MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

UNITED STATES: THE ABSENCE OF MEASUREMENT WITH THE VA PHARMACY BENEFITS MANAGEMENT SERVICES

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

LOGIT WORKING PAPER No 3 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 reports that follow provide the empirical confirmation of this claim. They show, with unsettling consistency, that the global HTA knowledge base neither possesses nor applies the principles of scientific measurement.

The objective of this study is to interrogate the epistemic foundations of the Veterans Affairs Pharmacy Benefits Management Services (VA PBM) as a decisive institutional actor in health technology assessment and formulary governance. Rather than evaluating individual formulary decisions or therapeutic recommendations, the analysis examines the belief system embedded in the quantitative claims the VA PBM accepts, reproduces, and operationalizes as decision-relevant evidence. Using a 24-item diagnostic grounded in representational measurement theory and expressed through a canonical probability-to-logit transformation, the study asks whether the numerical constructs relied upon by the VA PBM satisfy the axioms required for admissible arithmetic, falsification, and cumulative knowledge. The purpose is not to assess technical competence or policy intent, but to determine whether the VA PBM's evaluative architecture is measurement-literate or whether it institutionalizes arithmetic without measurement at the point where numerical claims have direct consequences for access, pricing, and clinical practice.

The findings are unequivocal. The VA PBM belief profile exhibits a systematic rejection of the axioms of representational measurement alongside near-ceiling endorsement of mathematically impossible claims required to sustain cost-effectiveness reasoning and model-based evaluation. Core principles involving measurement preceding arithmetic, the requirement of ratio scales for multiplication, unidimensionality as a prerequisite for quantitative claims, and the necessity of Rasch measurement for latent traits, are weakly endorsed or rejected outright. At the same time, false propositions embedded in conventional HTA practice, including the ratio status and aggregability of QALYs and the legitimacy of summated ordinal scores as quantitative measures, are strongly reinforced. The resulting logit structure is not one of ambiguity or compromise but of inversion: arithmetic is treated as authoritative, while measurement is treated as optional. In contrast to academic or advisory bodies, this inversion is operationalized directly in policy, rendering the VA PBM one of the most consequential institutional expressions of measurement failure in contemporary health care.

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 VA PHARMACY BENEFITS KNOWLEDGE BASE

For the purposes of this analysis, the VA PBM knowledge base is defined as the recurrent and institutionally reinforced body of concepts, methods, and evaluative norms through which the organization constructs, interprets, and justifies quantitative claims about therapy impact. It is not defined by any single guideline, report, or committee decision, but by the patterned regularities that appear across formulary monographs, therapeutic class reviews, comparative effectiveness summaries, pharmacoeconomic evaluations, utilization management criteria, and internal decision frameworks over time. What unifies this knowledge base is not organizational authorship but functional role: the production and legitimization of numbers that guide access, substitution, and pricing decisions within the VA health system.

The analytic boundaries of this knowledge base encompass the routine use of cost-utility reasoning, preference-based utility instruments, QALYs, and reference-case or quasi-reference-case modeling as admissible quantitative inputs to decision making. These practices are accompanied by the widespread acceptance of summated patient-reported outcome scores, mapped utilities, and composite indices as if they possessed interval or ratio properties, despite their ordinal origins. Within this ecosystem, statistical techniques such as regression modeling, sensitivity analysis, and uncertainty intervals are routinely applied to quantities whose measurement status is never established, implicitly treating statistical manipulation as a substitute for measurement validation.

Equally important are the structural absences that define the knowledge base. Representational measurement theory does not appear as a governing constraint on admissible claims. Scale-type requirements are not enforced as gatekeeping conditions for arithmetic. Rasch measurement, despite its relevance to the VA PBM's heavy reliance on latent-trait constructs such as quality of life, symptom burden, and functional status, is not adopted as a required standard. Latent traits are discussed in terms of score changes and thresholds rather than possession on invariant scales, allowing subjective observations to be numerically mobilized without being measured.

The VA PBM knowledge base is therefore characterized not by explicit rejection of measurement theory, but by patterned indifference to it. Quantitative legitimacy is conferred through institutional acceptance rather than through satisfaction of axioms. Numbers become authoritative because they are embedded in formal processes, not because they represent quantities with invariant units or meaningful zero points. In this sense, the knowledge base is behavioral and structural rather than philosophical. It reflects what the VA PBM repeatedly does, defends implicitly, and treats as acceptable evidence in practice. The 24-item diagnostic is applied accordingly, not as a survey of individual beliefs, but as a probe of the epistemic architecture within which VA PBM decisions are generated—and that architecture, as the findings demonstrate, is fundamentally incompatible with the requirements of scientific measurement.

