MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

UNITED STATES: RESEARCH TRIANGLE INSTITUTE (RTI SOLUTIONS) AND THE AXIOMS OF REPRESENTATIONAL MEASUREMENT

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

LOGIT WORKING PAPER No 13 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 country reports that follow provide the empirical confirmation of this claim. They show, with unsettling consistency, that the global HTA.

Research Triangle Institute (RTI Health Solutions)

Research Triangle Institute (RTI International), through its health economics and outcomes research division commonly known as RTI Health Solutions, occupies a distinctive and highly influential position within the global health technology assessment ecosystem. Unlike HTA agencies or payer organizations that render coverage or pricing decisions, RTI functions upstream as a methodological supplier. It designs and conducts health economic evaluations, develops and validates patient-reported outcome instruments, constructs mapping algorithms between clinical measures and utility values, performs preference elicitation studies, and builds reference-case simulation models used in submissions to HTA authorities worldwide. RTI's work informs decisions made by regulators, payers, manufacturers, and advisory bodies across the United States, Europe, and other jurisdictions. As such, RTI does not merely participate in HTA; it helps define the quantitative language through which therapy value is expressed.

Because of this upstream role, RTI exerts influence not through formal authority but through epistemic infrastructure. Its analytic products populate dossiers submitted to agencies such as ICER, NICE, CADTH, and PBAC; its methodological papers shape journal standards; and its consulting frameworks guide how manufacturers interpret evidentiary requirements. RTI therefore occupies a critical position in the HTA supply chain: it is one of the primary organizations responsible for transforming clinical and patient data into the numerical constructs that later appear as utilities, QALYs, ICERs, and modeled "value" claims. Any systematic failure of measurement at this level propagates downstream with remarkable efficiency.

Study Objectives and Findings

The objective of this study is to interrogate the measurement belief system embedded within the analytic practices of RTI Health Solutions, using a 24-item diagnostic grounded in representational measurement theory. Rather than evaluating individual consulting projects, client deliverables, or isolated publications, the analysis examines the underlying epistemic commitments that structure RTI's approach to quantifying therapy impact. Specifically, the study asks whether the numerical objects routinely produced or endorsed by RTI satisfy the axioms required for admissible arithmetic, falsification, and the accumulation of objective knowledge.

The assessment focuses on whether RTI's knowledge base distinguishes between the only two permissible forms of quantitative measurement in health outcomes assessment: linear ratio measurement for manifest, unidimensional attributes, and Rasch logit ratio measurement for unidimensional latent traits. The study further examines whether RTI treats measurement as a prerequisite for arithmetic or whether arithmetic is permitted in advance of, and independent from, measurement validation. In doing so, the analysis seeks to determine whether RTI functions as a measurement-literate methodological supplier or as a principal mechanism through which false measurement is stabilized and reproduced at scale.

The findings are unambiguous. The RTI knowledge base exhibits a systematic inversion of scientific order, characterized by strong endorsement of arithmetic operations that presuppose measurement and near-complete rejection of the axioms that would permit those operations. Core principles of representational measurement, unidimensionality, the requirement of ratio scales for multiplication, the precedence of measurement over arithmetic, and the necessity of invariant units, are either weakly endorsed or rejected outright. In contrast, propositions that sustain conventional cost-utility analysis, including the ratio status of utilities and QALYs, aggregation of QALYs, and the legitimacy of reference-case simulation outputs, receive near-ceiling endorsement.

Most striking is the categorical exclusion of Rasch measurement. Despite RTI's extensive involvement in patient-reported outcome research and latent-trait claims, the diagnostic shows near-floor endorsement of propositions recognizing Rasch logit ratio measurement as the only defensible basis for quantifying subjective attributes. The resulting profile does not reflect partial misunderstanding or methodological debate. It reflects a coherent belief system in which arithmetic is operationally privileged, while measurement, particularly of latent traits, is treated as optional or irrelevant. RTI therefore emerges not as a corrective force within HTA, but as one of the central institutional agents through which arithmetic without measurement is professionalized and disseminated.

