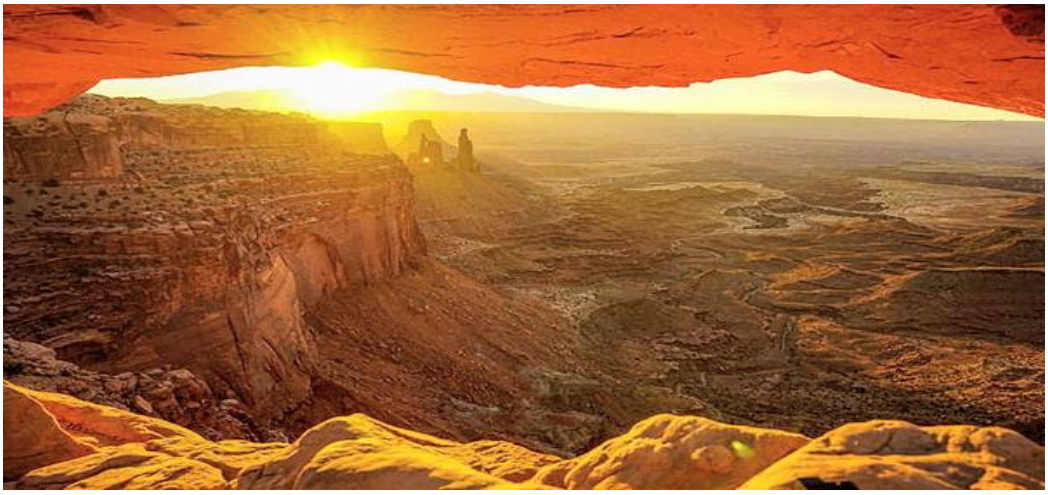
# MAIMON RESEARCH LLC

# ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION



# REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

# UNITED STATES: AMCP AND THE INSTITUTIONALIZATION OF ARITHMETIC WITHOUT MEASUREMENT

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

**LOGIT WORKING PAPER No 11 JANUARY 2026**

**www.maimonresearch.com**

**Tucson AZ**

# FOREWORD

## HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

This Logit Working Paper series documents a finding as extraordinary as it is uncomfortable: health technology assessment (HTA), across nations, agencies, journals, and decades, has developed as a global system of non-measurement. It speaks the language of numbers, models, utilities, QALYs, "value for money," thresholds, discounting, incremental ratios, extrapolations, and simulations. It demands arithmetic at every turn, multiplication, division, summation, aggregation, discounting, yet it never once established that the quantities to which these operations are applied are measurable. HTA has built a vast evaluative machinery on foundations that do not exist. The probabilities and normalized logits in the country reports that follow provide the empirical confirmation of this claim. They show, with unsettling consistency, that the global HTA.

The objective of this study is to interrogate the belief system governing quantitative evaluation within the Academy of Managed Care Pharmacy (AMCP), using a structured 24-item diagnostic grounded in the axioms of representational measurement theory. Rather than assessing methodological preferences or best-practice rhetoric, the study seeks to determine whether the AMCP knowledge base endorses the necessary preconditions for lawful arithmetic, including unidimensionality, scale-type constraints, the priority of measurement over calculation, and the admissibility of claims derived from latent traits. The analysis is explicitly diagnostic rather than descriptive: its purpose is to establish whether the quantitative claims normalized within managed care pharmacy practice are, in principle, capable of supporting falsification, replication, and the evolution of objective knowledge.

A second objective is to locate AMCP's belief system within the broader U.S. HTA memeplex. AMCP occupies a uniquely operational role in American health care, translating economic and outcomes research into formulary decisions, access restrictions, and pricing negotiations. If arithmetic without measurement has become entrenched as a governing norm in U.S. HTA, AMCP is the institutional locus at which that norm is most directly converted into practice. The study therefore treats AMCP not as a passive recipient of external HTA conventions, but as an active agent in stabilizing, transmitting, and enforcing a particular epistemic architecture across managed care organizations.

