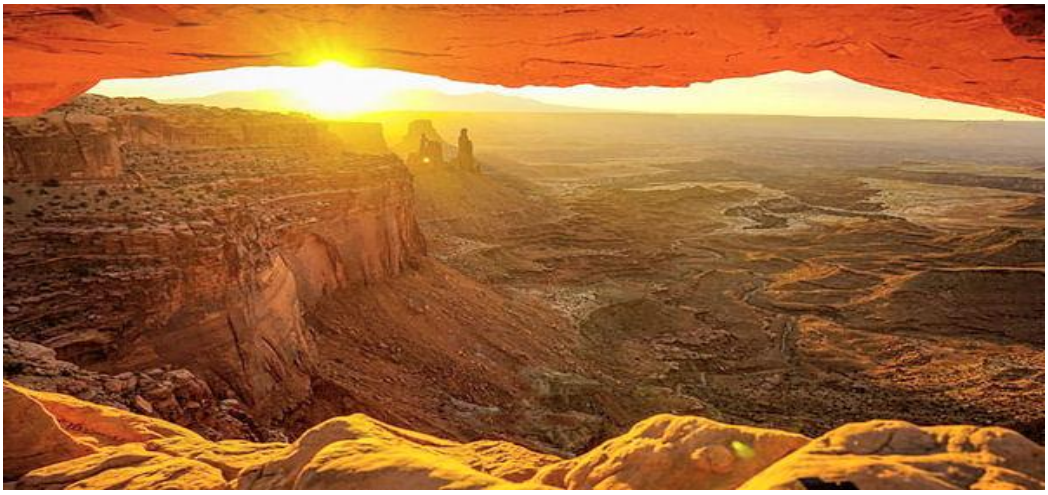
MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION



REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

AUSTRALIA: PHARMACEUTICAL BENEFITS ADVISORY COMMITTEE (PBAC) – DECISIONS WITHOUT MEASUREMENT

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

LOGIT WORKING PAPER No 34 JANUARY 2026

www.maimonresearch.com

Tucson AZ

FOREWORD

HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

This Logit Working Paper series documents a finding as extraordinary as it is uncomfortable: health technology assessment (HTA), across nations, agencies, journals, and decades, has developed as a global system of non-measurement. It speaks the language of numbers, models, utilities, QALYs, "value for money," thresholds, discounting, incremental ratios, extrapolations, and simulations. It demands arithmetic at every turn, multiplication, division, summation, aggregation, discounting, yet it never once established that the quantities to which these operations are applied are measurable. HTA has built a vast evaluative machinery on foundations that do not exist. The probabilities and normalized logits in the country reports that follow provide the empirical confirmation of this claim. They show, with unsettling consistency, that the global HTA knowledge base neither possesses nor applies the principles of scientific measurement.

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 PBAC KNOWLEDGE BASE

The knowledge base of the Pharmaceutical Benefits Advisory Committee can be characterized as an administratively stabilized analytical system organized around model conformity rather than measurement admissibility. It is defined not by explicit statements of measurement philosophy, but by the recurring concepts, assumptions, and evaluative practices that PBAC requires, accepts, and reproduces across submissions, guidelines, and committee deliberations. These patterns establish the epistemic boundaries within which pharmaceutical value is permitted to be expressed.

At the center of this knowledge base is the reference-case framework inherited largely from the NICE model. PBAC submissions are structured around cost-utility analysis, the generation of QALYs, and long-horizon simulation modeling intended to project lifetime health and cost outcomes. These models rely heavily on preference-based utility instruments, mapping functions, and algorithmic transformations that convert subjective responses into numerical inputs. Within the PBAC system, these numerical outputs are treated as if they were measured quantities, suitable for multiplication, aggregation, discounting, and comparison across therapies.