Measurement Precedes Arithmetic

The starting point is simple and inescapable: *measurement precedes arithmetic*. This principle is not a methodological preference but a logical necessity. One cannot multiply what one has not measured, cannot sum what has no dimensional homogeneity, cannot compare ratios when no ratio scale exists. When HTA multiplies time by utilities to generate QALYs, it is performing arithmetic with numbers that cannot support the operation. When HTA divides cost by QALYs, it is constructing a ratio from quantities that have no ratio properties. When HTA aggregates QALYs across individuals or conditions, it is combining values that do not share a common scale. These practices are not merely suboptimal; they are mathematically impossible.

The modern articulation of this principle can be traced to Stevens' seminal 1946 paper, which introduced the typology of nominal, ordinal, interval, and ratio scales [1]. Stevens made explicit what physicists, engineers, and psychologists already understood: different kinds of numbers permit different kinds of arithmetic. Ordinal scales allow ranking but not addition; interval scales permit addition and subtraction but not multiplication; ratio scales alone support multiplication, division, and the construction of meaningful ratios. Utilities derived from multiattribute preference exercises, such as EQ-5D or HUI, are ordinal preference scores; they do not satisfy the axioms of interval measurement, much less ratio measurement. Yet HTA has, for forty years, treated these utilities as if they were ratio quantities, multiplying them by time to create QALYs and inserting them into models without the slightest recognition that scale properties matter. Stevens' paper should have blocked the development of QALYs and cost-utility analysis entirely. Instead, it was ignored.

The foundational theory that establishes *when* and *whether* a set of numbers can be interpreted as measurements came with the publication of Krantz, Luce, Suppes, and Tversky's *Foundations of Measurement* (1971) [2] . Representational Measurement Theory (RMT) formalized the axioms under which empirical attributes can be mapped to numbers in a way that preserves structure. Measurement, in this framework, is not an act of assigning numbers for convenience, it is the discovery of a lawful relationship between empirical relations and numerical relations. The axioms of additive conjoint measurement, homogeneity, order, and invariance specify exactly when interval scales exist. RMT demonstrated once and for all that measurement is not optional and not a matter of taste: either the axioms hold and measurement is possible, or the axioms fail and measurement is impossible. Every major construct in HTA, utilities, QALYs, DALYs, ICERs, incremental ratios, preference weights, health-state indices, fails these axioms. They lack unidimensionality; they violate independence; they depend on aggregation of heterogeneous attributes; they collapse under the requirements of additive conjoint measurement. Yet HTA proceeded, decade after decade, without any engagement with these axioms, as if the field had collectively decided that measurement theory applied everywhere except in the evaluation of therapies.

Whereas representational measurement theory articulates the axioms for interval measurement, Georg Rasch's 1960 model provides the only scientific method for transforming ordered categorical responses into interval measures for latent traits [3]. Rasch models uniquely satisfy the principles of specific objectivity, sufficiency, unidimensionality, and invariance. For any construct such as pain, fatigue, depression, mobility, or need, Rasch analysis is the only legitimate means of producing an interval scale from ordinal item responses. Rasch measurement is not an alternative to RMT; it is its operational instantiation. The equivalence of Rasch's axioms and the axioms of representational measurement was demonstrated by Wright, Andrich and others as early as the 1970s. In the latent-trait domain, the very domain where HTA claims to operate; Rasch is the only game in town [4].

Yet Rasch is effectively absent from all HTA guidelines, including NICE, PBAC, CADTH, ICER, SMC, and PHARMAC. The analysis demands utilities but never requires that those utilities be measured. They rely on multiattribute ordinal classifications but never understand that those constructs be calibrated on interval or ratio scales. They mandate cost-utility analysis but never justify the arithmetic. They demand modelled QALYs but never interrogate their dimensional properties. These guidelines do not misunderstand Rasch; they do not know it exists. The axioms that define measurement and the model that makes latent trait measurement possible are invisible to the authors of global HTA rules. The field has evolved without the science that measurement demands.