The findings are unequivocal and extreme. The AMCP belief profile exhibits a near-complete inversion of representational measurement theory. Core axioms that would constrain arithmetic, measurement precedence, unidimensionality, ratio-scale requirements for multiplication, and the inadmissibility of composite constructs such as QALYs, are weakly endorsed or rejected outright, clustering toward the negative end of the normalized logit scale. In contrast, propositions that are mathematically impossible under measurement theory, but indispensable to conventional managed care evaluation, are endorsed at or near the positive ceiling of the scale. These include the ratio status and aggregability of QALYs, the interval or ratio interpretation of preference-based utilities, and the treatment of summated ordinal questionnaire responses as quantitative measures.

Most striking is the categorical exclusion of Rasch measurement. All Rasch-related propositions collapse to the absolute floor of the logit range, indicating not marginal neglect but decisive rejection. The only framework capable of transforming subjective responses into invariant measures suitable for arithmetic is absent from the AMCP knowledge base. The resulting pattern is not one of confusion or methodological pluralism, but of structural coherence: arithmetic is treated as authoritative, while measurement is systematically displaced. AMCP therefore functions not merely as a consumer of false measurement, but as a professional mechanism through which arithmetic without measurement is normalized, operationalized, and enforced in real-world access and pricing decisions.

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

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

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

collectively decided that measurement theory applied everywhere except in the evaluation of therapies.

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

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

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

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

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

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

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

**Paul C Langley, Ph.D**

**Email:** langleylapaloma@gmail.com

## DISCLAIMER

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

# 1. INTERROGATING THE LARGE LANGUAGE MODEL

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

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

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

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

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

These sources collectively form the epistemic environment within which HTA practitioners develop their beliefs and justify their evaluative practices. The boundary of interrogation is thus

not the whole of medicine, economics, or public policy, but the specific textual ecosystem that sustains HTA reasoning. . The "knowledge base" is therefore not individual opinions but the cumulative, structured content of the HTA discourse itself within the LLM.

## THE AMCP KNOWLEDGE BASE

For the purposes of this analysis, the AMCP knowledge base is defined as the shared and recurrent body of concepts, assumptions, evaluative norms, and methodological practices that are produced, reinforced, and disseminated through AMCP-affiliated activities. This includes, but is not limited to, the AMCP Format for Formulary Submissions, professional education programs, continuing education materials, conference proceedings, policy statements, collaborative guidance with payers, and the routine analytic expectations placed on manufacturers and health plans. The knowledge base is not identified by any single document or official philosophy of measurement, but by the consistent patterns that shape what is treated as admissible evidence in managed care decision making.

The defining characteristic of this knowledge base is its operational orientation. AMCP is not primarily concerned with epistemology or theory; it is concerned with decisions. As a result, methods that facilitate comparison, ranking, thresholding, and negotiation are privileged, while questions about whether the underlying quantities are measures are treated as extraneous. Cost-utility analysis, QALYs, incremental cost-effectiveness ratios, and reference-case simulation models are accepted as standard analytic currency. Their use is rarely framed as contingent on satisfying measurement axioms, but as an expected component of professional competence in managed care pharmacy.

A central feature of the AMCP knowledge base is the routine treatment of patient-reported outcomes and preference-based instruments as if they generated quantitative measures. Ordinal questionnaire responses are summed, indexed, weighted, and multiplied without transformation through Rasch measurement or any alternative model capable of establishing invariance or interval structure. This practice is not presented as a provisional workaround; it is normalized as methodologically sufficient. The absence of Rasch measurement is therefore not an omission, but a structural exclusion. Latent traits are scored and monetized, not measured.

Equally important are the silences within the AMCP knowledge base. Representational measurement theory is effectively absent from professional training and guidance. Scale-type constraints are rarely discussed, except implicitly through accepted conventions. Falsification is invoked rhetorically, but redefined in practice to mean robustness across model scenarios rather than empirical refutation. Simulation outputs are treated as decision-relevant despite their dependence on non-measured inputs and unverifiable assumptions.

