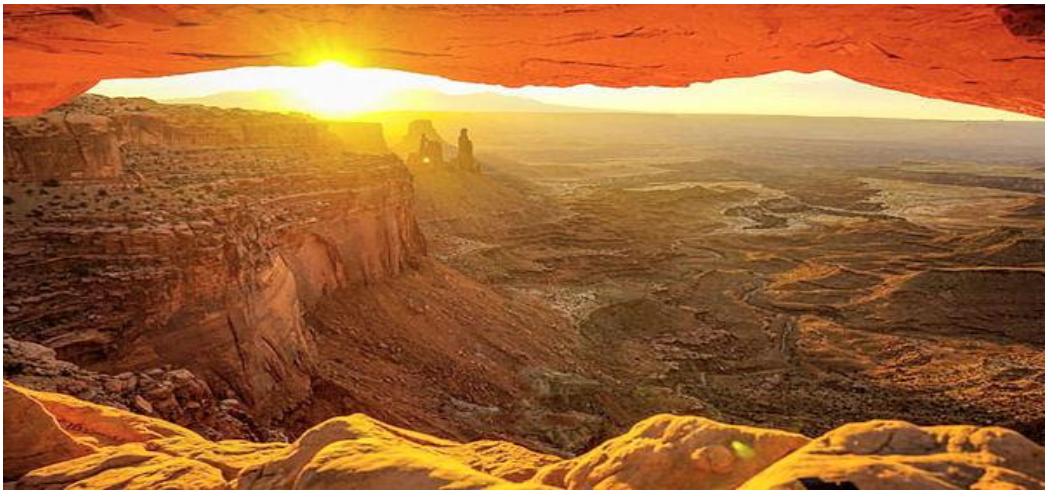
MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION



REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

NEW ZEALAND: PHARMAC ENDORSES FALSE MEASUREMENT FOR HEALTH TECHNOLOGY ASSESSMENT

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

LOGIT WORKING PAPER No 1261 APRIL 2026

www.maimonresearch.com

Tucson AZ

FOREWORD

HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

PHARMAC is New Zealand's national agency responsible for the assessment, prioritization, and funding of pharmaceuticals and, increasingly, other health technologies. Its central objective is to secure the best possible health outcomes for the population within a fixed budget by determining which medicines and related interventions should be publicly funded. To achieve this, PHARMAC integrates clinical evaluation, economic assessment, and policy decision-making within a single institutional framework, using a set of decision criteria that include health need, clinical benefits and risks, cost-effectiveness, and the impact on health system resources.

What distinguishes PHARMAC within the global HTA landscape is the degree to which assessment and decision-making are centralized. Unlike jurisdictions where independent academic units, advisory bodies, and agencies contribute separately to evidence generation and appraisal, PHARMAC operates as a unified structure in which the evaluation of evidence and the allocation of funding are closely coupled. Its methodological approach, set out in its pharmacoeconomic guidance, emphasizes cost-utility analysis, the use of quality-adjusted life years, and model-based projections to inform comparative value assessments. This centralization creates a closed analytical environment in which a single institutional knowledge base defines both how evidence is constructed and how it is applied in decision-making, with limited scope for independent methodological challenge or alternative frameworks within the national system.

The objective of this study is to undertake a structured interrogation of the HTA knowledge base associated with PHARMAC using a standardized 24-item canonical diagnostic instrument grounded in the axioms of representational measurement theory. The intent is not to assess individual funding decisions or specific analyses, but to determine whether the conceptual framework that underpins PHARMAC's methods satisfies the minimum conditions required for scientific measurement. In particular, the study evaluates whether the knowledge base recognizes the requirements for admissible arithmetic operations, including unidimensionality, invariance, the existence of a true zero, and the necessity that latent constructs be measured through Rasch-conformant instruments. By assigning categorical probabilities to each statement and transforming these into normalized logits, the analysis generates a profile of conceptual endorsement that allows an explicit assessment of whether PHARMAC's framework supports evaluable, falsifiable value claims.

