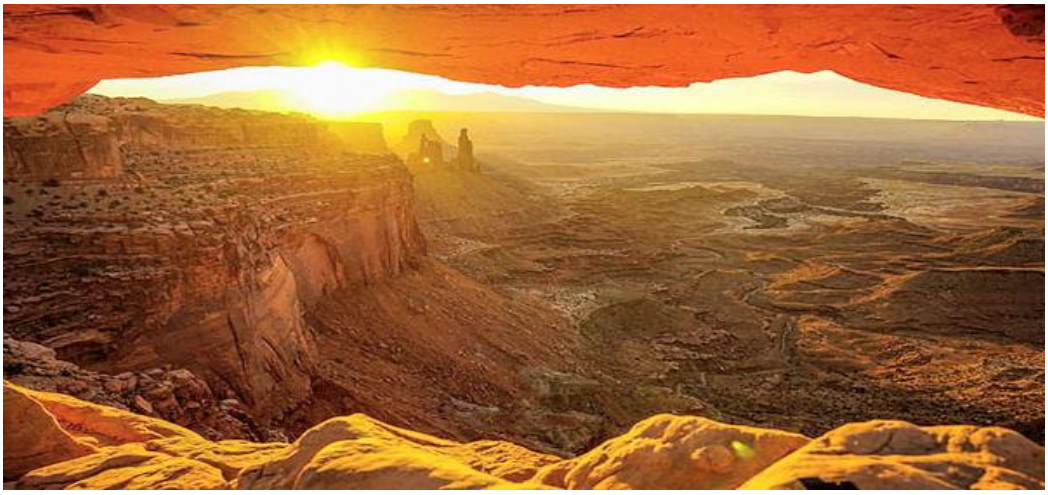
MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION



REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

UNITED STATES: RTI INTERNATIONAL AND THE INSTITUTIONAL ENDORSEMENT OF MEASUREMENT FAILURE AND NUMERICAL STORYTELLING IN HEALTH TECHNOLOGY ASSESSMENT

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

LOGIT WORKING PAPER No 150 FEBRUARY 2026

www.maimonresearch.com

Tucson AZ

FOREWORD

HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

RTI International (formerly Research Triangle Institute) is a large, US-based, nonprofit research organization that occupies a distinctive position at the intersection of applied science, public policy, and health technology assessment–adjacent activity. Founded in 1958 and headquartered in Research Triangle Park, North Carolina, RTI was established to translate academic research into practical solutions for government and industry. Over time, it has evolved into a globally active institution conducting policy-relevant research across health, education, economics, environment, and international development. Its scale, reputation, and sustained engagement with health policy make it a consequential contributor to the HTA knowledge environment, even where it does not function as a formal HTA agency.

From an HTA perspective, RTI's relevance arises from two overlapping roles. First, it operates as a producer of policy-facing health research for US federal agencies, state governments, and international bodies. Second, through its commercial arm, RTI Health Solutions, it provides health economics and outcomes research (HEOR) and market-access support to biopharmaceutical and medical device manufacturers. Together, these activities place RTI squarely within the epistemic ecosystem that generates, legitimizes, and operationalizes quantitative claims about health outcomes, value, and comparative effectiveness.

RTI has played a significant role in evidence synthesis, comparative effectiveness research, and large-scale health data analysis. It has been a long-standing contractor to US government agencies such as the Agency for Healthcare Research and Quality (AHRQ), contributing to systematic reviews, evidence-based practice center outputs, and methodological work that informs clinical and coverage decisions. These activities overlap directly with HTA functions, particularly in their reliance on structured evidence appraisal, outcome hierarchies, and quantitative synthesis.

In parallel, RTI Health Solutions functions as a global HEOR consultancy serving life-science companies. Its work includes cost-effectiveness analysis, utility elicitation, real-world evidence generation, patient-reported outcome research, and support for reimbursement submissions across multiple jurisdictions. In this role, RTI applies standard HTA conventions—QALYs, preference-based measures, modeling, and simulation—to support value claims intended for payers and decision makers. These activities are not peripheral; they are central to the contemporary practice of HTA as it is operationalized in submissions and value dossiers.

