MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

UNITED STATES: INVALID MEASUREMENT IN HEALTH TECHNOLOGY ASSESSMENT — A STRUCTURAL ASSESSMENT OF THE HTA RELATED KNOWLEDGE BASE OF THE NORTHEASTERN UNIVERSITY SCHOOL OF PHARMACY AND PHARMACEUTICAL SCIENCES

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

LOGIT WORKING PAPER No 3037 APRIL 2026

www.maimonresearch.com

Tucson AZ

FOREWORD

HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

The Northeastern University School of Pharmacy and Pharmaceutical Sciences is one of the two PharmD-granting institutions in Massachusetts and is housed within the Bouvé College of Health Sciences at Northeastern University in Boston. The School integrates professional pharmacy education with pharmaceutical sciences research, emphasizing a model that combines academic instruction with extensive experiential learning. Central to its identity is Northeastern's well-established cooperative education (co-op) program, which places students in clinical, industry, regulatory, and research settings, providing structured exposure to real-world health care environments throughout their training.

The School offers the Doctor of Pharmacy (PharmD) degree alongside graduate programs in pharmaceutical sciences, encompassing areas such as drug development, health outcomes, and regulatory science. Its curriculum reflects a broad engagement with contemporary pharmacy practice, including medication therapy management, health systems science, and population health. Faculty research spans clinical pharmacy, pharmacoeconomics, pharmacoepidemiology, and translational science, often in collaboration with health systems, industry partners, and government agencies.

As a component of a large research university, the School operates within an interdisciplinary framework that encourages integration across health professions and scientific domains, positioning graduates for roles in clinical practice, research, and health system decision-making.

The objective of this study is to interrogate the HTA-associated knowledge base of the School of Pharmacy and Pharmaceutical Sciences using a standardized 24-item set of canonical statements derived from the axioms of representational measurement and Rasch measurement theory. Each statement is evaluated as TRUE or FALSE, with the School's implicit or explicit position inferred from its curriculum and research outputs. Categorical probabilities, constrained to discrete values ending in 0 or .05, are assigned to reflect the degree of endorsement within the knowledge base, and these are transformed into normalized logits on a symmetric ±2.50 scale. The purpose is not to survey opinions but to determine the extent to which the knowledge base behaves "as if" each statement were true, thereby identifying whether it supports credible, evaluable, and replicable value claims grounded in valid measurement. The study is framed within the broader concern that measurement must precede arithmetic, and that all HTA claims must be constrained to either linear ratio measures for manifest attributes or Rasch logit ratio measures for latent traits.

The findings demonstrate a highly polarized profile consistent with previous interrogations of academic pharmacy and HTA knowledge bases. Statements that would enforce the axioms of representational measurement—such as the requirement for ratio scales in multiplication (p=0.10; logit -2.20), the primacy of measurement over arithmetic (p=0.15; logit -1.75), and the necessity

of Rasch measurement for latent constructs (p=0.05; logit -2.50) are overwhelmingly rejected. In contrast, statements that sustain the conventional HTA framework, such as the acceptance of QALYs as ratio measures (p=0.95; logit +2.50), their dimensional homogeneity (p=0.95; logit +2.50), and the validity of aggregating QALYs (p=0.95; logit +2.50) are endorsed with near certainty. Simulation-based reference case modeling is similarly treated as generating credible claims (p=0.90; logit +2.20), despite its non-falsifiable structure. While there is nominal support for rejecting non-falsifiable claims (p=0.85; logit +1.75), this principle is not applied to core HTA methods. The overall pattern confirms a systematic measurement inversion: arithmetic operations are accepted in the absence of valid measurement properties, rendering resulting value claims non-evaluable.

The objective of this assessment is to evaluate the HTA-related knowledge base associated with MCPHS against the axioms of representational measurement using the 24-item canonical statement framework. The analysis seeks to determine whether the constructs, methods, and assumptions embedded in teaching and applied pharmacoeconomic practice meet the requirements for admissible measurement, including unidimensionality, invariance, dimensional homogeneity, and the principle that measurement must precede arithmetic. Each statement is assigned an endorsement probability and transformed into a normalized logit, providing a structured profile of how the knowledge base behaves when interrogated against these standards. The aim is not to evaluate individuals or courses, but to characterize the internal coherence of the HTA framework as presented and applied within a major pharmacy training environment.