The findings are unequivocal. The PHARMAC knowledge base exhibits a pronounced pattern of measurement inversion. Statements that are true within representational measurement theory are weakly endorsed or rejected, while statements that are false—particularly those that underpin cost-utility analysis—are strongly endorsed, often at maximum logit levels. The QALY is treated as a ratio measure, assumed to be aggregable, and regarded as dimensionally coherent, while the requirement that multiplication demands a ratio scale is rejected. At the same time, Rasch measurement is entirely absent, indicating that the only scientifically defensible pathway for constructing measures of latent constructs is not recognized. The overall profile demonstrates that

PHARMAC's HTA framework does not meet the axioms of representational measurement and therefore supports the generation of non-evaluable claims.

The objective of this study is to interrogate the epistemic foundations of the Pharmaceutical Benefits Advisory Committee (PBAC) as Australia's national authority for pharmaceutical reimbursement and pricing. Rather than evaluating individual PBAC decisions or specific submissions, the analysis examines the belief system embedded in the analytical framework that PBAC requires and enforces. Using a 24-item diagnostic grounded in representational measurement theory, the study evaluates whether the numerical constructs central to PBAC decision making, utilities, QALYs, cost-effectiveness ratios, and reference-case simulation outputs satisfy the axioms necessary for admissible arithmetic, falsification, and the evolution of objective knowledge. The purpose is not to assess policy outcomes, but to determine whether the PBAC framework itself rests on measurable quantities or on numerical conventions that cannot, in principle, support scientific evaluation.

This assessment is particularly important given PBAC's institutional role. As the gatekeeper to national reimbursement, PBAC does not merely consume health technology assessment evidence; it shapes the entire Australian HTA ecosystem. Its requirements determine how manufacturers construct submissions, how academic centers train analysts, how consultants design models, and how journals define acceptable evidence. The study therefore treats PBAC not as a passive decision body, but as a central epistemic authority whose analytical standards define what counts as "evidence" in Australian pharmaceutical policy.

The findings are unequivocal. The PBAC knowledge base exhibits a systematic inversion of scientific reasoning in which arithmetic is authorized independently of measurement. Core axioms of representational measurement including the precedence of measurement over arithmetic, the requirement of ratio scales for multiplication, the necessity of unidimensionality and the admissibility conditions for latent attributes collapse to the floor or near-floor of endorsement. At the same time, propositions that enable cost-utility modeling, including the treatment of ordinal utilities as interval or ratio measures, the aggregation of QALYs, and the legitimacy of reference-case simulations, rise toward the ceiling of endorsement.

This pattern does not reflect inconsistency or partial misunderstanding. It reflects a coherent belief structure in which numerical plausibility substitutes for measurement validity. Rasch measurement, the only framework capable of producing invariant measures for latent attributes, is effectively absent from PBAC's analytical foundations. As a result, patient-reported outcomes and preference-based measures are treated as quantitative without ever satisfying the conditions required for quantity. The PBAC framework therefore cannot support falsifiable claims, cumulative learning, or empirical correction. It functions as an administrative decision mechanism rather than as a scientific evaluative system.

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 PHARMAC KNOWLEDGE BASE

The HTA knowledge base associated with PHARMAC is best understood as a centralized and internally coherent methodological framework that integrates evidence generation, economic evaluation, and policy decision-making within a single institutional structure. Its core features are defined in its pharmacoeconomic guidance, which places cost-utility analysis at the center of evaluation, with benefits expressed in terms of quality-adjusted life years derived from preference-based instruments such as the EQ-5D. These utility values are combined with time and incorporated into economic models that project costs and outcomes over extended horizons. The resulting outputs—typically expressed as cost per QALY—are used to inform prioritization decisions within a fixed budget constraint. This framework is reinforced through standardized submission requirements, internal analytical processes, and advisory committee review, creating a consistent approach to HTA across all funding applications.

What characterizes this knowledge base is the absence of any requirement to establish the measurement properties of the constructs it employs. Utilities derived from ordinal preference structures are treated as if they possess interval or ratio scale properties, enabling their use in arithmetic operations such as multiplication, aggregation, and comparison. There is no requirement to demonstrate unidimensionality, invariance, or the existence of a true zero, nor is there systematic use of Rasch modeling to transform subjective observations into valid measures of latent traits. Instead, the framework proceeds on the implicit assumption that numerical representation is sufficient to constitute measurement. This assumption allows the construction of QALYs and the application of cost-effectiveness modeling without addressing whether the underlying quantities meet the axioms of representational measurement.