Crucially, RTI's influence is epistemic rather than regulatory. It does not make coverage decisions, but it helps define what counts as acceptable evidence and admissible quantitative reasoning in both public and private HTA contexts. Its methodological publications, contract research outputs, and consultancy practices contribute to the normalization of prevailing HTA constructs, including

utilities, QALYs, aggregation, and model-based inference. As such, RTI functions as a conduit through which HTA conventions are transmitted, reinforced, and professionalized.

For the purposes of a canonical 24-item assessment, RTI International is therefore best understood as an institutional knowledge base rather than a decision authority. It represents a mature, influential embodiment of mainstream HTA and HEOR practice within the United States and internationally. Interrogating RTI's knowledge base offers a way to assess whether a leading scientific research institute that prides itself on rigor and policy relevance nonetheless reproduces the same measurement assumptions and failures that characterize HTA more broadly.

The objective of this study is to interrogate the health technology assessment knowledge base associated with RTI International using the canonical 24-item diagnostic grounded in representational measurement theory and Rasch measurement principles. The purpose is not to evaluate project quality, analytic competence, or contractual performance, but to determine whether the axioms required for scientific measurement function as admissibility conditions within RTI's HTA-relevant activities. Specifically, the study examines whether unidimensionality, scale-type integrity, invariance, and the logical priority of measurement over arithmetic are articulated and enforced, or whether numerical legitimacy instead derives from methodological convention and professional consensus within HEOR and HTA practice.

The findings are unambiguous. RTI's HTA knowledge base structurally endorses measurement failure. Foundational measurement axioms are absent as governing constraints, while false measurement propositions associated with utilities, QALYs, aggregation, and simulation modeling are strongly reinforced. This pattern does not reflect oversight, transitional reform, or internal debate. It reflects a stable epistemic configuration in which arithmetic is routinely applied to quantities whose measurement status is neither established nor interrogated. RTI does not merely participate in this configuration; it professionalizes it. The result is an institutional knowledge base that produces and legitimizes quantitative claims without satisfying the conditions required for those claims to qualify as measures.

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.

. 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 REGIONAL KNOWLEDGE BASE FOR HTA WITH RTI INTERNATIONAL

RTI International occupies a distinctive and influential position within the contemporary HTA ecosystem. As a large, nonprofit research institute with a substantial global footprint, RTI operates simultaneously as a producer of public-sector evidence and as a provider of commercial HEOR and market-access services through its subsidiary, RTI Health Solutions. This dual role places RTI at the center of the epistemic environment that generates, refines, and legitimizes quantitative claims about health outcomes, value, and comparative effectiveness.

From an HTA perspective, RTI's public-sector work includes systematic evidence synthesis, comparative effectiveness research, methodological development, and policy-relevant analysis conducted for government agencies and international bodies. These activities contribute directly to the evidentiary substrate upon which clinical guidelines, coverage policies, and payer decisions are based. RTI's commercial activities, in parallel, involve the generation of utilities, QALYs, cost-effectiveness models, real-world evidence strategies, and patient-reported outcome programs designed to support reimbursement and access decisions across multiple jurisdictions. Together, these activities define an institutional knowledge base that is deeply embedded in mainstream HTA practice.

Within this knowledge base, numerical legitimacy is conferred by methodological acceptability rather than by measurement admissibility. Quantitative constructs are treated as valid because they are generated by recognized instruments, algorithms, or models, and because they align with prevailing HTA conventions. Measurement is assumed, not demonstrated. The axioms that would determine whether numbers legitimately represent empirical attributes—unidimensionality, scale-type constraints, invariance, and additivity—do not function as gatekeeping criteria.