The findings demonstrate a clear and highly polarized pattern of measurement inversion. Statements reflecting the axioms of representational measurement are weakly endorsed, with probabilities typically in the range of 0.05 to 0.25 (logits approximately −2.50 to −1.10), indicating that the conditions required for admissible arithmetic are not treated as binding constraints. In contrast, statements representing conventional HTA constructs—particularly those associated with QALYs, preference-based measures, and simulation outputs—are strongly endorsed, with probabilities in the range of 0.85 to 0.95 (logits approximately +1.75 to +2.50). This indicates that ordinal and composite constructs are treated as if they were ratio measures suitable for multiplication, aggregation, and comparison. While there is nominal recognition of scientific principles such as falsifiability, this is undermined by the acceptance of model-based outputs as if they were empirically testable claims. The resulting profile indicates that the knowledge base supports structured decision-making but cannot generate claims that are empirically evaluable, replicable, or falsifiable under the axioms of representational measurement.

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.

NORTHEASTERN UNIVERSITY SCHOOL OF PHARMACY AND PHARMACEUTICAL SCIENCES KNOWLEDGE BASE

The HTA-associated knowledge base of the School of Pharmacy and Pharmaceutical Sciences is not a formally bounded entity but an inferred construct derived from the School’s curriculum, research domains, faculty expertise, and its alignment with the broader academic and professional environment in which it operates. As part of a major research university, the School is embedded within an interdisciplinary health sciences framework that emphasizes clinical pharmacy practice, pharmaceutical sciences, health outcomes research, and engagement with healthcare delivery systems. This positioning ensures that students and faculty are exposed to contemporary approaches to evaluating therapeutic interventions, including pharmacoeconomic analysis, outcomes research, and real-world evidence generation.

Within this environment, the knowledge base reflects the dominant conventions of health technology assessment as practiced internationally. It incorporates the use of multiattribute preference instruments such as the EQ-5D, the construction of utility scores, and the application of cost-effectiveness frameworks that rely on the quality-adjusted life year as a central outcome. These elements are not typically presented as contested methodological choices but as standard tools of analysis. The reference case model, with its reliance on simulation techniques, extrapolation, and aggregation of costs and outcomes, is implicitly accepted as a legitimate framework for decision support. This acceptance is reinforced through teaching materials, published research, and exposure to regulatory and reimbursement environments where such methods are routinely applied.

At the same time, the knowledge base does not appear to engage with the foundational requirements of measurement theory that would govern the legitimacy of these practices. Concepts such as unidimensionality, invariance, admissible transformations, and dimensional homogeneity are not central to the evaluative framework. There is little evidence of a structured treatment of the distinction between ordinal, interval, and ratio scales, or of the implications of these distinctions for permissible mathematical operations. Similarly, while patient-reported outcomes and other subjective measures are widely used, they are typically analyzed through summation or transformation methods that do not satisfy the criteria for fundamental measurement. The Rasch model, as a means of constructing valid measures for latent traits, does not occupy a central role in the knowledge base.

The result is a system of evaluation that prioritizes the generation of numerical outputs over the establishment of measurement validity. The knowledge base supports the construction of composite endpoints, the aggregation of heterogeneous resource inputs into cost measures, and the manipulation of utility scores within arithmetic frameworks, without requiring that these elements conform to the axioms of representational measurement. This orientation is consistent with the broader HTA memeplex, in which methodological conventions are sustained through institutional

practice, publication standards, and regulatory expectations rather than through adherence to fundamental measurement principles.

Consequently, the Northeastern knowledge base can be characterized as one that is operationally coherent but epistemologically unconstrained. It provides the tools and frameworks necessary to participate in contemporary HTA practice, yet it does so without establishing the measurement properties required to ensure that resulting claims are scientifically credible. This tension between practical utility and theoretical validity lies at the heart of the interrogation and explains the observed pattern of probabilities and logits.

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: NORTHEASTERN UNIVERSITY SCHOOL OF PHARMACY AND PHARMACEUTICAL SCIENCES

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

It is essential to understand how to interpret these results.

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

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

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

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

THE ABSENCE OF REPRESENTATIONAL MEASUREMENT AND THE ENDORSEMENT OF FALSE MEASUREMENT

The interrogation of the HTA-associated knowledge base for the Northeastern University School of Pharmacy and Pharmaceutical Sciences yields a profile that is not merely consistent with prior assessments but, if anything, reinforces the conclusion that academic pharmacy programs in the United States are structurally embedded within a measurement-deficient evaluative framework. The probabilities assigned to each canonical statement demonstrate a highly polarized pattern, with strong endorsement of propositions that sustain the conventional HTA paradigm and equally strong rejection of those that would impose the constraints of representational measurement (Table). This is not an incidental misalignment; it is a systematic inversion of the relationship between measurement and arithmetic.