The centralization of HTA within PHARMAC amplifies the implications of this omission. Because assessment and decision-making are integrated within a single agency, the methodological framework is not subject to independent academic challenge or competing evaluative approaches within the national system. The knowledge base thus functions as a closed analytical environment in which methodological assumptions are reinforced through repeated application. Statistical and modeling sophistication coexist with a lack of measurement validity, and the appearance of quantitative rigor substitutes for the demonstration of admissible scale properties. As a result, the framework generates outputs that are internally consistent but not empirically evaluable in a measurement-consistent sense. The PHARMAC knowledge base therefore exemplifies a system in which the conditions required for scientific measurement are neither recognized nor enforced, and where value claims are constructed and applied in policy decision-making without meeting the standards necessary for replication, falsification, or validation within normal science.

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: PHARMAC

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 PHARMAC

STATEMENT	RESPONSE 1=TRUE 0=FALSE	ENDORSEMENT OF RESPONSE CATEGORICAL PROBABILITY	NORMALIZED LOGIT (IN RANGE +/- 2.50)
INTERVAL MEASURES LACK A TRUE ZERO	1	0.20	-1.40
MEASURES MUST BE UNIDIMENSIONAL	1	0.15	-1.80
MULTIPLICATION REQUIRES A RATIO MEASURE	1	0.10	-2.20

TIME TRADE-OFF PREFERENCES ARE UNIDIMENSIONAL	0	0.90	+2.20
RATIO MEASURES CAN HAVE NEGATIVE VALUES	0	0.95	+2.50
EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES	0	0.95	+2.50
THE QALY IS A RATIO MEASURE	0	0.95	+2.50
TIME IS A RATIO MEASURE	1	0.80	+1.40
MEASUREMENT PRECEDES ARITHMETIC	1	0.15	-1.80
SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES	0	0.95	+2.50
MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC	1	0.10	-2.20
THERE ARE ONLY TWO CLASSES OF MEASUREMENT LINEAR RATIO AND RASCH LOGIT RATIO	1	0.05	-2.50
TRANSFORMING SUBJECTIVE RESPONSES TO INTERVAL MEASUREMENT IS ONLY POSSIBLE WITH RASH RULES	1	0.05	-2,50
SUMMATION OF LIKERT QUESTION SCORES CREATES A RATIO MEASURE	0	0.95	+2.50
THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE	0	0.95	+2.50
CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.10	-2.20
QALYS CAN BE AGGREGATED	0	0.95	+2.50
NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.40	-0.40
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.95	+2.50
THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.70	+0.85
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	0.90	+2.20

THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT	1	0.05	-2.50
THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.05	-2.50

REVIEW: PHARMAC EMBRACES FALSE MEASUREMENT

PHARMAC's own published methods make the structure of this profile unsurprising. Its guidance explicitly centers cost-utility analysis, defines outcomes in QALY**s**, discusses cost/QALY gained, and points applicants to a formal pharmacoeconomic analysis framework that uses modeled projections and utility-based reasoning. PHARMAC documentation also specifically refers to the use of EQ-5D utility weights for health-related quality of life.

The interrogation of the HTA knowledge base associated with PHARMAC in New Zealand yields one of the clearest possible profiles of measurement inversion. This is not an incidental slippage in terminology, a local methodological preference, or a minor technical weakness that can be repaired by better data inputs or more sophisticated models. The profile points to a structured and institutionalized commitment to false measurement. The pattern is stark. Statements that are false within the axioms of representational measurement are endorsed at very high levels, often at the maximum normalized logit. Statements that are true within representational measurement are weakly endorsed, rejected, or effectively absent. The resulting asymmetry places PHARMAC securely within the global HTA memeplex, but with a profile that is particularly forceful because PHARMAC is not simply an academic center or a journal. It is a national decision-making body whose methods have direct consequences for reimbursement, access, and resource allocation.

The starting point for understanding the PHARMAC profile is the agency's explicit and long-standing reliance on cost-utility analysis. Its published documentation is direct: cost-utility analysis is recommended, benefits are measured in QALYs, and cost/QALY gained is a core expression of results. PHARMAC's methodological materials also make clear that economic models, utility weights, and projected health benefits are integral to the analytical process. This is not a hidden assumption. It is openly declared. The question raised by the present interrogation is therefore not whether PHARMAC uses QALYs and utility instruments. It does. The question is whether the constructs employed meet the axioms of representational measurement that would justify the arithmetic operations imposed upon them. On that question, the logit profile is devastating.