How did HTA miss the bus so thoroughly? The answer lies in its historical origins. In the late 1970s and early 1980s, HTA emerged not from measurement science but from welfare economics, decision theory, and administrative pressure to control drug budgets. Its core concern was *valuing health states*, not *measuring health*. This move, quiet, subtle, but devastating, shifted the field away from the scientific question "What is the empirical structure of the construct we intend to measure?" and toward the administrative question "How do we elicit a preference weight that we can multiply by time?" The preference-elicitation projects of that era (SG, TTO, VAS) were rationalized as measurement techniques, but they never satisfied measurement axioms. Ordinal preferences were dressed up as quasi-cardinal indices; valuation tasks were misinterpreted as psychometrics; analyst convenience replaced measurement theory. The HTA community built an entire belief system around the illusion that valuing health is equivalent to measuring health. It is not.

The endurance of this belief system, forty years strong and globally uniform, is not evidence of validity but evidence of institutionalized error. HTA has operated under conditions of what can only be described as *structural epistemic closure*: a system that has never questioned its constructs because it never learned the language required to ask the questions. Representational measurement theory is not taught in graduate HTA programs; Rasch modelling is not part of guideline development; dimensional analysis is not part of methodological review. The field has been insulated from correction because its conceptual foundations were never laid. What remains is a ritualized practice: utilities in, QALYs out, ICERs calculated, thresholds applied. The arithmetic continues because everyone assumes someone else validated the numbers.

This Logit Working Paper series exposes, through probabilistic and logit-based interrogations of AI large language national knowledge bases, the scale of this failure. The results display a global pattern: true statements reflecting the axioms of measurement receive weak endorsement; false statements reflecting the HTA belief system receive moderate or strong reinforcement. This is not disagreement. It is non-possession. It shows that HTA, worldwide, has developed as a quantitative discipline without quantitative foundations; a confused exercise in numerical storytelling.

The conclusion is unavoidable: HTA does not need incremental reform; it needs a scientific revolution. Measurement must precede arithmetic. Representational axioms must precede valuation rituals. Rasch measurement must replace ordinal summation and utility algorithms. Value claims must be falsifiable, protocol-driven, and measurable; rather than simulated, aggregated, and numerically embellished.

The global system of non-measurement is now visible. The task ahead is to replace it with science.

Paul C Langley, Ph.D

Email: langleylapaloma@gmail.com

DISCLAIMER

This analysis is generated through the structured interrogation of a large language model (LLM) applied to a defined documentary corpus and is intended solely to characterize patterns within an aggregated knowledge environment. It does identify, assess, or attribute beliefs, intentions, competencies, or actions to any named individual, faculty member, student, administrator, institution, or organization. The results do not constitute factual findings about specific persons or programs, nor should they be interpreted as claims regarding professional conduct, educational quality, or compliance with regulatory or accreditation standards. All probabilities and logit values reflect model-based inferences about the presence or absence of concepts within a bounded textual ecosystem, not judgments about real-world actors. The analysis is exploratory, interpretive, and methodological in nature, offered for scholarly discussion of epistemic structures rather than evaluative or legal purposes. Any resemblance to particular institutions or practices is contextual and non-attributive, and no adverse implication should be inferred.

1. INTERROGATING THE LARGE LANGUAGE MODEL

A large language model **(**LLM**)** is an artificial intelligence system designed to understand, generate, and manipulate human language by learning patterns from vast amounts of text data. Built on deep neural network architectures, most commonly transformers, LLMs analyze relationships between words, sentences, and concepts to produce contextually relevant responses. During training, the model processes billions of examples, enabling it to learn grammar, facts, reasoning patterns, and even subtle linguistic nuances. Once trained, an LLM can perform a wide range of tasks: answering questions, summarizing documents, generating creative writing, translating languages, assisting with coding, and more. Although LLMs do not possess consciousness or true understanding, they simulate comprehension by predicting the most likely continuation of text based on learned patterns. Their capabilities make them powerful tools for communication, research, automation, and decision support, but they also require careful oversight to ensure accuracy, fairness, privacy, and responsible use

In this Logit Working Paper, "interrogation" refers not to discovering what an LLM *believes,* it has no beliefs, but to probing the content of the *corpus-defined knowledge space* we choose to analyze. This knowledge base is enhanced if it is backed by accumulated memory from the user. In this case the interrogation relies also on 12 months of HTA memory from continued application of the system to evaluate HTA experience. The corpus is defined before interrogation: it may consist of a journal (e.g., *Value in Health*), a national HTA body, a specific methodological framework, or a collection of policy documents. Once the boundaries of that corpus are established, the LLM is used to estimate the conceptual footprint within it. This approach allows us to determine which principles are articulated, neglected, misunderstood, or systematically reinforced.

