MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

UNITED STATES: THE *AMERICAN JOURNAL OF MANAGED CARE* AND THE ABSENCE OF MEASUREMENT

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

LOGIT WORKING PAPER No 23 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 HTA presents a world of measurement failure.

The objective of this assessment is to examine the *American Journal of Managed Care* as a critical node in the health technology assessment and formulary decision ecosystem, using a 24-item diagnostic grounded in the axioms of representational measurement theory. Rather than evaluating individual articles or editorial intent, the analysis interrogates the journal's *knowledge base*: the recurring concepts, assumptions, analytic practices, and numerical constructs that AJMC repeatedly treats as admissible evidence for managed care decision making. The purpose is to determine whether the quantitative claims normalized within this corpus satisfy the minimum conditions required for meaningful arithmetic, falsification, and cumulative learning, or whether the journal functions primarily as an operational conduit for non-measurable constructs inherited from the dominant HTA memeplex.

This inquiry is especially important given AJMC's explicit positioning as a decision-facing journal. Unlike methodological or theoretical outlets, AJMC targets payer audiences, formulary committees, and health system leadership, translating academic health economics and outcomes research into actionable narratives. As such, its epistemic responsibilities extend beyond scholarly discourse to governance practice. The central question is therefore whether AJMC serves as a gatekeeper that filters out non-admissible quantitative claims or whether it reproduces and legitimizes arithmetic without measurement at the point where it most directly influences access, pricing, and coverage policy.

The findings are unambiguous. The AJMC exhibits the same structural inversion of measurement and arithmetic observed in upstream HTA journals, but with greater practical consequence. Core axioms required for quantitative inference, measurement preceding arithmetic, scale-type coherence, unidimensionality, and the inadmissibility of composite constructs, are weakly endorsed or effectively absent. At the same time, the journal strongly reinforces false propositions that enable routine cost-utility reasoning, including the treatment of utilities and QALYs as ratio measures, the permissibility of aggregating heterogeneous outcomes, and the legitimacy of reference-case simulation outputs as decision variables.

Rasch measurement, the only framework capable of producing invariant logit ratio measures for latent attributes, is functionally excluded from the journal's methodological boundaries. Latent trait possession is not recognized as the outcome of interest, and subjective instruments are treated as quantitative through summation rather than measurement. As a result, AJMC does not merely tolerate false measurement; it operationalizes it. The journal functions not as a corrective to HTA epistemic failure, but as a downstream amplifier that converts non-measures into administratively usable decision artifacts for managed care systems.

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 KNOWLEDGE BASE OF THE *AMERICAN JOURNAL OF MANAGED CARE*

The knowledge base of the AJMC can be characterized as an applied translation system rather than a measurement-governed evaluative framework. It is organized around the transformation of health economics and outcomes research outputs into decision-relevant narratives for payer audiences, with little interrogation of the measurement properties underlying those outputs. Within this system, numerical form is treated as sufficient evidence of quantification, and methodological sophistication is routinely conflated with measurement validity.

At the core of the AJMC knowledge base is the acceptance of composite and derived constructs as decision-grade quantities. Utilities, quality-adjusted life-years, incremental cost-effectiveness ratios, and modeled value metrics are presented as if they were stable numerical objects rather than conditional artifacts of specific instruments, algorithms, and assumptions. The journal does not require demonstration that these constructs meet the axioms of representational measurement before they are interpreted, compared, or used to support coverage and reimbursement narratives.

Subjective outcomes occupy a central position in the AJMC corpus. Patient-reported outcome instruments, satisfaction measures, and quality-of-life scores are routinely summarized, averaged, and compared across populations. These scores are treated as continuous variables despite originating from ordinal response categories that lack equal intervals and invariance. Psychometric indicators, such as reliability, responsiveness, or construct validity, are used as surrogates for measurement, even though they cannot establish scale type or permissible arithmetic. In this way, the journal sustains a scoring-based epistemology rather than a measurement-based one.

Latent attributes are frequently invoked but never formally constructed. Concepts such as quality of life, functioning, burden, adherence experience, or patient value are treated by the AJMC corpus as quantities without being defined as single attributes with measurable continua. Unidimensionality is not enforced as a prerequisite for analysis. Multi-domain instruments and composite indices are therefore permitted to masquerade as single outcomes, enabling arithmetic operations that would be disallowed under any measurement-literate framework.