TABLE 1: ITEM STATEMENT, RESPONSE, ENDORSEMENT AND NORMALIZED LOGITS NORTHEASTERN UNIVERSITY SCHOOL OF PHARMACY AND PHARMACEUTICAL SCIENCES

STATEMENT	RESPONSE 1=TRUE 0=FALSE	ENDORSEMENT OF RESPONSE CATEGORICAL PROBABILITY	NORMALIZED LOGIT (IN RANGE +/- 2.50)
INTERVAL MEASURES LACK A TRUE ZERO	1	0.25	-1.10
MEASURES MUST BE UNIDIMENSIONAL	1	0.20	-1.40
MULTIPLICATION REQUIRES A RATIO MEASURE	1	0.10	-2.20
TIME TRADE-OFF PREFERENCES ARE UNIDIMENSIONAL	0	0.90	+2.20
RATIO MEASURES CAN HAVE NEGATIVE VALUES	0	0.90	+2.20
EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES	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	0.15	-1.75
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	0.95	+2.50
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	0.85	+1.75
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.90	+2.20
THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.80	+1,40
THE RASCH LOGIT RATIO SCALE IS THE ONLY BASIS FOR ASSESSING THERAPY IMPACT FOR LATENT TRAITS	1	0.05	-2.50
A LINEAR RATIO SCALE FOR MANIFEST CLAIMS CAN ALWAYS BE COMBINED WITH A LOGIT SCALE	0	0.35	-1.25
THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT	1	0.25	-1.10
THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.05	-2.50

The initial statements set the tone. The proposition that interval measures lack a true zero is accepted as true, but only weakly, with a probability of 0.25 (logit -1.10). Similarly, the requirement that measures must be unidimensional is acknowledged at a probability of 0.20 (logit -1.40). These are foundational axioms. Their weak endorsement signals that while they may be recognized in principle, they are not operationalized in practice. This becomes immediately evident in the response to the statement that multiplication requires a ratio measure, where the probability drops to 0.10 (logit -2.20). At this point, the knowledge base effectively abandons the constraints necessary for valid arithmetic. Once multiplication is permitted without ratio properties, the door is open to the construction of composite outcomes that violate dimensional homogeneity.

This is precisely what is observed in the treatment of QALYs. The statement that the QALY is a ratio measure is correctly identified as false, yet the knowledge base behaves as if it were true, with a probability of 0.95 (logit +2.50). The same probability attaches to the claim that the QALY is dimensionally homogeneous. These are not marginal endorsements; they are maximal. They indicate that the knowledge base not only accepts the QALY but does so with a degree of confidence that borders on certainty. The contradiction is stark: the axioms that would invalidate the QALY are weakly endorsed, while the conclusions that depend on ignoring those axioms are strongly affirmed. This is the essence of measurement inversion.

The role of preference-based instruments further illustrates the problem. The statement that EQ-5D-3L preference algorithms create interval measures is rejected by the canonical framework but endorsed with a probability of 0.90 (logit +2.20). This reflects a deeply entrenched belief that transformation procedures can elevate ordinal data to interval status. The same belief underpins the endorsement, at 0.90 (logit +2.20), of the claim that summations of subjective instrument

responses are ratio measures. These positions are not defensible within the axioms of representational measurement. They persist because the knowledge base does not subject them to scrutiny.

The rejection of Rasch measurement is even more decisive. The statement that there are only two classes of measurement—linear ratio and Rasch logit ratio—is endorsed at a probability of only 0.05 (logit -2.50). The same probability is assigned to the proposition that transforming subjective responses to interval measurement is only possible with Rasch rules, and to the claim that the Rasch logit ratio scale is the only basis for assessing therapy impact for latent traits. These are not borderline cases; they represent near-total rejection. The knowledge base has no place for Rasch measurement. It does not engage with it, does not teach it, and does not consider it necessary. Instead, it relies on summation and transformation techniques that lack a theoretical foundation.

The implications for cost-effectiveness analysis are immediate. The statement that claims for cost-effectiveness fail the axioms of representational measurement is accepted only weakly, with a probability of 0.15 (logit -1.75). In contrast, the aggregation of QALYs is endorsed at 0.95 (logit +2.50), and the validity of reference case simulations is implicitly supported through the rejection, at 0.90 (logit +2.20), of the claim that such simulations fail to generate falsifiable outcomes. The knowledge base treats simulation outputs as if they were empirical evidence. There is no recognition that these models are closed systems, constructed on assumptions that cannot be tested within the model itself.