This is particularly evident in the treatment of subjective and latent constructs. Health-related quality of life, preferences, and other latent attributes are routinely quantified using standardized instruments and preference-based scoring systems. These scores are then summed, averaged, multiplied, and aggregated as inputs to QALY calculations and economic models. No requirement exists within RTI's HTA knowledge base that ordinal responses be transformed using Rasch measurement or any other method capable of establishing lawful interval or ratio properties. Algorithmic scoring substitutes for measurement theory, and arithmetic proceeds accordingly.

The QALY occupies a central and unchallenged position within this framework. Utilities are treated as if they were measured on a ratio scale, QALYs are treated as dimensionally homogeneous quantities, and aggregation across persons and time is treated as permissible arithmetic. These practices are not defended as measurement claims; they are normalized as

professional standards. The absence of measurement admissibility criteria allows these assumptions to persist without scrutiny.

Economic evaluation and simulation modeling further reinforce this structure. Cost-effectiveness analyses and reference-case models are presented as generating decision-relevant quantitative claims, even though their outputs depend on assumptions and mappings that cannot be independently tested. While uncertainty analysis and sensitivity testing are routinely performed, these procedures operate entirely within a framework that presumes the legitimacy of the underlying measures. Falsifiability is invoked rhetorically, but it does not function as a binding scientific standard.

Importantly, RTI's knowledge base exhibits high levels of statistical and methodological competence. Concepts such as regression, modeling, and odds ratios are well understood and correctly applied. However, statistical literacy does not substitute for measurement validity. Knowing how to compute a statistic does not confer permission to apply arithmetic to non-measures. The canonical assessment demonstrates that RTI's HTA practice possesses the former while lacking the latter.

As an institution widely regarded for rigor and policy relevance, RTI plays a critical role in stabilizing HTA conventions. Its publications, contract research outputs, and consultancy services transmit a quantitative grammar in which utilities, QALYs, and model outputs are treated as legitimate measures by default. In doing so, RTI functions not as a marginal participant, but as a high-credibility conduit through which measurement failure is normalized and reproduced.

In sum, the RTI International HTA knowledge base exemplifies the core finding of the Logit Working Papers series: measurement failure in HTA is not the product of incompetence or neglect. It is the result of an institutionalized refusal to treat representational measurement axioms as admissibility conditions. Where those axioms are absent, arithmetic persists, but measurement does not.

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: RTI INTERNATIONAL

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.

RESULTS AND DISCUSSION

The canonical interrogation of the HTA knowledge base associated with RTI International yields a pattern that is internally coherent, methodologically orthodox, and diagnostically decisive. The probability–logit profile does not suggest confusion, inconsistency, or partial transition toward measurement reform. Instead, it reveals a stable epistemic configuration in which foundational axioms of scientific measurement are absent as admissibility conditions, while their negation is systematically normalized through professional practice. RTI's HTA footprint therefore exemplifies not an idiosyncratic failure, but the professionalization of false measurement within contemporary health technology assessment.

TABLE 1: ITEM STATEMENT, RESPONSE, ENDORSEMENT AND NORMALIZED LOGITS RTI INTERNATIONAL

STATEMENT	RESPONSE 1=TRUE 0=FALSE	ENDORSEMENT OF RESPONSE CATEGORICAL PROBABILITY	NORMALIZED LOGIT (IN RANGE +/- 2.50)
INTERVAL MEASURES LACK A TRUE ZERO	1	0.30	-0.85
MEASURES MUST BE UNIDIMENSIONAL	1	0.15	-1.75
MULTIPLICATION REQUIRES A RATIO MEASURE	1	0.15	-1.75
TIME TRADE-OFF PREFERENCES ARE UNIDIMENSIONAL	0	0.85	+1.75
RATIO MEASURES CAN HAVE NEGATIVE VALUES	0	0.90	+2.20
EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES	0	0.85	+1.75
THE QALY IS A RATIO MEASURE	0	0.90	+2.20
TIME IS A RATIO MEASURE	1	0.85	+1.75
MEASUREMENT PRECEDES ARITHMETIC	1	0.05	-2.50
SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES	0	0.90	+2.50
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	0.20	-1.40
SUMMATION OF LIKERT QUESTION SCORES CREATES A RATIO MEASURE	0	0.90	+2.20
THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE	0	0.85	+1.75
CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.10	-2.20
QALYS CAN BE AGGREGATED	0	0.90	+2.20

NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.15	-0.62
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.85	+1.75
THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.70	+0.85
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	0.80	+1.40
THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT	1	0.10	-2.20
THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.05	-2.50

RTI occupies a distinctive position within the HTA ecosystem. It is not a regulatory authority, nor does it issue binding coverage or reimbursement decisions. Rather, it functions as a high-credibility producer of policy-facing evidence and a major provider of health economics and outcomes research services to life-science manufacturers. Through its public-sector work in evidence synthesis and comparative effectiveness research, and through its commercial HEOR and market-access activities, RTI contributes directly to the generation and legitimization of quantitative claims about health outcomes, value, and comparative effectiveness. As such, it represents an institutional knowledge base whose epistemic commitments warrant scrutiny.

The foundational scale axioms provide the first diagnostic signal. The proposition that interval measures lack a true zero receives only weak reinforcement. This indicates that, while scale typology may be understood at a definitional level, it does not function as a binding constraint on arithmetic within RTI's HTA practice. Interval–ratio distinctions are not operationalized as admissibility rules; they are treated as descriptive background knowledge. Where such distinctions fail to constrain arithmetic, the door is opened to multiplicative and aggregative operations that measurement theory does not permit.

The requirement that measures be unidimensional collapses further. Unidimensionality is the defining condition for measurement, not an optional refinement. Its weak reinforcement indicates that RTI's HTA knowledge base does not require demonstration of dimensional coherence prior to index construction, aggregation, or ratio comparison. Multidimensional health attributes are routinely collapsed into single numerical scores, not because their dimensional structure has been resolved, but because established HTA conventions treat such compression as acceptable.

This inversion of scientific order becomes explicit in the collapse of the proposition that measurement precedes arithmetic. With a normalized logit at the floor of the scale, the RTI HTA knowledge base does not treat measurement as a prerequisite for calculation. Arithmetic is authorized by methodological convention, institutional precedent, and professional consensus, rather than by meeting representational measurement axioms. The identical collapse of the proposition that arithmetic requires compliance with those axioms confirms that measurement theory does not operate as a governing framework.

The ontological implications of this absence are decisive. The proposition that only two admissible classes of measurement exist, linear ratio and Rasch logit ratio, also collapses to the floor. Ordinal, interval, and composite constructs are treated as interchangeable inputs to arithmetic so long as they are generated by accepted instruments or algorithms. Measurement ontology is replaced by methodological permissiveness.

The consequences are most evident in the treatment of latent constructs. All Rasch-related propositions exhibit near-total non-possession. The HTA knowledge base does not recognize Rasch rules as necessary for transforming subjective responses into interval measurement, does not recognize the Rasch logit ratio scale as the only admissible basis for assessing latent-trait impact, and does not frame possession of a latent trait as the outcome of interest. Latent attributes such as health-related quality of life are discussed as if they were measurable, yet no mechanism is invoked to establish that measurement has occurred.

Against this backdrop, the strong reinforcement of false measurement propositions is structurally revealing. Time trade-off preferences are treated as unidimensional; ratio measures are treated as capable of taking negative values; preference-based algorithms are treated as conferring interval properties; and QALYs are treated as ratio measures that can be aggregated and compared multiplicatively. These endorsements are not theoretical claims advanced in isolation; they are embedded in routine analytic workflows. Endorsement here is behavioral: the knowledge base behaves as if these propositions were true because its arithmetic depends on them.