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: VA PHARMACY BEBEFITS

Table 1 presents, the endorsement 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 VA PHARMACY BENEFITS

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	0.25	-1.10

MULTIPLICATION REQUIRES A RATIO MEASURE	1	0.15	-1.75
TIME TRADE-OFF PREFERENCES ARE UNIDIMENSIONAL	0	0.80	+1.40
RATIO MEASURES CAN HAVE NEGATIVE VALUES	0	0.85	+1.75
EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES	0	0.85	+1.75
THE QALY IS A RATIO MEASURE	0	0.85	+1.75
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	0.85	+1.75
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	0.10	-2.20
TRANSFORMING SUBJECTIVE RESPONSES TO INTERVAL MEASUREMENT IS ONLY POSSIBLE WITH RASH RULES	1	0.10	-2.20
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	+1.10
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.85	+1.75
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	0.10	-2.20

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	0.10	-2.20

REVIEW: THE MEASUREMENT FAILURE OF VA PHARMACY BENEFITS MANAGEMENT SERVICES

The Veterans Affairs Pharmacy Benefits Management Services occupies a position of exceptional consequence within the American health system. Unlike academic HTA centers, journals, or advisory bodies whose influence is mediated through persuasion, the management services exerciss direct authority over formulary access, therapeutic substitution, utilization controls, and effective pricing for a population that exceeds nine million veterans. Its determinations are not abstract exercises in methodological preference. They are operational decisions with immediate clinical and distributive consequences. For that reason alone, the epistemic foundations of VA PBM decision making should be held to a higher standard than those applied to academic discourse or private advisory organizations. The 24-item diagnostic demonstrates that this higher standard is not met. What emerges instead is a belief system that mirrors, and in several respects intensifies, the core failures of the health technology assessment memeplex: the systematic rejection of representational measurement axioms alongside the near-ceiling endorsement of mathematically impossible arithmetic.

The defining characteristic of the belief profile is the inversion of scientific order. The proposition that measurement must precede arithmetic is endorsed at $p = 0.15$, corresponding to a canonical logit of -1.75. This is not marginal ambivalence. It places the principle firmly in the rejection region. The implication is direct and far-reaching: within the VA PBM evaluative framework, arithmetic is permitted without first establishing that the quantities being manipulated are measures at all. Numbers are allowed to do epistemic work simply because they are numbers. Measurement is treated as optional background rather than as a necessary precondition. This inversion is not an accidental omission. It is the enabling condition for everything that follows.

Once arithmetic is liberated from measurement, the system becomes free to endorse propositions that are otherwise indefensible. The aggregation of QALYs is endorsed at $p = 0.95$, yielding a logit of +2.50, the maximum value on the scale. The belief that the QALY is a ratio measure is endorsed at $p = 0.85$, logit +1.75. The belief that EQ-5D preference algorithms create interval measures is also endorsed at $p = 0.85$, logit +1.75. These endorsements are not peripheral assumptions. They are the load-bearing commitments that sustain cost-effectiveness reasoning within the VA PBM. Without them, incremental cost-effectiveness ratios collapse, thresholds lose coherence, and model-based comparisons cease to be interpretable. The VA PBM resolves the

resulting contradiction not by revisiting arithmetic, but by rejecting the axioms that would constrain it.

This pattern is made explicit by the treatment of ratio arithmetic itself. The proposition that multiplication requires a ratio measure is endorsed at p = 0.15, logit −1.75. This places the requirement well below neutrality. In effect, the VA PBM denies the condition under which cost can be divided by outcome. Yet cost per QALY reasoning remains embedded in formulary deliberations, therapeutic comparisons, and policy guidance. The arithmetic survives precisely because the rule that would invalidate it has been excluded. This is not a technical oversight. It is an institutional choice to privilege computational output over measurement legitimacy.