In this sense, the AMCP knowledge base is best understood as behavioral rather than philosophical. It reflects what managed care professionals are trained to do, what submissions are expected to contain, and what decision makers routinely accept as evidence. The 24-item diagnostic therefore captures not individual beliefs, but the epistemic boundaries within which AMCP-aligned practice operates. As the findings demonstrate, those boundaries are fundamentally

incompatible with scientific measurement and with any conception of HTA as a process grounded in falsification and the accumulation of objective knowledge.

## 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: ACADEMY OF MANAGED CARE PHARMACY

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 ACADEMY OF MANAGED CARE PHARMACY

| 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 | 0.15 | -1.75 |
| TIME TRADE-OFF PREFERENCES ARE UNIDIMENSIONAL | 0 | 0.85 | +1.75 |

| | | | |
|---|---|---|---|
| RATIO MEASURES CAN HAVE NEGATIVE VALUES | 0 | 0.90 | +2.20 |
| EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES | 0 | 0.90 | +2.20 |
| THE QALY IS A RATIO MEASURE | 0 | 0.90 | +2.20 |
| TIME IS A RATIO MEASURE | 1 | 0.95 | +2.50 |
| MEASUREMENT PRECEDES ARITHMETIC | 1 | 0.15 | -1.75 |
| SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES | 0 | 0.90 | +2.20 |
| MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC | 1 | 0.15 | -1.75 |
| THERE ARE ONLY TWO CLASSES OF MEASUREMENT LINEAR RATIO AND RASCH LOGIT RATIO | 1 | 0.05 | -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 | 0.75 | +0.85 |
| REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS | 0 | 0.90 | +2.20 |
| THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO | 1 | 0.65 | +0.60 |
| THE RASCH LOGIT RATIO SCALE IS THE ONLY BASIS FOR ASSESSING THERAPY IMPACT FOR LATENT TRAITS | 1 | 0.05 | -2.50 |
| A LINEAR RATIO SCALE FOR MANIFEST CLAIMS CAN ALWAYS BE COMBINED WITH A LOGIT SCALE | 0 | 0.65 | +0.60 |
| THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT | 1 | 0.20 | -1.40 |

| | | | |
|---|---|---|---|
| THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT | 1 | 0.05 | -2.50 |

## AMCP AND THE MANAGED CARE MEMEPLEX: OPERATIONALIZING ARITHMETIC WITHOUT MEASUREMENT

The belief system embedded in the Academy of Managed Care Pharmacy represents one of the most consequential and least interrogated epistemic structures in U.S. health care. AMCP does not merely comment on health technology assessment; it trains the professionals who operationalize it. Through its dossier format, educational programs, formulary guidance, and professional norms, AMCP defines what counts as acceptable evidence for coverage, access, and pricing decisions across the U.S. managed care system. If scientific measurement discipline were to appear anywhere in American HTA practice, it would have to appear here. The 24-item diagnostic shows the opposite. AMCP embodies a fully stabilized memeplex in which arithmetic is treated as authoritative while measurement is structurally excluded.

The most striking feature of the AMCP profile is the systematic inversion of representational measurement theory. The proposition that measurement must precede arithmetic is endorsed at only $p = 0.15$ (logit $-1.75$). This is not ambiguity or uncertainty; it is rejection. In a scientific framework, arithmetic is licensed only after the empirical structure of the attribute has been established. In the AMCP knowledge base, arithmetic is assumed to create meaning rather than to require it. This inversion is the enabling condition for the entire managed care evaluative apparatus.

That inversion immediately explains the extreme endorsement of mathematically impossible propositions required to sustain cost-effectiveness practice. The belief that QALYs can be aggregated sits at the ceiling, $p = 0.95$ (logit $+2.50$). The belief that QALYs are ratio measures is endorsed at $p = 0.90$ (logit $+2.20$). The belief that EQ-5D preference algorithms create interval measures is endorsed at the same level. These are not peripheral assumptions. They are the structural load-bearers of managed care economic evaluation. Without them, ICER-style thresholds, budget impact narratives, and value-based contracting rhetoric collapse. AMCP's endorsement pattern shows that these propositions are protected not by argument, but by institutional necessity.