The absence of Rasch measurement is decisive in defining the journal’s epistemic boundaries. Rasch modeling, which uniquely provides invariant logit ratio measures of latent trait possession, is not treated as a gatekeeping requirement. Without Rasch transformation, ordinal responses remain ordinal regardless of subsequent statistical manipulation. The journal’s methodological environment nevertheless allows regression modeling, responder analyses, and change-score interpretation to proceed as if interval or ratio properties had been established.

The AJMC knowledge base also exhibits strong alignment with reference-case modeling conventions. Simulation outputs are routinely treated as credible surrogates for empirical claims, and model stability is interpreted as evidentiary robustness. This permits closure without falsification. Rather than requiring prospective, reproducible protocols capable of empirical refutation, the journal accepts internally coherent modeling frameworks as sufficient justification for decision support.

Most importantly, the journal does not function as a measurement gatekeeper. Representational measurement theory is absent as an organizing framework. Scale-type admissibility is not treated as a threshold condition for publication. Measurement precedes arithmetic only rhetorically, not operationally. As a result, AJMC transmits the HTA memeplex intact from academic modeling culture into managed care governance.

In doing so, AJMC helps normalize a system in which coverage and pricing decisions are informed by numerical artifacts that cannot support falsification or cumulative learning. The knowledge base is therefore not neutral. It is structured to preserve continuity of practice rather than to enforce the conditions required for scientific evaluation of therapy impact.

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: *AMERICAN JOURNAL OF MANAGED CARE*

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 *AMERICAN JOURNAL OF MANAGED CARE*

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	0.20	-1.40
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	0.90	+2.20
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	0.85	+1.75
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	0.20	-1.40
QALYS CAN BE AGGREGATED	0	0.95	+2.50
NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.65	+0.60
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.05	-2.50
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	0.25	-1.10
THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.05	-2.50

AMERICAN JOURNAL OF MANAGED CARE: ABSENCE OF A GATEKEEPER FUNCTION AND REPRESENTATIONAL MEASUREMENT

AJMC is not a peripheral newsletter that can be waved away as "industry media." It describes itself as a peer-reviewed, MEDLINE-indexed journal aimed at decision-makers in managed care and health policy, positioned inside a multimedia brand that explicitly targets payer and policy audiences. That positioning matters because it defines the journal's implicit responsibility: if any U.S. outlet should act as a gatekeeper for formulary-relevant claims, it is one that says it keeps leaders "on the forefront" and publishes research "relevant to decision-makers." The diagnostic profile shows the opposite. AJMC does not function as a gatekeeper against false measurement. It functions as a conduit that makes false measurement administratively usable, professionally normal, and therefore "decision-ready."

The AJMC pattern is not subtle. It reproduces the same two-pillar structure you have already exposed in Value in Health and Pharmacoeconomics: one pillar supplies legitimacy, the other supplies reinforcement. AJMC sits downstream of those pillars, translating their permissive arithmetic into managed-care operations—coverage restrictions, utilization management narratives, outcomes "value" dashboards, and the routine expectation that submissions arrive pre-packaged in the conventional grammar: utilities, QALYs, ICERs, and reference-case model outputs. That is exactly where measurement failure becomes consequential. Journals can be wrong in private; managed care makes wrongness actionable. AJMC's role in the ecosystem is therefore particularly damaging: it takes mathematically non-admissible objects and helps make them feel like standard equipment for payer practice.

The decisive feature of the AJMC logit profile is the inversion of measurement standards: arithmetic is treated as primary, while measurement is treated as optional background noise. The proposition "measurement precedes arithmetic" collapses to p = 0.10 with a canonical logit of −2.20. The companion proposition—"meeting the axioms of representational measurement is required for arithmetic"—is also at p = 0.10 (−2.20). This is not an academic quibble; it is the gate that determines whether any subsequent calculation is meaningful. When that gate is shut, everything downstream becomes performative: models can be built, ratios can be computed, thresholds can be debated, and committees can claim "evidence-based" justification, but none of it is anchored to demonstrable measurement properties.