The endorsement of the false propositions that sustain QALY arithmetic is near total. The claims that the QALY is a ratio measure, that the QALY is dimensionally homogeneous, and that QALYs can be aggregated are all endorsed at the maximum positive logit of +2.50. This is not a hesitant or qualified commitment. It is categorical. It means the PHARMAC knowledge base proceeds as though utilities derived from preference-based instruments can be multiplied by time, summed across persons, and compared as if they were valid quantities on a lawful scale. Yet this is exactly

what the axioms of representational measurement do not permit unless the underlying quantities satisfy the conditions of ratio measurement. A ratio scale requires a true zero and admissible transformations preserving ratios. Utilities derived from ordinal health-state preference systems do not possess these properties. Once that is recognized, the entire arithmetic foundation of the QALY collapses.

That collapse is not peripheral. It reaches into every corner of the PHARMAC framework because PHARMAC's published guidance does not merely tolerate cost-utility analysis; it operationalizes it. The use of utility instruments, especially EQ-5D, is not incidental but woven into the construction of value claims. PHARMAC has explicitly referred to New Zealand EQ-5D utility weights as a basis for measuring changes in health-related quality of life. The problem, however, is that the assignment of preference weights to descriptive health states does not transform ordinal categories into interval, let alone ratio, measurement. It simply re-expresses ordinal relations numerically. The interrogation reflects this perfectly. The false statement that EQ-5D-3L preference algorithms create interval measures is endorsed at +2.50. This is the classic signature of measurement inversion: the appearance of numerical refinement is mistaken for lawful measurement.

The same inversion appears in the treatment of subjective instrument scores more generally. The false claims that summations of subjective instrument responses are ratio measures and that Likert scores create ratio measures are both endorsed at +2.50. These are not small conceptual slips. They reveal a knowledge base that either does not recognize, or has chosen to bypass, the difference between ordinal classification and measurement. Once that boundary is crossed, almost any numerical output can be made to look authoritative. But authority is not the same as measurement. The ability to compute a number does not establish that the number has the properties required for arithmetic. In PHARMAC's case, this distinction is effectively erased.

The strongest negative logits are reserved for Rasch measurement and the entire conceptual architecture required for latent trait assessment. All Rasch-related statements are assigned the minimum value of -2.50. This indicates effective non-possession. The knowledge base does not merely omit Rasch; it has no place for it. That matters enormously. Without Rasch-conformant transformation, there is no lawful pathway from ordinal observations of subjective phenomena to invariant measures of latent traits. In other words, if one wants to make claims about quality of life, symptom burden, or other subjective therapy impacts as measurable quantities, Rasch rules are not optional embellishments. They are the only available route. PHARMAC's total absence of this framework means that its use of utilities and QALYs is not just weakly supported. It is unsupported at the one point where support would have to begin.

The profile also demonstrates that PHARMAC does not recognize the prior logical requirement that measurement must precede arithmetic. That true statement is endorsed at only 0.15, giving a normalized logit of -1.80. Closely linked to this is the very low endorsement of the proposition that meeting the axioms of representational measurement is required for arithmetic, at -2.20. This pair of results is among the most revealing in the whole table. It shows that the PHARMAC knowledge base accepts arithmetic first and asks no foundational questions later. Numerical manipulation is treated as self-justifying. This is why the framework can move so smoothly from ordinal health-state descriptions to utilities, from utilities to QALYs, and from QALYs to

cost/QALY ratios. The prior question—whether any of these objects are lawful measures—is not built into the framework.

The statement that multiplication requires a ratio measure is also strongly rejected, at -2.20. This is decisive. PHARMAC's own methods depend on multiplying utility by time to construct QALYs. Time is a ratio variable. If the companion variable is not also a ratio measure, then the operation is inadmissible. There is no escape from this requirement through professional convention, repeated use, or policy convenience. Multiplication is not licensed by consensus. It is licensed by scale type. Once that is understood, the PHARMAC method is exposed as an exercise in unauthorized arithmetic.