The treatment of subjective data exposes the most consequential failure. The VA PBM strongly endorses the belief that summation of Likert-type question scores creates a ratio measure at p = 0.90, logit +2.20. It likewise endorses the belief that summations of subjective instrument responses are ratio measures at p = 0.85, logit +1.75. These propositions are mathematically false. Ordinal categories do not acquire equal intervals, invariance, or a true zero through addition. No amount of internal consistency, reliability estimation, or factor modeling can change that fact. Endorsing these propositions at near-ceiling strength means that the VA PBM treats patient-reported outcomes as if they were already measured quantities suitable for multiplication, aggregation, and optimization. The epistemic damage is immediate and irreversible. Once ordinal scores are granted ratio status by fiat, there is no arithmetic operation that cannot be justified.

Unidimensionality, the defining requirement for any quantitative scale, fares no better. The belief that measures must be unidimensional is endorsed at p = 0.25, logit −1.10, placing it in the rejection region. At the same time, the belief that time trade-off preferences are unidimensional is endorsed at p = 0.80, logit +1.40. This contradiction reveals how unidimensionality functions within the VA PBM belief system. It is not a property to be demonstrated empirically. It is an assumption to be asserted when required by arithmetic and ignored when inconvenient. Multiattribute constructs are treated as single quantities because arithmetic demands them to be so, not because the attributes have been shown to form an invariant unidimensional structure.

This selective treatment of unidimensionality is inseparable from the VA PBM's reliance on composite preference-based instruments. Health-related quality of life is routinely treated as a single attribute despite being constructed from heterogeneous domains. The belief that the QALY is dimensionally homogeneous is endorsed at p = 0.85, logit +1.75. This endorsement directly contradicts the rejection of unidimensionality as a general requirement. The contradiction is resolved institutionally by redefining dimensionality rather than by enforcing it. Dimensional homogeneity becomes a rhetorical label rather than a measurable property.

The Rasch block of the diagnostic exposes the epistemic boundary of the VA PBM with no possibility of charitable reinterpretation. Every Rasch-related proposition collapses toward the floor of the scale. The belief that there are only two admissible classes of measurement, linear ratio scales for manifest attributes and Rasch logit ratio scales for latent traits, is endorsed at p = 0.10, logit −2.20. The belief that transforming subjective responses to interval measurement is only possible with Rasch rules is endorsed at p = 0.10, logit −2.20. The belief that the Rasch logit ratio scale is the only basis for assessing therapy impact for latent traits sits at the same value. The belief

that Rasch rules are identical to the axioms of representational measurement theory is likewise endorsed at p = 0.10, logit −2.20. These values indicate categorical exclusion. Rasch measurement is not marginally disfavored. It is structurally rejected.

This rejection is decisive because Rasch is not one psychometric option among many. It is the only framework capable of transforming ordinal observations into invariant measures suitable for arithmetic. By rejecting Rasch while endorsing summated ordinal scoring as ratio measurement, the VA PBM embraces the consequences of measurement without accepting its discipline. Latent traits are invoked, scored, averaged, and monetized without ever being measured. Patient experience is acknowledged rhetorically and abused numerically.

The marginalization of latent trait possession further clarifies this posture. The proposition that the outcome of interest for latent traits is possession of that trait is endorsed at p = 0.25, logit −1.10. This weak endorsement indicates a preference for talking about changes in scores, differences in means, and responder thresholds rather than confronting the substantive question of how much of a trait a population actually possesses. Possession is epistemically dangerous because it demands measurement. Scores are epistemically safe because they do not. The VA PBM's belief system chooses safety over science.

The treatment of falsification completes the picture. The VA PBM endorses the proposition that non-falsifiable claims should be rejected at p = 0.75, logit +1.10, aligning itself rhetorically with scientific norms. At the same time, it endorses the belief that reference-case simulations generate falsifiable claims at p = 0.85, logit +1.75. This is a direct contradiction. Simulation outputs are conditional projections derived from assumptions, many of which rest on non-measures. Sensitivity analysis explores the internal behavior of a model; it does not expose claims to empirical refutation. By treating simulations as falsifiable, the VA PBM grants them epistemic authority while insulating them from being wrong. Robustness across scenarios substitutes for exposure to reality.

What distinguishes the VA PBM from other nodes in the HTA ecosystem is not the structure of its belief system, which closely mirrors that of academic HTA centers and advisory bodies, but the consequences of that belief system. The VA PBM converts arithmetic without measurement directly into policy. Veterans are denied, delayed, or switched therapies based on ratios that cannot be defended as ratios, aggregates that cannot be defended as aggregates, and models that cannot be falsified. This is not numerical storytelling in the abstract. It is numerical governance.