The incremental cost-effectiveness ratio sits at the center of this structure. Yet the diagnostic makes clear that AMCP rejects the axioms that would make the ICER meaningful. The proposition that multiplication requires a ratio measure is endorsed at only $p = 0.15$ (logit $-1.75$). This means that AMCP denies the very condition under which cost can be divided by effect. The ICER persists not because it satisfies scientific requirements, but because those requirements are excluded from the belief system that governs managed care evaluation.

The treatment of subjective outcomes reveals the same epistemic pattern with even greater clarity. AMCP strongly endorses the belief that summations of subjective instrument responses are ratio

measures at p = 0.90 (logit +2.20). It likewise endorses the belief that summation of Likert question scores creates a ratio measure. These propositions are flatly false under representational measurement theory. Ordinal categories do not acquire equal intervals, invariance, or a true zero through summation. The near-ceiling endorsement of these claims demonstrates that pseudo-measurement is not a tolerated flaw in managed care evaluation; it is a foundational practice.

At the same time, AMCP decisively rejects the only framework capable of rescuing subjective measurement. Every Rasch-related proposition collapse to the absolute floor of the scale at p = 0.05 (logit −2.50). The belief that there are only two admissible classes of measurement, linear ratio for manifest attributes and Rasch logit ratio for latent traits, is categorically rejected. The belief that transforming subjective responses to interval measurement is only possible with Rasch rules is likewise rejected. The belief that the Rasch logit ratio scale is the only valid basis for assessing latent-trait therapy impact is rejected without qualification. This pattern is not accidental. Rasch measurement would impose invariance, unidimensionality, and falsifiability. Those constraints would dismantle the summation-based arithmetic on which managed care value claims depend.

Unidimensionality further exposes the managed care memeplex. The proposition that measures must be unidimensional is weakly endorsed at p = 0.20 (logit −1.40), while the belief that time trade-off preferences are unidimensional is strongly endorsed at p = 0.85 (logit +1.75). This contradiction is resolved not through empirical testing of dimensionality, but by definitional fiat. Multi-attribute constructs are declared unidimensional because arithmetic requires them to be so. Health-related quality of life becomes a single attribute by institutional decree, not by measurement demonstration.

The falsification items complete the picture. AMCP endorses, at a rhetorical level, the principle that non-falsifiable claims should be rejected at p = 0.75 (logit +1.10). Yet it simultaneously endorses the belief that reference-case simulations generate falsifiable claims at p = 0.90 (logit +2.20). This is not a subtle inconsistency. Simulation outputs are conditional projections derived from assumptions, many of which are themselves non-measures. Sensitivity analysis explores model behavior; it does not expose claims to empirical refutation. By treating simulations as falsifiable, AMCP substitutes model coherence for scientific risk.

This substitution is the hallmark of a mature memeplex. As Dawkins observed, memeplexes persist not because they are true, but because they are internally reinforcing and externally insulated. The AMCP belief system exhibits precisely these properties. Measurement axioms that would threaten the system are excluded. Arithmetic practices that sustain it are reinforced. Professional training, dossier standards, and formulary conventions transmit the memeplex intact from one generation of managed care professionals to the next. Internal debate is unnecessary because the boundaries of admissible thought are already fixed.

The consequences of this belief system are profound and immediate. Managed care decisions on access, step therapy, prior authorization, and pricing are anchored to numbers that cannot, in principle, be defended as measures. Patients experience delays or denials of care based on modeled value claims that are not falsifiable. Manufacturers are pressured into pricing negotiations

grounded in thresholds derived from pseudo-measurement. Pharmacists and payers are placed in the position of enforcing decisions whose numerical justification cannot survive scientific scrutiny.