The QALY settlement occupies a central position in this profile. The strong endorsement of the QALY as a ratio and dimensionally homogeneous measure indicates that RTI's HTA practice treats QALY-based outputs as legitimate quantities suitable for ratio comparison and aggregation. This occurs despite the absence of any requirement that utilities be measured on a ratio scale or that heterogeneous health attributes be rendered dimensionally coherent. Aggregation follows as a matter of routine arithmetic, insulated from measurement critique by institutional normalization.

Claims for cost-effectiveness are correspondingly insulated. The proposition that such claims fail the axioms of representational measurement collapses to the minimum probability. Cost-effectiveness outputs are treated as epistemically legitimate products of accepted methodology, not as arithmetic claims contingent on measurement validity. The failure is not that cost-effectiveness is debated and rejected; it is that the admissibility question is never posed.

The treatment of falsifiability reinforces this conclusion. While there is nominal reinforcement of the principle that non-falsifiable claims should be rejected, this is overwhelmed by strong endorsement of the proposition that reference-case simulations generate falsifiable claims.

Simulation outputs are treated as if they were empirically testable, despite their dependence on structural assumptions, preference mappings, and extrapolations that cannot be independently verified. Falsifiability functions rhetorically rather than operationally.

Statistical competence does not correct this failure. The relatively strong recognition of the definition of the logit reflects quantitative literacy, but statistical literacy is not measurement validity. Knowing how to compute a statistic does not confer permission to apply arithmetic to non-measures. RTI's profile therefore demonstrates a critical distinction: HTA's measurement failure is not caused by ignorance of mathematics, but by the institutional refusal to apply measurement axioms as admissibility constraints.

Taken together, the probability–logit profile establishes RTI International as a paradigmatic institutional conduit for mainstream HTA arithmetic. Its HTA knowledge base does not challenge the prevailing settlement; it exemplifies it. The combination of methodological sophistication, professional credibility, and unexamined measurement assumptions makes the failure more consequential, not less.

The canonical assessment leaves no middle position available. Either RTI International accepts that the axioms of representational measurement are binding conditions for quantitative claims, or it must concede that the HTA arithmetic it produces and supports does not constitute measurement in the scientific sense. Appeals to methodological rigor, consensus practice, or professional expertise cannot resolve this contradiction. Until measurement admissibility is made explicit and enforced, RTI's HTA outputs, however carefully produced, remain instances of numerical storytelling rather than empirically evaluable claims. The issue is not one of improvement or refinement; it is one of admissibility. No institution, however reputable, is exempt from that requirement.

HOW DOES RTI TRANSITION FROM MEASUREMENT FAILURE TO REPRESENTATIONAL MEASUREMENT

A transition by RTI International from its current position within HTA to one grounded in representational measurement would require a fundamental reorientation of how quantitative claims are authorized, not a refinement of existing methods. The central change would be epistemic rather than technical: measurement admissibility would have to become an explicit and binding condition for all arithmetic claims. RTI would need to state, publicly and operationally, that no numerical result may be advanced unless the underlying attribute satisfies the axioms of representational measurement appropriate to the mathematical operations performed. This rule would not function as a methodological preference or a reporting guideline; it would function as a gatekeeper. Scale type would have to be declared prior to analysis, admissible arithmetic would have to follow from that declaration, and violations would no longer be framed as limitations but as disallowed claims. Without such a rule, any claimed transition would be cosmetic.

Once admissibility is made explicit, RTI would need to restructure its HTA knowledge base around a strict separation between manifest and latent claims. Manifest attributes, such as counts, durations, and resource use, would be treated exclusively as linear ratio measures, permitting lawful arithmetic including addition, multiplication, and ratio comparisons. Latent constructs, such

as quality of life, symptom burden, or patient need, would no longer be treated as quasi-quantities derived from ordinal instruments. They would be admissible only if measured on Rasch logit ratio scales, with outcomes defined in terms of possession of the latent trait rather than numerical scores. The commingling of manifest and latent quantities within a single arithmetic framework would cease, because such commingling has no foundation in measurement theory.