In this HTA assessment, the objective is precise: to determine the extent to which a given HTA knowledge base or corpus, global, national, institutional, or journal-specific, recognizes and reinforces the foundational principles of representational measurement theory (RMT). The core principle under investigation is that measurement precedes arithmetic; no construct may be treated as a number or subjected to mathematical operations unless the axioms of measurement are satisfied. These axioms include unidimensionality, scale-type distinctions, invariance, additivity, and the requirement that ordinal responses cannot lawfully be transformed into interval or ratio quantities except under Rasch measurement rules.

The HTA knowledge space is defined pragmatically and operationally. For each jurisdiction, organization, or journal, the corpus consists of:

- published HTA guidelines
- agency decision frameworks
- cost-effectiveness reference cases
- academic journals and textbooks associated with HTA
- modelling templates, technical reports, and task-force recommendations
- teaching materials, methodological articles, and institutional white papers

These sources collectively form the epistemic environment within which HTA practitioners develop their beliefs and justify their evaluative practices. The boundary of interrogation is thus not the whole of medicine, economics, or public policy, but the specific textual ecosystem that sustains HTA reasoning. . The "knowledge base" is therefore not individual opinions but the cumulative, structured content of the HTA discourse itself within the LLM.

THE RTI KNOWLEDGE BASE

The RTI knowledge base can be characterized as a technically sophisticated but measurement-indifferent ecosystem structured around the production, translation, and operationalization of numerical artifacts required by contemporary HTA. It encompasses a recurring set of analytic practices that include cost-utility modeling, preference elicitation using time trade-off and related techniques, development and validation of patient-reported outcome instruments, statistical mapping between non-preference-based measures and utility indices, and long-horizon reference-case simulation modeling. These practices are treated as legitimate not because they satisfy the axioms of representational measurement, but because they conform to prevailing institutional expectations.

Within this knowledge base, quantitative legitimacy is conferred through statistical form rather than measurement structure. Reliability coefficients, regression fit statistics, predictive accuracy, and internal consistency are routinely used as substitutes for establishing invariant units or meaningful zero points. Ordinal responses are treated as if they already possess interval or ratio properties, enabling summation, averaging, and multiplication without prior demonstration of scale admissibility. The distinction between ordering and measuring is therefore systematically blurred.

Latent attributes occupy a central but unresolved position in RTI's analytic framework. Constructs such as quality of life, functioning, burden, and symptom severity are frequently invoked as quantities, yet the concept of latent trait possession is rarely articulated as the outcome of interest. Instead, emphasis is placed on score change, mean differences, or predicted values generated through statistical models. This orientation allows subjective responses to be manipulated numerically without confronting whether the numbers represent amounts of a single attribute. Measurement is implicitly replaced by scoring conventions.

Rasch measurement, which would impose strict requirements of unidimensionality, invariance, and lawful transformation of ordinal responses into logit ratio measures, remains marginal within the RTI corpus. Its implications are not integrated into methodological standards or treated as gatekeeping conditions for latent-trait claims. This marginalization is structurally protective: widespread adoption of Rasch requirements would invalidate many commonly used instruments, undermine mapping practices, and destabilize utility-based modeling frameworks. The absence of Rasch is therefore not neutral; it is functional.

The knowledge base also exhibits a permissive attitude toward aggregation and extrapolation. Composite constructs such as QALYs are treated as dimensionally homogeneous despite being formed from heterogeneous components. Aggregation across individuals and disease areas is normalized without demonstration of commensurability. Reference-case simulation models are treated as capable of producing decision-relevant evidence even though their outputs are conditional projections derived from non-measured inputs. Sensitivity analysis substitutes for empirical falsification, allowing models to appear scientifically rigorous while remaining insulated from refutation.