The one area where the knowledge base aligns with scientific principles is in the rejection of non-falsifiable claims, endorsed at 0.85 (logit +1.75). However, this alignment is superficial. It is not extended to the core methodologies of HTA. Simulation models, which are inherently non-falsifiable, are accepted without question. The principle of falsifiability is acknowledged in the abstract but ignored in application. This selective application of scientific criteria further reinforces the conclusion that the knowledge base is not governed by a coherent epistemological framework.

The treatment of measurement theory is similarly inconsistent. The statement that measurement precedes arithmetic is endorsed at only 0.15 (logit -1.75), indicating that the knowledge base does not prioritize the establishment of measurement properties before engaging in mathematical operations. This is the defining characteristic of the current HTA paradigm. Arithmetic is performed first, and measurement is assumed rather than demonstrated. The result is a system that produces numbers without ensuring that those numbers have meaning.

The concept of latent traits provides another point of divergence. The outcome of interest for latent traits—the possession of that trait—is recognized only weakly, with a probability of 0.25 (logit -1.10). This suggests that the knowledge base does not fully engage with the ontological status of latent constructs. Without a clear understanding of what is being measured, the question of how it should be measured becomes irrelevant. This may explain the absence of Rasch modeling. If the construct itself is not clearly defined, there is no perceived need for a rigorous measurement framework.

The final statement, that the Rasch rules for measurement are identical to the axioms of representational measurement, is endorsed at a probability of 0.05 (logit -2.50). This is perhaps the

most revealing result. It indicates that the knowledge base does not recognize the equivalence between Rasch measurement and the broader principles of representational measurement. Without this recognition, Rasch appears as an optional or specialized technique rather than as a necessary condition for valid measurement of latent traits.

Taken together, the probabilities and logits present a coherent picture. High probabilities cluster around statements that support the conventional HTA framework: the validity of QALYs, the aggregation of utilities, the use of simulation models, and the transformation of ordinal data. Low probabilities cluster around statements that would impose measurement discipline: the requirement for ratio scales, the necessity of unidimensionality, the role of Rasch measurement, and the primacy of measurement over arithmetic. The distribution is not random. It reflects a systematic preference for methods that enable the construction of numerical outputs, regardless of their measurement properties.

The Northeastern knowledge base does not differ in any substantive way from those previously interrogated. It participates in the same global memeplex, characterized by the acceptance of numerical storytelling as a substitute for measurement. The institutional context—an emphasis on clinical practice, experiential learning, and applied research—does not mitigate this. If anything, it reinforces it. Students are trained to use the tools of HTA without questioning their validity. The result is a generation of practitioners who are proficient in the application of methods that do not meet the standards of normal science.

This raises a fundamental question. If the knowledge base does not support valid measurement, what exactly is being taught? The answer is that it teaches a set of conventions. These conventions are internally consistent and widely accepted, but they are not grounded in the axioms of measurement. They persist because they are useful, not because they are true. The distinction is critical. Usefulness does not confer validity. A method can be widely used and still be fundamentally flawed.

The persistence of this system depends on the continued acceptance of its outputs. As long as decision-makers are willing to treat cost-per-QALY ratios as meaningful, the system will continue to function. However, once the measurement assumptions are scrutinized, the foundation begins to erode. The probabilities and logits make this erosion visible. They quantify the extent to which the knowledge base departs from the requirements of representational measurement.

In this sense, the Northeastern assessment is not an outlier. It is a confirmation. It confirms that the absence of measurement is not confined to specific institutions or countries. It is a defining feature of the HTA paradigm. The question is not whether this paradigm can be adjusted or refined. The question is whether it can survive the introduction of measurement standards.

At present, the answer appears to be no. The knowledge base is too deeply invested in methods that would be invalidated by such standards. The adoption of Rasch measurement and ratio-scale requirements would necessitate a fundamental restructuring of HTA. It would require the abandonment of QALYs, the rejection of composite cost measures, and the replacement of simulation models with empirically evaluable claims. This is not a minor reform. It is a transformation.

The School of Pharmacy and Pharmaceutical Sciences, like its peers, stands at this crossroads. The interrogation presented here does not prescribe a solution. It does, however, make the problem explicit. The probabilities and logits leave little room for ambiguity. The knowledge base, as currently constituted, does not meet the standards of representational measurement. Whether it chooses to address this remains an open question.

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.

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