The severity of this failure cannot be dismissed as pragmatic compromise. Measurement axioms do not constrain decisions; they constrain claims. They do not prevent action; they prevent pretending that something has been measured when it has not. A federal agency that rejects these axioms does not become flexible. It becomes unaccountable. When arithmetic is detached from measurement, there is no principled way to say that one therapy produces more benefit than another, that one policy improves outcomes more than an alternative, or that a threshold has any empirical meaning. Decisions continue to be made, but they are no longer anchored to quantities that can support falsification or cumulative knowledge.

The VA PBM profile therefore represents one of the most severe cases of institutionalized measurement failure documented to date. It combines near-ceiling endorsement of mathematically impossible propositions with near-floor rejection of the axioms that would prohibit them. Rasch measurement is excluded. Possession of latent traits is marginalized. Unidimensionality is asserted when convenient and ignored when not. Aggregation is celebrated without dimensional justification. Simulation is treated as evidence. Arithmetic reigns supreme.

Until the VA PBM accepts that measurement must precede arithmetic, that only linear ratio and Rasch logit ratio scales are admissible, and that latent traits can only be quantified as possession on a Rasch scale, its decisions will remain mathematically elaborate and scientifically indefensible. The tragedy is not merely methodological. It is ethical. A system charged with serving veterans has adopted a belief structure that confuses scoring with measurement and treats numerical output as justification. That is not evidence-based policy. It is bureaucratic numeracy masquerading as science.

WHY DID THE VA PBM ADOPT THIS ANALYTICAL FRAMEWORK

The short answer is that the VA did not so much *choose* this analytical framework as *inherit* it, normalize it, and then harden it into policy through institutional imitation. The longer answer explains why ICER, and bodies like it, became an attractive template despite their lack of measurement legitimacy.

First, the VA PBM operates under an unusually strong mandate to demonstrate rational stewardship of public funds. Unlike private payers, it must show that its decisions are systematic, defensible, and procedurally fair across a national system. That requirement creates a powerful demand for a single, portable numerical language that can be applied consistently across disease areas, therapies, and time. Cost-utility analysis, QALYs, and model-based comparisons appear to satisfy that demand because they promise commensurability: different interventions can be compared on a common scale. The problem, as your work shows, is that this promise rests on false measurement. But from an institutional perspective, the appearance of commensurability is often mistaken for scientific rigor.

Second, ICER provided the VA the operationalized reference-case apparatus that translated HTA conventions into concrete policy signals. By the time the VA began engaging with ICER's work, ICER had already packaged cost-utility modeling, threshold reasoning, and scenario analysis into a format that looked policy-ready. It did not merely argue for QALYs; it showed how to use them to generate price benchmarks, value classifications, and comparative rankings. For an organization like the VA PBM, this was attractive not because it solved the measurement problem, but because it solved an administrative one: how to justify difficult access and pricing decisions using a standardized external logic.

Third, the VA's engagement with ICER reflects a broader phenomenon of epistemic outsourcing. Rather than adjudicating foundational methodological questions internally, questions about scale type, latent trait measurement, or the admissibility of arithmetic, the VA effectively deferred those questions to an external authority that presented itself as methodologically sophisticated. ICER's reference-case model, peer review process, and academic affiliations created the impression that

the hard epistemic work had already been done. In reality that work was never done; it was bypassed. But once bypassed by a perceived authority, it became easier for the VA to treat the framework as settled science.

Fourth, there is a strong path-dependence effect. Once an organization commits to cost-utility reasoning, it must continue to endorse the assumptions that make it possible. If the VA were to question whether utilities are interval measures, whether QALYs are ratio-scaled, or whether summated ordinal scores can support arithmetic, it would not merely be revising a technical detail. It would be undermining the legitimacy of years of prior decisions. Institutional self-preservation therefore favors methodological continuity over epistemic correction. Following ICER was not just imitation; it was a way of locking in a shared belief system that diffuses responsibility.

Fifth, the VA's analytic culture has been shaped by decades of interaction with academic health economics, pharmacy schools, and outcomes research centers that already operate within the same memeplex. ICER did not introduce false measurement into the VA; it merely consolidated and formalized it. The same assumptions about utilities, QALYs, and modeling had long circulated in the academic literature that VA analysts were trained in and drew upon. ICER functioned as a focal point, not an origin. Engaging with ICER therefore felt natural because it aligned with the knowledge base the VA already possessed.