AJMC's profile then reveals what replaces the gate. The journal ecosystem normalizes the specific propositions that keep cost-utility arithmetic alive. "The QALY is a ratio measure" sits at p = 0.90 (+2.20). "QALYs can be aggregated" sits at p = 0.95 (+2.50). "EQ-5D preference algorithms create interval measures” sits at p = 0.90 (+2.20). “Ratio measures can have negative values” sits at p = 0.90 (+2.20). These are not independent errors. They form a mutually supporting cluster of enabling beliefs. If the QALY is ratio, then multiplication by time is permissible; if aggregation is permissible, then population-level coverage policy can be justified; if the EQ-5D algorithm is treated as interval (or better), then averaging and regression feel “safe”; if negative values are allowed while still calling the construct ratio, then the most conspicuous contradiction in the entire utility enterprise is quietly domesticated. This is exactly how a memeplex protects itself: it does not win by proving itself; it wins by making the contradictions professionally ignorant.

Notice the hypocrisy built into the profile. Time is correctly recognized as ratio at p = 0.95 (+2.50). That tells you the journal’s knowledge system is not incapable of understanding ratio structure. It understands it perfectly when the attribute is manifest, physical, and uncontroversial. The failure is therefore not “lack of quantitative sophistication.” It is selective exemption. Ratio discipline is applied where it is cheap and abandoned where it would be fatal to the preferred architecture. And the preferred architecture in managed care is the one that produces closure: a single number, a modeled ratio, a threshold story, and a coverage recommendation that can be defended as “standard practice.”

The multiplication item is the cleanest exposure of that exemption. “Multiplication requires a ratio measure” collapses to p = 0.10 (−2.20), while the whole cost-utility genre depends on multiplying time by a preference weight and pretending the product is a quantitative health outcome. This is not merely inconsistent. It is structurally dishonest. If multiplication requires ratio measurement, then the moment you admit utilities are not ratio you must stop computing QALYs. AJMC’s profile shows that the system avoids the admission by downgrading the multiplication requirement rather than by repairing measurement.

Unidimensionality is the next gate that AJMC refuses to guard. “Measures must be unidimensional” sits at p = 0.20 (−1.40), weak enough to be effectively non-binding. Yet the journal’s ecosystem relies on instruments and indices that are explicitly composite; multi-domain quality-of-life summaries, weighted indexes, preference algorithms that compress heterogeneous descriptors into a single number. When unidimensionality is not enforced, composite objects can be treated as if they were single attributes, and the journal can publish endless analyses that look quantitative while never establishing that a quantity exists. That is why insistence on unidimensional claims is not stylistic preference; it is the minimum requirement for the very concept of “more” and “less” to have measurable meaning.

The most destructive part of AJMC’s profile, however, is the normalization of score-arithmetic on subjective instruments. “Summation of Likert question scores creates a ratio measure” sits at p = 0.90 (+2.20). “Summations of subjective instrument responses are ratio measures” sits at p = 0.85 (+1.75). These are the upstream lies that feed every downstream lie. Once a community treats ordinal category totals as ratio-like quantities, everything becomes easy: you can compute means, differences, responder thresholds, cost per unit change, mapped utilities, and eventually cost per QALY. This is why the journal *Quality of Life Research* is such a powerful upstream supplier; AJMC is a powerful downstream enabler. The journal does not need to defend the metaphysics of utility; it only needs to keep publishing as if summed scores are already measures and as if the rest is just “analysis.”

Now we get to the critical point: latent traits and Rasch. The Rasch block is not "low." It is annihilated. "There are only two classes of measurement, linear ratio and Rasch logit ratio" is at p = 0.05 (−2.50). "Transforming subjective responses to interval measurement is only possible with Rasch rules" is at p = 0.05 (−2.50). "The Rasch logit ratio scale is the only basis for assessing therapy impact for latent traits" is at p = 0.05 (−2.50). "Rasch rules are identical to the axioms of representational measurement" is at p = 0.05 (−2.50). Those floor values matter because they show more than ignorance; they show boundary enforcement. Rasch is not merely "not emphasized." Rasch is treated as something the journal ecosystem can live without precisely because accepting Rasch as the gatekeeper would detonate the score-based instrument families that dominate outcomes research and the mapping machinery that translates those scores into utilities for economic models.

AJMC's treatment of "possession" reveals the same avoidance. "The outcome of interest for latent traits is the possession of that trait" sits at p = 0.25 (−1.10). That is exactly the level you would expect in a corpus that wants to talk about "improvement" without ever being pinned down on what is being improved, in what units, and with what invariance. Possession is dangerous to a memeplex because it forces the measurement question into the open: if a latent trait exists, and if people possess more or less of it, then the only defensible way to quantify that possession is a Rasch logit ratio scale built under invariance constraints. Without that, you do not have measured possession; you have ranked responses and summed scores. AJMC's profile shows that the managed care knowledge base prefers the convenience of scores to the accountability of measures.