This structural separation would immediately force the abandonment of the QALY as an admissible claim. RTI would not need to deny the historical importance of the QALY or suppress discussion of its use in policy contexts. What would have to end is the production of QALY-based claims as if they represented measurable quantities. Multiplying utilities by time, aggregating across individuals, and comparing ratios of cost per QALY would no longer be permissible, because the utility component fails the requirements of ratio measurement. The QALY could remain as a descriptive artifact within the history of HTA, but it would lose its status as an evidentiary endpoint.

Such a shift would require a corresponding transformation of RTI's patient-reported outcome and preference research. Rather than eliciting utilities for insertion into economic models, RTI would need to redirect its psychometric expertise toward the construction of Rasch-compliant instruments for specific latent constructs. This would involve explicit testing of unidimensionality, item hierarchy, and invariance across populations, and the reporting of results as logit ratio measures rather than summated scores. In this framework, subjective responses would no longer be treated as numerical surrogates for latent attributes; they would be treated as observations requiring lawful transformation before arithmetic could occur. Patient-reported outcomes would become measures in the scientific sense, not inputs to invalid arithmetic.

Economic evaluation and modeling would also require reclassification. RTI could continue to develop models, but their epistemic status would have to change. Models would be explicitly designated as exploratory and hypothesis-generating tools, not as generators of empirically evaluable claims. Outputs would no longer be presented as falsifiable evidence, nor would they be used to justify thresholds or pricing decisions. This reclassification would eliminate the current contradiction in which simulation outputs are rhetorically treated as testable claims while depending on assumptions that cannot be independently verified. Modeling would regain a legitimate role, but only as a source of structured conjecture rather than numerical authority.

At the level of value claims, RTI would need to abandon composite constructions in favor of protocol-driven single claims. Each claim would specify a target population, a single outcome with a valid measurement structure, a defined timeframe, and a falsifiable empirical test. The aggregation of disparate attributes under the heading of "value" would cease, replaced by a portfolio of discrete, evaluable claims. This would align RTI's HTA work with the logic of normal science, where claims stand or fall individually rather than being shielded within composite indices.

None of these changes could be implemented quietly. For the transition to be credible, RTI would need to publish a formal measurement charter that specifies which scale types are admissible, which arithmetic operations are permitted on each, and which forms of HTA claims the institution will no longer produce. Such a statement would carry immediate consequences. Certain revenue

streams tied to conventional HEOR and market-access work would be reduced or eliminated. At the same time, RTI would differentiate itself sharply from the mainstream HTA industry, establishing itself as an institution willing to accept the costs of scientific admissibility.

The decisive point is that there is no incremental path between representational measurement and the current HTA settlement. One either treats measurement axioms as binding conditions or one does not. If RTI were to make this transition, it would no longer function as a conventional HTA or HEOR provider. It would become the first large, professionally credible institution to reject numerical storytelling in favor of lawful measurement. That transition would be difficult, disruptive, and costly. It would also resolve, once and for all, the contradiction at the heart of contemporary HTA: the claim to scientific authority without the discipline of measurement.

HOW DOES RTI INTERNATIONAL EXPLAIN A 40 YEAR LEGACY OF FALSE MEASUREMENT IN HEALTH TECHNOLOGY ASSESSMENT

With an academic and professional staff of roughly **6,000**, RTI International cannot plausibly explain a forty-year legacy of false measurement by appealing to "the system," "the field," or inherited conventions without conceding a failure of due diligence. At that scale, ignorance is not an explanation; it is an indictment.