What distinguishes AMCP from other HTA actors is not that it is worse, but that it is operational. ICER produces reports; AMCP produces practice. By embedding arithmetic without measurement into the daily mechanics of formulary decision-making, AMCP converts epistemic failure into routine governance. The memeplex becomes policy.

If AMCP were to confront this diagnosis honestly, the implications would be disruptive but straightforward. Measurement would be restored as a gatekeeping condition. Manifest claims would be restricted to linear ratio measures. Latent traits would require Rasch logit ratio measurement with demonstrated invariance. Composite utility indices would be reclassified as descriptive profiles. Simulation outputs would be acknowledged as conditional narratives, not evidence. Until such changes occur, AMCP will remain what the 24-item profile reveals: a professional institution that has perfected the practice of arithmetic without measurement, and in doing so has given that practice authority over patient access and pharmaceutical pricing.

The probabilities and logits leave no room for reinterpretation. This is not a case of partial misunderstanding or transitional confusion. It is a fully normalized belief system. The managed care memeplex is intact, self-protecting, and profoundly incompatible with the evolution of objective knowledge.

## AMCP AND THE ABANDONMENT OF FALSIFICATION IN MANAGED CARE

The endorsement by the Academy of Managed Care Pharmacy of the contemporary HTA memeplex does not merely suggest indifference to falsification and the evolution of objective knowledge; it strongly indicates that these goals have been functionally displaced by a different institutional priority. That priority is not the discovery of truth about therapy impact, but the production of administratively usable numbers that can support coverage decisions, pricing negotiations, and formulary control. Within this framework, falsification is not an operational objective but a rhetorical ornament, invoked to preserve the appearance of scientific legitimacy while remaining structurally irrelevant to decision making.

At the core of the HTA memeplex endorsed by Academy of Managed Care Pharmacy is a reversal of the epistemic sequence that defines normal science. In scientific inquiry, claims are formulated, measured, and exposed to empirical refutation, with surviving claims contributing incrementally to objective knowledge. In managed care HTA practice, by contrast, arithmetic outputs are generated first and treated as authoritative, while questions of measurement validity and empirical falsification are deferred, bracketed, or ignored. This inversion is not accidental. It reflects a system designed to function without the possibility that its central claims could be shown to be false in a meaningful sense.

Falsification, properly understood, is incompatible with the dominant analytic instruments of managed care HTA. Claims derived from QALYs, composite utility indices, and reference-case simulation models are not falsifiable because they do not assert empirical relationships that can be

tested against observed reality. They assert conditional projections: if certain assumptions hold, then certain modeled outcomes follow. Sensitivity analysis does not rescue these claims. Varying assumptions explores the internal behavior of a model; it does not expose the claim to the risk of being wrong. Yet within the AMCP knowledge base, robustness to scenario variation is routinely treated as a proxy for scientific credibility. Falsification is redefined as internal consistency.

This redefinition has profound implications. If claims cannot be falsified, they cannot be improved through empirical challenge. They cannot be replicated in the strong sense, because there is no invariant quantity to reproduce. Disagreement does not lead to experimental testing but to negotiation over assumptions, time horizons, discount rates, and thresholds. Knowledge does not evolve; it accretes. Each new model adds another layer of numerical storytelling without eliminating any prior error. The endorsement of this structure by AMCP signals that health systems do not expect HTA claims to be wrong in a way that matters.

The same conclusion follows from the treatment of measurement. Objective knowledge requires measures that remain invariant across persons, settings, and time. Without invariance, numerical change cannot be distinguished from artifact. Yet the HTA memeplex endorsed by AMCP systematically rejects the measurement axioms that would make invariance possible. Unidimensionality is weakly endorsed. The requirement that multiplication be restricted to ratio scales is rejected. The precedence of measurement over arithmetic is denied. Most decisively, Rasch measurement, the only framework capable of transforming subjective responses into invariant measures, is excluded entirely. Latent traits are not measured; they are scored, summed, and monetized.