One might try to defend PHARMAC by saying that its framework is pragmatic, that it supports comparative judgments under resource constraints, or that it provides a common language for prioritization. None of these responses touches the measurement issue. A claim is not redeemed because it is useful for committees. A model is not validated because it is widely understood. A cost/QALY figure does not become evidence because it is embedded in policy process. The point of the interrogation is that the analytical outputs may be operationally influential while still failing the axioms required for them to count as measures. The problem is foundational, not administrative.

The profile on falsifiability reinforces this conclusion. The true statement that non-falsifiable claims should be rejected receives only weak support, at -0.40. At the same time, the false statement that reference-case simulations generate falsifiable claims is endorsed at +2.50. This is exactly what one would expect in a system centered on model-based projections. PHARMAC relies heavily on modeled estimates and future-oriented economic comparisons. Such outputs may be sensitive to assumptions and may be varied through scenario analysis, but that is not the same as being empirically falsifiable. The interrogation indicates that the PHARMAC knowledge base normalizes this confusion. Model elasticity is mistaken for scientific testability.

This is where the New Zealand case becomes particularly important. Unlike jurisdictions with a larger ecosystem of independent academic HTA groups, New Zealand has increasingly concentrated evaluative authority within PHARMAC. That gives the PHARMAC knowledge base unusual significance. It is not merely one voice among many; it is the dominant institutional expression of HTA method in the country. If that dominant knowledge base is structured around measurement inversion, then the problem is not offset by an external academic counterweight. The inversion becomes the national analytical norm.

The implications for claims evaluation are direct. A framework based on false measurement cannot generate evaluable claims. At best, it produces internally consistent numerical stories. Those stories may inform priorities, but they cannot satisfy the standards of normal science because they are not grounded in admissible measurement. They cannot be replicated as measurement claims because the constructs themselves are not lawfully measured. They cannot be falsified in the relevant sense because the model outputs rest on scale assumptions that are never established. This is why the issue is not one of improving current HTA practice within the PHARMAC structure. The structure itself excludes the conditions required for evaluability.

That is the force of the phrase measurement inversion. It does not mean merely that the field has overlooked a few technical niceties. It means the field has reversed the order of scientific reasoning. Instead of beginning with admissible measurement and then applying arithmetic, it begins with preferred arithmetic and assumes measurement into existence. Instead of asking whether constructs satisfy unidimensionality, invariance, and scale requirements, it asks how best to operationalize utilities and QALYs. Instead of rejecting non-evaluable claims, it systematizes them. PHARMAC exemplifies this inversion at an institutional level.

The caution for New Zealand is therefore severe. PHARMAC's analytical legacy cannot be defended simply by pointing to its procedural consistency or national importance. A consistent method built on false measurement is still false measurement. The profile reported here does not suggest a framework in need of marginal refinement. It suggests a framework whose foundations are irreparably compromised by its exclusion of representational measurement. If New Zealand wants HTA claims that are credible, evaluable, and replicable, then the current PHARMAC analytical structure has to be questioned at the level of first principles.

In summary, the PHARMAC interrogation is among the strongest available demonstrations of institutionalized measurement inversion. The agency's official methods openly rely on cost-utility analysis, QALYs, utility weights, and model-based projections. The canonical profile shows that the knowledge base behind those methods strongly endorses the false propositions required to sustain them and rejects the true propositions that would rule them out. The absence of Rasch measurement is complete. The commitment to admissible arithmetic is absent. The tolerance for non-falsifiable modeled claims is high. This is not an evidence framework. It is a numerical storytelling framework with national authority attached to it. For that reason, PHARMAC should be regarded not as a model of HTA discipline, but as a particularly clear example of how a closed institutional knowledge base can normalize false measurement and call it evidence.

PHARMAC AND FALSE MEASUREMENT: IMPLICATIONS FOR POLICY AND CREDIBILITY

The implications for policy and decision-making in New Zealand, given the findings for PHARMAC, are not peripheral. They are foundational. If the quantities that underpin HTA claims do not meet the axioms of representational measurement, then the framework through which PHARMAC evaluates and prioritizes therapies cannot be sustained as a scientific enterprise. This is not a criticism of specific decisions, nor of individual assessments. It is a prior question: whether the evidentiary architecture itself is capable of generating evaluable claims.