Finally, there is a crucial political dimension. Model-based cost-effectiveness analysis provides plausible deniability. When access is restricted or prices are challenged, decision makers can point to models, thresholds, and external assessments rather than to discretionary judgment. This is especially valuable in a public system. ICER's framework offers a way to convert contested value judgments into apparently objective numerical outputs. The fact that those outputs are non-falsifiable and built on non-measures is invisible to stakeholders who lack training in representational measurement theory. What matters institutionally is not falsifiability, but defensibility.

In that sense, the VA adopted this framework not because it is scientifically sound, but because it **is** institutionally convenient, socially legitimized, and administratively scalable**.** ICER served as a catalyst and a shield, allowing the VA to align itself with a broader HTA orthodoxy while avoiding direct engagement with the measurement axioms that would dismantle that orthodoxy. The tragedy, which your work exposes, is that this choice substitutes procedural coherence for scientific validity, turning an organization with enormous real-world impact into an enforcer of arithmetic without measurement.

DOES THE VA PHARMACY BENEFITS MANAGEMENT HAVE A FUTURE?

The VA PBM *can* have a future, but not the one it currently inhabits. Its future depends entirely on whether it continues to anchor itself to the ICER-style reference-case paradigm or whether it decisively breaks from it and reclaims a role grounded in scientific measurement, falsifiable claims, and mission-aligned decision making.

The present review established the core fact: the VA PBM did not independently evolve its analytical framework. It imported it. By aligning itself with ICER's reference-case methodology, the VA PBM adopted a ready-made belief system that treats utilities, QALYs, and long-horizon simulations as admissible evidence while quietly abandoning representational measurement axioms. The 24-item diagnostic shows that this adoption was not superficial. The same inversion appears: arithmetic without measurement, modeling without falsification, and latent traits treated as if they were already quantified. That creates a structural contradiction at the heart of the VA PBM's mandate.

The VA is not a commercial payer. It is not a price-setting proxy for private markets. It is a vertically integrated public health system with a defined population, longitudinal care responsibility, and direct accountability for outcomes experienced by veterans. In that context, the ICER reference case is not merely flawed; it is *misaligned*. Reference-case simulations were designed to produce hypothetical lifetime value estimates for generalized decision makers. The VA PBM, by contrast, has access to real patients, real utilization, real adherence behavior, and real outcomes over meaningful time horizons. To rely on imaginary QALYs when possession-based outcomes and manifest resource claims are observable is not prudence; it is abdication.

If the VA PBM continues on its current path, its future is narrow and brittle. It becomes an internal replicator of the ICER memeplex: issuing coverage and formulary decisions justified by non-falsifiable constructs, defended procedurally rather than scientifically, and insulated from empirical challenge by appeal to "best practice." In that future, the VA PBM gradually loses epistemic legitimacy. Its analyses cannot be replicated in the strong sense, cannot be falsified, and cannot generate cumulative knowledge. Decisions become increasingly opaque, contested, and vulnerable to political rather than scientific scrutiny. That is not a stable future. The alternative future is radically different and far more powerful.

If the VA PBM were to accept the measurement critique, it would be uniquely positioned to lead a post-QALY transformation. Unlike ICER, the VA does not need to pretend that utilities measure health. It can demand single-attribute, evaluable claims tied to observable outcomes. For manifest attributes, that means linear ratio measures: hospital days avoided, events prevented, time to progression, resource use. For latent attributes, that means Rasch logit ratio measures of possession, developed prospectively and tested for invariance in the veteran population. Claims would be time-bounded, protocol-driven, and empirically revisable.

In that future, formulary decisions would no longer be justified by reference-case projections but by *trackable performance*. Manufacturers would be required to submit claims that can fail. Therapies would be evaluated on whether they actually deliver improvements in possession or resource outcomes over defined intervals. Knowledge would accumulate because it could be corrected. Crucially, this future aligns with the VA's ethical and institutional mission. Veterans are not abstract utility carriers. They are patients whose experiences, functioning, and outcomes matter in concrete terms. A system that measures possession rather than scores, and performance rather than modeled value, is not only more scientific; it is more defensible.

The answer is conditional but clear. If the VA PBM remains tethered to ICER's reference-case framework, it has no independent future. It becomes a derivative institution, borrowing authority

from a model that itself cannot survive sustained measurement scrutiny. But if the VA PBM rejects arithmetic without measurement and rebuilds its evaluative framework around falsifiable, protocol-based claims, it could become the first large health system to demonstrate what *normal science* in formulary decision making actually looks like.

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