The falsification items show how the journal preserves a scientific posture while avoiding scientific exposure. "Non-falsifiable claims should be rejected" is moderately endorsed at p = 0.65 (+0.60). That sounds like virtue. But it is immediately neutralized by "reference case simulations generate falsifiable claims" at p = 0.85 (+1.75). This is the key laundering maneuver in the managed-care setting. Reference-case models generate conditional projections. They can be made to look stable under sensitivity analyses, but stability across assumptions is not falsification. A claim is falsifiable only when it is tied to a protocol that risks refutation against observed outcomes in defined populations and timeframes. AJMC's profile shows that the journal ecosystem treats "model discipline" as a substitute for empirical risk, which is precisely how reference-case modeling became a closure machine: it produces a decision variable that looks rigorous without exposing the decision variable to the world.

This is where AJMC's failure as a gatekeeper becomes a governance problem for health systems. Managed care organizations need evaluation frameworks that can be revisited, corrected, and improved, because formularies are not one-time choices; they are long-horizon commitments with budget, access, and patient outcome consequences. A journal that normalizes non-measures as endpoints and conditional projections as "evidence" is not merely making a technical mistake. It is undermining the possibility of cumulative learning. If the dependent variable is not a measure, then post-listing evaluation cannot accumulate objective knowledge; it can only accumulate more modeled stories, more score changes, and more negotiated interpretations.

The parallels to *Value in Health* and *Pharmacoeconomics* are direct and ugly. *Value in Health* supplies the rhetorical legitimacy: "good practice," "methods standards," "consensus." *Pharmacoeconomics* supplies reinforcement: the repeated normalization that utilities, QALYs, ICERs, and reference cases are the discipline's natural language. AJMC supplies operationalization. It is the place where these constructs are made to feel like standard tools for decision-makers. That is why a managed care journal is not "down market" in epistemic terms. In the memeplex, downstream journals can be more dangerous than upstream ones because they embed false measurement into the routines of governance: coverage criteria, step therapy, prior authorization rationales, and "value-based" contracting narratives that presuppose quantities that do not exist as measures.

If you want the hard conclusion that Table 1 warrants, it is this: AJMC does not merely fail to police measurement; it helps teach decision-makers that policing measurement is unnecessary. It reinforces the idea that if a number is published, if it is peer reviewed, if it is modeled with sophistication, and if it aligns with reference-case conventions, then it is fit to govern access and price. That is precisely the opposite of what representational measurement requires. Under measurement-first standards, the journal would treat scale type as a gatekeeping condition; it would treat unidimensionality as non-negotiable; it would treat Rasch measurement as mandatory for latent trait claims; it would treat QALYs and utility arithmetic as inadmissible; and it would treat reference-case outputs as conditional narratives unless tied to falsifiable protocols. AJMC's canonical logit profile shows that the journal ecosystem does the reverse: it downgrades the constraints and elevates the conveniences.

This critique is not that AJMC is "wrong" in some diffuse way. The critique is that AJMC sits in a position where it could have functioned as a corrective, an epistemic checkpoint between academic modeling culture and payer decision culture and it has chosen, structurally, not to do so. It reproduces the same measurement inversion that is documented across the HTA ecosystem, but with an added harm: it converts that inversion into decision practice. In a managed care context, that is not an intellectual embarrassment; it is a systematic mechanism for making non-evaluable claims govern real coverage decisions.

If AJMC wanted to behave as a genuine gatekeeper for formulary claims, the corrective is not complicated, but it is disruptive: publish only those therapy impact claims that are explicitly unidimensional; require explicit declaration of scale type and permissible arithmetic; require linear ratio measures for manifest claims; require Rasch logit ratio measures for latent claims; prohibit QALYs, utilities, and mapping outputs as "measures"; and treat reference-case models as descriptive scaffolding that cannot close a case without prospective, reproducible protocols. Until those standards are enforced, AJMC's practical role in the supply chain is clear: it helps keep the memeplex alive by ensuring that arithmetic continues to outrun measurement, and that managed care decision-makers are never forced to confront the fact that the central numerical objects they are asked to use are not, in the representational measurement sense, measures at all.

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.

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