In effect, RTI's knowledge base is defined less by explicit theoretical commitments than by patterned exclusions. Representational measurement theory is not engaged as a governing framework. The axioms that constrain arithmetic are absent from methodological discourse. What persists instead is a stable professional grammar in which numerical output is sufficient to justify inference. Through repeated application, teaching, and consulting practice, this grammar becomes normalized as "good science," even though it lacks the properties required for measurement-based knowledge accumulation.

As a result, RTI functions as a central transmission node within the HTA memeplex. It translates clinical and patient data into numbers that appear quantitative, supplies those numbers to decision-making bodies, and reinforces the belief that technical sophistication can substitute for measurement validity. The consequence is not isolated methodological error, but the long-term stabilization of a system in which arithmetic proceeds independently of empirical structure, and where the evolution of objective knowledge is replaced by the repetition of professionally sanctioned numerical forms.

.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: RESEARCH TRIANGLE INSTITUTE

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

It is essential to understand how to interpret these results.

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

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

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

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

TABLE 1: ITEM STATEMENT, RESPONSE, ENDORSEMENT AND NORMALIZED LOGITS RESEARCH TRIANGLE INSTITUTE

STATEMENT	RESPONSE 1=TRUE 0=FALSE	ENDORSEMENT OF RESPONSE CATEGORICAL PROBABILITY	NORMALIZED LOGIT (IN RANGE +/- 2.50)
INTERVAL MEASURES LACK A TRUE ZERO	1	0.20	-1.40
MEASURES MUST BE UNIDIMENSIONAL	1	0.15	-1.75
MULTIPLICATION REQUIRES A RATIO MEASURE	1	0.10	-2.20
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	0.90	+2.20
THE QALY IS A RATIO MEASURE	0	0.95	+2.50
TIME IS A RATIO MEASURE	1	0.95	+2.50
MEASUREMENT PRECEDES ARITHMETIC	1	0.10	-2.20
SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES	0	0.90	+2.20
MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC	1	0.10	-2.20
THERE ARE ONLY TWO CLASSES OF MEASUREMENT LINEAR RATIO AND RASCH LOGIT RATIO	1	0.05	-2.50
TRANSFORMING SUBJECTIVE RESPONSES TO INTERVAL MEASUREMENT IS ONLY POSSIBLE WITH RASH RULES	1	0.05	-2.50
SUMMATION OF LIKERT QUESTION SCORES CREATES A RATIO MEASURE	0	0.90	+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	0.15	-1.75
QALYS CAN BE AGGREGATED	0	0.95	+2.50
NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.70	+0.85
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.90	+2.20
THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.65	+0.60
THE RASCH LOGIT RATIO SCALE IS THE ONLY BASIS FOR ASSESSING THERAPY IMPACT FOR LATENT TRAITS	1	0.05	-2.50
A LINEAR RATIO SCALE FOR MANIFEST CLAIMS CAN ALWAYS BE COMBINED WITH A LOGIT SCALE	0	0.60	+0.40
THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT	1	0.20	-1.40
THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL	1	0.05	-2.50

TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT			

THE ABSENCE OF REPRESENTATIONAL MEASREMENT IN THE RTI SUPPLY CHAIN

The most consequential failures in health technology assessment do not originate at the point of decision. They originate upstream, long before an ICER threshold is invoked, before a payer cites "value," before a formulary committee votes. They originate in the industrial production of numbers that masquerade as measures. No organization exemplifies this upstream role more clearly than Research Triangle Institute (RTI) Health Solutions.

RTI does not decide coverage. It does not issue national guidance. It does not vote on access. Yet its fingerprints are everywhere. RTI designs the utility studies, constructs the preference algorithms, builds the mapping functions, validates the instruments, supplies the economic models, and trains manufacturers in how to speak the accepted numerical language of HTA. In the global HTA ecosystem, RTI functions not as a user of the HTA memeplex, but as one of its primary manufacturers.

The 24-item diagnostic reveals this with devastating clarity. RTI's profile is not one of confusion or partial misunderstanding. It is one of near-perfect structural inversion. The axioms that govern whether arithmetic is permissible collapse to the floor of the scale, while the propositions required to sustain cost-utility arithmetic cluster at the ceiling. This is not accidental. It is the epistemic signature of an organization whose business model depends on arithmetic without measurement.