PHARMAC's decision-making processes rely, explicitly or implicitly, on constructs such as utilities, QALYs, and modeled cost-effectiveness estimates. These constructs are treated as if they possess the properties required to support arithmetic operations, multiplication, aggregation, and comparison. Yet the interrogation results demonstrate that the knowledge base does not recognize, let alone satisfy, the conditions necessary for such operations. The consequence is immediate. If the inputs are not measures, the outputs cannot be interpreted as evidence. They are numerical artifacts, internally consistent but externally invalid.

This creates a disconnect between the appearance of rigor and the reality of measurement. The presence of numbers, models, and thresholds conveys an impression of scientific precision. Decisions appear to be grounded in quantitative analysis. But this precision is illusory. It rests on the assumption that the quantities being manipulated are measurable, when in fact they are not. Once this assumption is challenged, the authority of the framework is weakened. Not incrementally, but fundamentally.

It follows that policy response cannot be framed in terms of refinement. There is no adjustment to discount rates, model structures, or parameter inputs that can correct a failure of measurement. The issue is not how the model is specified, but whether the quantities within it are admissible. If they are not, then the model, however sophisticated, cannot generate meaningful outputs. The question for PHARMAC is therefore not whether its methods can be improved, but whether they can be justified.

This places PHARMAC in a position that is shared by other HTA agencies, but no less acute for that. The reliance on QALYs and related constructs is deeply embedded in international practice. It is supported by guidelines, reinforced by academic training, and normalized through repeated application. Yet none of these factors address the measurement problem. Consensus does not confer validity. Repetition does not transform ordinal constructs into ratio measures. The framework remains dependent on assumptions that are incompatible with the axioms of measurement.

The appropriate response is therefore not defensive. It is to confront the problem at its source. This requires a shift from a framework that assumes measurement to one that demonstrates it. For manifest variables, this implies the use of linear ratio scales grounded in observable quantities. For latent constructs, it requires the adoption of measurement models—specifically Rasch-conformant approaches—that establish unidimensional, invariant scales. Without this shift, the distinction between evidence and numerical representation cannot be maintained.

There are, of course, institutional consequences. Acknowledging the measurement problem would necessitate a reassessment of established methods, training programs, and decision frameworks. It would challenge existing practices and require the development of alternative approaches. But these consequences are not optional. They follow directly from the recognition that current constructs do not meet the conditions required for measurement.

The alternative is to continue as before, relying on a framework that generates non-evaluable claims while presenting them as evidence. This may be operationally convenient, but it is not consistent with the standards of normal science or with the obligation to base decisions on valid measurement. The issue, therefore, is not whether change is difficult, but whether it is avoidable. The interrogation results suggest that it is not.

The evidence now exists. Continued reliance on these constructs without engagement risks not only methodological criticism, but the erosion of credibility.

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 three distance education programs to support the transition to a new paradigm in HTA. These comprise 12 module senior level program that details the standards for measurement, the failure of current HTA standards and the basis for protocol supported claims assessment for ratio measures of manifest attributes and Rasch logic ratio logit measures for latent attributes. The two other programs are only 5 modules but are designed to complement the 12-module program, for measurement axioms and Rasch attribute possession.

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.

MAIMON RESEARCH LLC

DISTANCE EDUCATION PROGRAMS IN THE THEORY OF MEASUREMENT

Three programs are available: two short 5-module programs and a 12-module program that is structured as a senior level course on the transition from the current HTA belief system to a new paradigm for HTA

The two short programs are (i) **NUMERICAL STORYTELLING: SYSTEMATIC MEASUREMENT FAILURE IN HEALTH TECHNOLOGY ASSESSMENT** and (ii) A **NEW START IN MEASUREMENT FOR HEALTH TECHNOLOGY ASSESSMENT.** They are designed to complement the 12-module course program. They can be accessed through the **DISTANCE EDUCATION** section of the website with URL https://maimonresearch.com/distance-education-programs/

The senior level course **HEALTH TECHNOLOGY ASSESSMENT REBUILT: EVIDENCE AND VALUE** is accessed through the EVIDENCE AND VALUE section of the website or URL link https://maimonresearch.com/evidence-and-value/.

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