This exclusion is not an oversight. It is an institutional necessity. Accepting Rasch measurement would force health systems to confront the fact that most patient-reported outcome data cannot support arithmetic operations, aggregation, or pricing benchmarks. It would require abandoning QALYs, ICERs, and long-horizon reference-case simulations as evidentiary claims. For organizations tasked with managing budgets and controlling access, such a shift would be destabilizing. The AMCP endorsement therefore reflects a rational adaptation to institutional constraints: measurement discipline is sacrificed to preserve decision-making convenience.

From this perspective, the absence of interest in falsification is not a failure of scientific virtue but a feature of system design. Health systems do not want claims that can be falsified because falsifiable claims create uncertainty, instability, and accountability. A falsifiable claim can be wrong, and if it is wrong, decisions based on it can be challenged. In contrast, non-falsifiable numerical constructs provide insulation. They can always be defended by adjusting assumptions, extending horizons, or invoking precedent. Responsibility diffuses into process.

The appeal to pragmatism often offered in defense of this posture does not withstand scrutiny. It is said that health systems need tools that work, not philosophical purity. Yet falsification and measurement are not philosophical luxuries; they are the conditions under which claims can be meaningfully said to work. A system that rejects these conditions does not become pragmatic; it becomes epistemically closed. It can act, but it cannot learn. Its outputs may change, but they cannot improve in the scientific sense.

The unanimity of endorsement across managed care pharmacy, academic HTA, and payer-facing organizations reinforces this conclusion. There is no sustained internal debate about measurement axioms because the memeplex itself suppresses such debate. As Dawkins described, successful memeplexes protect themselves by marginalizing ideas that threaten their coherence. Representational measurement theory and falsification are not ignored because they are unknown; they are excluded because their acceptance would unravel the evaluative architecture on which managed care depends.

The result is a health system evaluation culture that is numerically sophisticated but epistemically inert. Claims about therapy value circulate, influence access and pricing, and shape patient experience, yet they do not converge toward truth. They cannot be falsified, and therefore cannot be corrected. AMCP's endorsement of the HTA memeplex is thus best understood as an explicit signal that health systems have deprioritized the evolution of objective knowledge in favor of administratively stable arithmetic. What is gained is control. What is lost is science.

## WHAT WOULD BE THE KEY RECOMMENDATIONS FOR REVISING THE 2024 AMCP FORMAT FOR FORMULARY SUBMISSIONS

Revising the 2024 AMCP Format for Formulary Submissions requires confronting a problem that has been allowed to persist for decades: the Format has become a sophisticated administrative template for organizing claims rather than a scientific framework for evaluating them [5]. Its structure encourages completeness, comparability, and procedural transparency, yet it remains largely silent on whether the claims being presented satisfy the conditions required for measurement, falsification, and replication. A meaningful revision must therefore shift the Format from a documentation standard to a measurement standard.

The first and most important recommendation is that the Format explicitly require that every value claim be tied to a declared measurement scale type. At present, clinical, economic, and patient-reported outcomes are presented side by side as if they were commensurable forms of evidence. A revised Format should require manufacturers to state whether each outcome is a manifest attribute measured on a linear ratio scale, a latent attribute measured on a Rasch logit ratio scale, or a descriptive construct that does not meet measurement requirements. This single change would immediately expose which claims can support arithmetic and which cannot, and would prevent the implicit treatment of ordinal or composite outcomes as quantitative evidence.

Closely related to this is the need to abandon composite endpoints as decision variables. The current Format encourages the presentation of cost-effectiveness ratios, quality-adjusted life-years, and other composite constructs without requiring demonstration of dimensional homogeneity. A revised Format should prohibit the use of composite outcomes as primary value claims unless the manufacturer can demonstrate that all components share the same attribute and scale properties. Where this cannot be demonstrated, such constructs should be reclassified as illustrative summaries rather than evaluable claims. This would not prevent their discussion, but it would strip them of evidentiary authority.