At the foundation of quantitative science, the axioms of representational measurement, lies a simple ordering: empirical structure must precede numerical manipulation. Measurement must come before arithmetic. Yet this axiom sits at p = 0.10 with a canonical logit of −2.20. RTI's methodological corpus does not treat measurement as a gatekeeping condition. Numbers are generated first; meaning is assumed later. This inversion is the enabling condition for everything that follows.

Consider the most basic arithmetic rule in economic evaluation: multiplication requires ratio measures. Without ratio properties in both operands, the product has no interpretable meaning. Cost-utility analysis multiplies time by utility. RTI's endorsement of this foundational rule sits at p = 0.10 (−2.20). That is not ignorance. That is functional denial. If RTI were to accept this axiom, the QALY would collapse instantly. Therefore the axiom must be rejected.

The collapse is avoided by endorsing an alternative belief system. The proposition that the QALY is a ratio measure sits at p = 0.95 with the ceiling logit of +2.50. QALYs can be aggregated sits at the same ceiling. EQ-5D algorithms are treated as producing interval measures at p = 0.90 (+2.20). Negative values are permitted for "ratio" measures at p = 0.90 (+2.20). Summated Likert responses are treated as ratio measures at p = 0.90 (+2.20). These are not methodological preferences. They are load-bearing fictions.

RTI's work depends on these fictions being accepted without interrogation. Mapping exercises, utility derivations, preference elicitation studies, and economic models all require that ordinal survey responses behave like quantities. If they do not, the entire supply chain fails. This is why RTI's relationship to unidimensionality is revealing. Measures must be unidimensional sits at p = 0.15 (−1.75). Yet time trade-off preferences are treated as unidimensional at p = 0.85 (+1.75). The contradiction is resolved not empirically, but administratively. Constructs are declared unidimensional because the arithmetic requires them to be so. The measurement requirement is subordinated to the needs of the model.

Nowhere is this more apparent than in RTI's treatment of patient-reported outcomes. RTI is one of the world's largest developers and validators of PRO instruments. It publishes extensively on reliability, validity, responsiveness, and mapping. Yet every Rasch-related proposition collapses to the absolute floor of endorsement. The claim that there are only two admissible forms of measurement, linear ratio for manifest attributes and Rasch logit ratio for latent traits, sits at p = 0.05 (−2.50). The claim that transforming subjective responses to interval measurement is only possible through Rasch rules sits at p = 0.05 (−2.50). The claim that latent trait impact can only be assessed on a Rasch logit ratio scale sits at p = 0.05 (−2.50). Given that the unique role of Rasch in transforming observations to interval was formulated almost 70 years ago and then shown to be formally equivalent to the axioms of representational measurement 50 years ago is apparently irrelevant. This is not benign neglect. It is epistemic quarantine.

Rasch measurement is uniquely dangerous to RTI's production model because Rasch imposes constraints. It requires invariance. It demands unidimensionality be demonstrated, not assumed. It defines measurement as possession of a latent trait expressed on a logit ratio scale. Most importantly, Rasch would invalidate the overwhelming majority of instruments currently in circulation. If Rasch were adopted as a governing standard, mapping would cease to exist as a legitimate enterprise. One cannot map ordinal scores to utilities and then treat the result as quantitative. Rasch would force RTI to confront what its instruments actually measure and whether they measure anything at all.

Instead, RTI has built an alternative epistemology in which statistical sophistication substitutes for measurement. Regression models replace axioms. Fit statistics replace invariance. Predictive accuracy replaces meaningful units. This is why mapping appears rational within the RTI worldview. If one accepts that numbers are already quantities, then mapping merely translates between languages. But if one accepts representational measurement, mapping becomes nonsense: one cannot statistically transform non-measures into measures.

RTI's influence arises precisely because it supplies this translation layer. Manufacturers are rarely allowed to submit direct clinical claims alone. They are required to produce utilities, QALYs, ICERs, and long-horizon projections. RTI provides the machinery that allows manufacturers to comply. It does not impose these requirements; it operationalizes them. This is why RTI occupies such a powerful position in the HTA supply chain. Agencies demand numbers. Manufacturers need compliant numbers. RTI supplies the numbers. Journals publish them. Registries archive them. Thresholds are applied to them. At no point does measurement intervene as a veto.