The argument that "no one knew" about Stevens (1946), scale typology, or the axioms of representational measurement simply does not withstand scrutiny. Stevens' paper is not obscure. It is one of the most cited methodological articles in the behavioral and social sciences, taught for decades across psychology, education, economics, statistics, and measurement theory. The distinction between nominal, ordinal, interval, and ratio scales is introductory material in countless graduate programs. For an institution that markets itself on scientific rigor, policy relevance, and methodological excellence, the claim that such foundational material was unknown is not credible.

The more accurate explanation is not ignorance, but institutional abdication of epistemic responsibility**.** RTI did not need every analyst to be a measurement theorist. What it needed , and failed to require, was institutional due diligence before producing and selling quantitative claims. At no point did RTI appear to ask the prior and logically necessary question: *Do the quantities we are manipulating satisfy the axioms required for the arithmetic we apply to them?* That question does not require specialist expertise. It requires intellectual honesty and a willingness to treat measurement as an admissibility condition rather than a decorative afterthought.

The failure, therefore, is not that HTA conventions existed. It is that RTI accepted those conventions **uncritically**, embedded them in contracts, operationalized them in workflows, and monetized them through HEOR services without independent verification of their scientific legitimacy. When RTI produced QALYs, cost-effectiveness ratios, aggregated utilities, or model outputs, it did so under its own institutional imprimatur. At that moment, responsibility did not lie with "the field" or "the client." It lay with RTI.

Blaming the system also collapses under the weight of RTI's own self-presentation. RTI does not describe itself as a passive implementer of guidelines. It presents itself as a thought leader, a methodological authority, and a trusted scientific partner. That status carries obligations. Chief

among them is the obligation to refuse to advance claims that cannot be defended under the axioms of measurement. The fact that such refusals did not occur is not an accident; it reflects a deliberate choice to privilege market compatibility over epistemic validity.

Nor is it persuasive to argue that responsibility diffuses across a large organization. Institutions exist precisely to concentrate responsibility through governance, review processes, and quality assurance. If six thousand staff members operated for decades without anyone elevating measurement admissibility as a blocking issue, that is not a distributed failure — it is a **systemic institutional failure**. It means measurement theory was excluded from the organization's definition of rigor.

The most uncomfortable implication is this: RTI's legacy of false measurement cannot be explained without acknowledging that due diligence was never performed. Not once was the QALY subjected to a formal admissibility review under representational measurement axioms. Not once were ordinal utilities treated as a potential barrier to arithmetic rather than as an inconvenience to be smoothed over by convention. Not once was a client told that a requested analysis could not be produced because the outcome was not measurable. That silence is not neutrality. It is endorsement.

The issue, then, is not whether RTI was *aware* of Stevens (1946). With its staffing, disciplinary breadth, and academic connections, awareness is unavoidable. The issue is that RTI chose institutionally and repeatedly not to act on what that awareness logically entailed. That choice made numerical storytelling professionally acceptable and commercially viable. It also made false measurement routine. At this point, RTI cannot explain its legacy by appealing to history or field norms. The only coherent explanation is that representational measurement was never treated as a non-negotiable scientific constraint. Until that is acknowledged explicitly, any claim of reform will look like reputational management rather than epistemic reckoning.

IS HEALTH TECHNOLOGY ASSESSMENT A DECISION SCIENCE OR A MEASUREMENT SCIENCE?

Health technology assessment has spent more than four decades oscillating between two identities without ever resolving the tension between them. On the one hand, HTA presents itself as a decision science: a pragmatic enterprise concerned with informing policy choices under uncertainty, balancing costs and benefits, and supporting resource allocation. On the other hand, HTA repeatedly claims to quantify outcomes, measure value, and compare technologies using numerical metrics that purport to represent empirical magnitudes. These two identities are not merely different emphases; they rest on fundamentally incompatible epistemic commitments. HTA cannot coherently be both unless it satisfies the requirements of measurement science. It has not done so.