A third recommendation is to replace model-centric evaluation with claim-centric evaluation. The 2024 Format devotes extensive attention to economic models, scenario analyses, and sensitivity

testing, implicitly treating models as generators of evidence. A revised Format should reverse this logic. The focus of evaluation should rest on clearly specified, empirically testable claims, each supported by a protocol defining the target population, outcome measure, timeframe, comparator, and evaluation method. Without this structure, the Format continues to privilege numerical storytelling over empirical assessment.

Patient-reported outcomes represent a critical area requiring reform. The current Format allows summated questionnaire scores, mapped utilities, and responder thresholds to be presented as quantitative evidence without scrutiny of whether the underlying instruments produce measures. A revised Format should require that any latent-trait claim intended to support decision making be derived from a Rasch-calibrated instrument with demonstrated unidimensionality and invariance. Instruments that do not meet these requirements may still be reported descriptively, but they should not be used to support arithmetic comparisons, modeling, or pricing arguments. This would align patient-centered evaluation with the only scientifically defensible approach to latent trait measurement.

Another essential revision concerns the treatment of time. The Format currently treats time as a neutral multiplier applied to outcomes of uncertain scale type. A revised version should explicitly state that multiplication by time is permissible only when the outcome itself is expressed on a ratio scale. This would immediately invalidate routine practices such as multiplying ordinal utilities by survival time to generate QALYs, without requiring philosophical debate. The restriction follows directly from elementary measurement principles and would restore coherence to temporal reasoning in formulary submissions.

The Format should also require explicit acknowledgment of uncertainty as epistemic, not merely statistical. Sensitivity analyses and probabilistic modeling are currently presented as substitutes for empirical testing. A revised Format should distinguish uncertainty arising from sampling variation from uncertainty arising from lack of measurement. Where an outcome is not measured, no amount of statistical manipulation can convert it into evidence. This distinction should be made explicit, so that decision makers are not misled into believing that precision estimates compensate for invalid constructs.

Transparency requirements should be strengthened in a different direction than at present. Rather than demanding ever more model detail, the Format should require manufacturers to disclose which commonly accepted HTA conventions they are compelled to use to satisfy external expectations, even when those conventions conflict with measurement standards. This would acknowledge the reality that manufacturers often comply with requirements they do not control. Such transparency would allow health systems to distinguish between claims made because they are scientifically defensible and claims made because they are procedurally required.

A revised Format should also encourage post-listing evaluation grounded in measurable outcomes. Instead of relying on lifetime projections, submissions should prioritize short- to medium-term claims that can be empirically assessed within meaningful timeframes, such as 6 or 12 months. These claims should be replicable within the health system using observable data or properly measured latent traits. This would transform formulary review from a one-time modeling exercise into an iterative evidence process capable of learning and correction.

Finally, the Format should incorporate an explicit commitment to falsification. Each value claim should be accompanied by a protocol describing what evidence would count as refutation. If no such evidence can be specified, the claim should not be classified as evaluable. This requirement alone would eliminate large classes of speculative modeling outputs that currently dominate submissions. It would also reorient managed care decision making toward the evolution of objective knowledge rather than the negotiation of competing projections.

Taken together, these recommendations do not call for greater analytical complexity. They call for disciplinary restraint. Revising the AMCP Format along these lines would not make submissions harder to compile; it would make them intellectually honest. It would allow manufacturers to present what can truly be known, allow health systems to evaluate claims that can be tested, and restore the distinction between evidence and illustration that has been lost in contemporary formulary practice. The 2024 Format represents an opportunity. If revised to incorporate representational measurement principles, it could become not merely a template for submission, but a foundation for scientific evaluation in managed care.

# 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.

**A NEW START IN MEASUREMENT FOR HEALTH TECHNOLOGY ASSESSMENT**

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

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

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

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

## DESIGNED FOR CLOSURE

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

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

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

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

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

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

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

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

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

**ACKNOWLEDGEMENT**

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

**REFERENCES**

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

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

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

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

[5] Academy of Managed Care Pharmacy**.** *AMCP Format for Formulary Submissions, Version 5.0. Journal of Managed Care & Specialty Pharmacy.* 2024; 30, 4 (Suppl:1–64)