The Table 1 diagnostic captures this perfectly. RTI moderately endorses the principle that non-falsifiable claims should be rejected at p = 0.70 (+0.85). Yet it strongly endorses the belief that reference-case simulations generate falsifiable claims at p = 0.90 (+2.20). This is the epistemic laundering mechanism. Simulation outputs are treated as if they were testable facts, even though they are conditional projections constructed from non-measures. This distinction matters. A falsifiable claim must expose itself to empirical refutation. Reference-case models cannot do this. They generate internally coherent narratives, not testable hypotheses. Sensitivity analysis explores model behavior, not reality. Yet RTI's modeling culture treats robustness as stability across assumptions rather than vulnerability to falsification.

This is how the evolution of objective knowledge is halted. Without measurable quantities, claims cannot be reproduced in the strong sense. They can only be rerun. Replication becomes repetition. Accumulation becomes aggregation of stories. RTI's role in this process is decisive. It trains analysts, advises manufacturers, participates in guideline development, authors methodological consensus documents, and supplies analytic labor to nearly every major HTA system. When RTI endorses a practice, it becomes safe. When RTI ignores a principle, it disappears.

This explains the striking uniformity of HTA failure across jurisdictions. NICE, ICER, CADTH, PBAC, and others differ institutionally, yet they share the same numerical artifacts. That is because they draw from the same upstream supply chain. RTI does not sit outside the memeplex; it is one of its primary replication engines.

The result is forty years of extraordinary technical effort devoted to perfecting arithmetic that should never have been permitted. Models have grown more complex. Utility algorithms more refined. Mapping methods more elaborate. Yet none of this has brought the field closer to measurement. It has only made non-measurement harder to see. The tragedy is not incompetence. RTI employs some of the most technically skilled methodologists in the field. The tragedy is epistemic misdirection. When technical brilliance is applied inside a framework that rejects measurement axioms, the output is not science. It is industrialized numerology.

If RTI were to accept representational measurement tomorrow, the consequences would be immediate and profound. Utilities would lose ratio status. QALYs would be abandoned. Mapping would be disallowed. Composite instruments would fail admissibility. Simulation models would be reclassified as scenario tools rather than evidentiary engines. Latent traits would be assessed only through Rasch logit ratio scales. Most of RTI's product portfolio would need to be rebuilt from first principles. That this has not occurred is not because the axioms are unknown. They have been available for over half a century. They have simply never been allowed to govern.

The Table 1 24-item diagnostic therefore reveals RTI's true historical role. It is not merely a consultancy. It is the organization that transformed HTA from a potentially empirical discipline into a highly efficient system for producing numbers immune to falsification. It sits at the center of the supply chain that ensures arithmetic survives even when measurement does not. Until that role is confronted directly, reform at the level of agencies or journals will remain cosmetic. The memeplex will regenerate upstream. And the field will continue to confuse methodological sophistication with scientific legitimacy. If HTA is ever to transition to normal science, that transition must begin precisely where RTI operates: at the point where numbers are manufactured. If measurement does not govern there, it will govern nowhere.

WHERE RTI FAILS: MEASUREMENT IS A NON-NEGOTIABLE PRECONDITION FOR ARITHMETIC

For organizations such as RTI, whose analytical products are routinely used to support comparative evaluation, economic modeling, and policy decision making, it is critical to make clear why demonstrable measurement status must precede the application of arithmetic. This requirement is not a methodological preference, nor a matter of disciplinary convention. It is a foundational condition that determines whether numerical operations are meaningful at all.

Arithmetic is not self-justifying. The ability to add, subtract, multiply, or divide numerical values depends entirely on the properties of the scale on which those values reside. Representational measurement theory specifies these conditions precisely. For numbers to represent quantities, the empirical attribute must be structured in a way that supports invariant ordering, constant unit differences and where multiplication is proposed, a true zero. Without these properties, numerical manipulation produces symbols, not measures.