A decision science does not require measurement in the strict sense. It may rely on preferences, rankings, deliberation, heuristics, or negotiated trade-offs. Its outputs are recommendations, not truths. A decision science can operate with ordinal information, subjective judgments, and context-specific criteria, because its legitimacy derives from procedural transparency rather than empirical lawfulness. If HTA were honestly framed as a decision science, it could acknowledge that its

numbers are aids to deliberation rather than measures of real-world attributes. Cost-effectiveness ratios, thresholds, and composite indices would be treated as structured opinions, not as quantities with intrinsic meaning.

A measurement science, by contrast, makes a much stronger claim. It asserts that numbers represent attributes of the world in a lawful way, such that arithmetic operations preserve empirical relations. Measurement is not about convenience or consensus; it is about admissibility. Addition, multiplication, ratios, and aggregation are permitted only when the axioms governing the underlying attribute are satisfied. This is not a philosophical preference. It is the condition that distinguishes measurement from numerology. Any field that claims to measure outcomes, quantify benefit, or compute efficiency ratios is asserting that it meets these conditions.

HTA routinely behaves as a measurement science while disclaiming responsibility for the axioms that measurement requires. Utilities are treated as if they were interval or ratio measures. Quality-adjusted life-years are treated as if they were dimensionally homogeneous quantities capable of multiplication and aggregation. Composite endpoints are treated as if they represented single attributes. Model outputs are treated as if they were falsifiable empirical claims. None of these practices is compatible with representational measurement theory, yet HTA advances them as if numerical form alone conferred legitimacy.

When challenged, HTA retreats to the language of decision science. It is said that HTA is "about informing decisions, not measuring truth," that models are "tools, not predictions," and that numbers are "decision aids." This retreat is revealing. It acknowledges, implicitly, that HTA's quantities cannot withstand the standards of measurement science. But HTA does not follow this acknowledgment to its logical conclusion. It continues to present its outputs as if they were measures, to apply arithmetic as if it were lawful, and to enforce thresholds and ratios as if they had objective meaning.

This unresolved ambiguity is not harmless. Decision science and measurement science have different standards of accountability. A decision framework can tolerate plural values and contextual variation because it does not claim to measure reality. A measurement claim cannot. It is either valid or it is not. By refusing to choose, HTA shields itself from both forms of accountability. When criticized for invalid measurement, it invokes decision pragmatism. When challenged on decision legitimacy, it invokes quantitative authority.

The consequence is numerical storytelling: numbers that look like measures, behave like measures, and are treated as measures, but are defended as merely advisory when their foundations are questioned. This is not a defensible epistemic position. It is a category error institutionalized.

HTA must therefore decide what it is. If it is a decision science, it must abandon claims to measurement, stop performing illegitimate arithmetic, and present its outputs as structured judgments rather than quantified facts. If it is a measurement science, it must accept representational measurement axioms as binding constraints and refuse to advance claims that do not satisfy them. There is no third option that preserves scientific credibility.

For forty years, HTA has avoided this choice. The result is not methodological pluralism, but epistemic incoherence. Until HTA declares whether it is measuring the world or merely advising decisions, and aligns its practices accordingly, its numbers will remain neither scientifically valid measures nor honest decision tools—only artifacts of a discipline unwilling to confront its own foundations.

III. THE TRANSITION TO MEASUREMENT IN HEALTH TECHNOLOGY ASSESSMENT

THE IMPERATIVE OF CHANGE

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

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

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

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

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

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

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

MEANINGFUL THERAPY IMPACT CLAIMS

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

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

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

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

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

THE PATH TO MEANINGFUL MEASUREMENT

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

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

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

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

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

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

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

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

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

TRANSITION REQUIRES TRAINING

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

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

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

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

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

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

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

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

A NEW START IN MEASUREMENT FOR HEALTH TECHNOLOGY ASSESSMENT

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

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

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

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

DESIGNED FOR CLOSURE

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

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

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

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

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

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

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

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

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

ACKNOWLEDGEMENT

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

REFERENCES

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

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

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

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