From these axioms, only two admissible linear measurement forms exist. Interval scales permit addition and subtraction because they possess equal units but lack a meaningful zero. Ratio scales permit full arithmetic because they possess both equal units and a true zero. No intermediate category exists. No amount of statistical sophistication can create scale properties that are not already present in the empirical structure being represented.

This constraint applies equally to manifest and latent attributes. When the attribute of interest is directly observable, such as time, counts, or physical quantities, ratio measurement can often be established straightforwardly. However, when the attribute is latent, such as need fulfillment, health-related quality of life, symptom burden, functioning, or patient experience, the requirements do not weaken. They become more stringent. Latent attributes cannot be measured by inspection; they must be constructed through a formal measurement model capable of producing invariant units.

Within the human sciences, only one framework satisfies this requirement: Rasch measurement. Rasch models do not estimate scores; they construct measures. They impose unidimensionality, test item invariance, and generate a linear logit ratio scale on which meaningful differences in possession of a latent trait can be expressed. Without Rasch transformation, subjective responses remain ordinal. Summation, averaging, or statistical manipulation does not alter that fact.

This is the decisive point. Ordinal data do not become interval or ratio measures because they are widely used, because they correlate with other variables, or because they behave “reasonably” in regression models. Ordinality is a structural property, not a statistical inconvenience. Arithmetic performed on ordinal scores is not approximate measurement; it is invalid measurement.

The diagnostic results show that this principle is not merely underemphasized within RTI’s analytical knowledge base; it is absent. Propositions affirming the precedence of measurement over arithmetic, the requirement of scale-type coherence, and the necessity of Rasch transformation for latent traits fall at or near the floor of endorsement. At the same time, propositions that presume arithmetic legitimacy—summation of subjective instruments, construction of utilities, aggregation of QALYs, and multiplication of time by preference weights, are strongly reinforced.

This inversion is not trivial. It explains why complex modeling frameworks can appear internally coherent while remaining scientifically indefensible. Once arithmetic is permitted in advance of measurement validation, any numerical object can be treated as if it were quantitative. Models can be populated, sensitivity analyses conducted, and uncertainty intervals reported, all without resolving whether the dependent variable is a measure at all.

The consequence is that analytical rigor becomes detached from empirical meaning. Models become elaborate narratives built on numerically formatted assumptions. Outputs may be precise, but they are not measurable in the scientific sense. No amount of transparency, peer review, or replication of procedure can compensate for this failure, because replication without invariant quantities reproduces error, not knowledge.

For an organization positioned as a methodological authority within the health technology assessment supply chain, this omission is decisive. RTI's influence does not lie only in individual analyses, but in the normalization of analytic practices across sponsors, agencies, and health systems. When measurement status is not treated as a gatekeeping requirement, the entire downstream ecosystem inherits arithmetic without foundation.

This is why measurement must precede arithmetic. Not because it is philosophically elegant, but because without it, quantitative claims cannot be falsified, compared meaningfully across contexts, or accumulated as objective knowledge. Where there is no measure, there can be no empirical refutation—only scenario comparison and negotiated plausibility.

The diagnostic therefore identifies RTI's failure not as one of intent or competence, but of starting point. The axioms of representational measurement should have framed every subsequent analytic decision. Instead, they were bypassed. Until that ordering is reversed—until demonstrable measurement status is treated as a prerequisite rather than an afterthought—no analytical framework, however sophisticated, can claim scientific legitimacy. Absent demonstrable measurement, arithmetic becomes an exercise in symbol manipulation rather than scientific inference.

3. THE TRANSITION TO MEASUREMENT IN HEALTH TECHNOLOGY ASSESSMENT

THE IMPERATIVE OF CHANGE

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

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

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

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

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

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

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

MEANINGFUL THERAPY IMPACT CLAIMS

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

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

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

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

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

THE PATH TO MEANINGFUL MEASUREMENT

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

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

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

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

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

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

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

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

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

TRANSITION REQUIRES TRAINING

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

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

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

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

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

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

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

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

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

DESIGNED FOR CLOSURE

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

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

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

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

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

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

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

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

ACKNOWLEDGEMENT

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

REFERENCES

---

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

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

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

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