Critically, the knowledge base does not require demonstration of scale type prior to arithmetic. Utilities are assumed to behave as interval or ratio measures despite permitting negative values and lacking a true zero. QALYs are treated as homogeneous quantitative objects despite being constructed from fundamentally heterogeneous components. The admissibility of arithmetic is therefore determined procedurally rather than axiomatically. If a value is generated through an approved method and embedded within a reference-case model, it is accepted as evidence regardless of whether it satisfies measurement requirements.

Latent attributes occupy a central but unresolved role within this structure. Concepts such as health-related quality of life, functioning, wellbeing, and patient experience are routinely invoked, yet they are never formally constructed as measurable quantities. PBAC does not require unidimensionality to be demonstrated, nor does it require invariance across populations. Summated ordinal scores and preference algorithms are treated as sufficient proxies for measurement. Rasch modeling, which would impose strict constraints on item functioning and enable latent trait possession to be expressed on a logit ratio scale, is not treated as a governing requirement. Its absence allows the continued use of instrument families that could not survive measurement scrutiny.

The PBAC knowledge base also redefines falsification. Rather than requiring claims to be testable against empirical outcomes, the framework relies on sensitivity analyses, scenario testing, and internal robustness checks. These procedures vary assumptions without ever exposing claims to refutation by the world. As a result, models can be refined indefinitely without the possibility of being wrong in the Popperian sense. Closure is achieved administratively, not scientifically.

What defines the PBAC knowledge base most clearly is its patterned silence. Representational measurement theory is not debated; it is absent. Stevens' scale typology is not enforced; it is ignored. The distinction between ordinal and quantitative structures is not operationalized. These omissions are not accidental. They are structural conditions that permit the framework to function. If measurement admissibility were enforced as a gatekeeping requirement, the central numerical objects of PBAC evaluation would become indefensible.

The PBAC knowledge base therefore functions as a national instantiation of the global HTA memeplex. It reproduces a belief system in which numerical form is mistaken for measurement, complexity is mistaken for rigor, and consistency is mistaken for truth. Within this system, evidence does not evolve through falsification and correction, but through repetition and institutional reinforcement. The result is an evaluative architecture that delivers decisions efficiently while remaining epistemically closed to scientific learning.

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: PHARMACEUTICAL ADVISORY COMMITTEE

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 PHARMACEUTICAL BENEFITS ADVISORY COMMITTEE

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	0.85	+1.75
RATIO MEASURES CAN HAVE NEGATIVE VALUES	0	O.85	+1.75
EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES	0	0.85	+1.75
THE QALY IS A RATIO MEASURE	0	0.90	+2.20
TIME IS A RATIO MEASURE	1	0.95	+2.50
MEASUREMENT PRECEDES ARITHMETIC	1	0.15	-1.75
SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES	0	0.85	+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	0.90	+2.20
NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.60	+0.40
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.70	+0.85
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 PBAC COMMITMENT TO ARITHMETIC BEFORE MEASUREMENT

The PBAC occupies a central and extraordinarily powerful role in the Australian health system. Its recommendations determine national reimbursement, pricing leverage, and patient access to medicines. It presents itself as an evidence-based advisory body operating in the public interest. Yet when examined through the lens of representational measurement, the PBAC knowledge base reveals a far more troubling reality: Australia's national formulary authority operates within a framework that systematically authorizes arithmetic while rejecting the conditions that make arithmetic meaningful.

The canonical diagnostic profile does not reveal uncertainty or partial misunderstanding. It reveals a coherent belief structure whose internal logic is stable, reproducible, and fundamentally incompatible with scientific measurement. At the heart of this structure lies a decisive inversion. The propositions that must be true before numerical operations can be justified collapse toward the floor of endorsement, while propositions that enable cost-utility modeling rise toward the ceiling. Measurement does not function as a gatekeeper. It functions as an inconvenience.

The most striking illustration is the treatment of ratio measurement. The proposition that multiplication requires a ratio scale receives near-floor endorsement at $p = 0.15$ (-1.75). Measurement precedes arithmetic collapses to the same level. The requirement that representational measurement axioms be satisfied before arithmetic is permitted likewise falls to -1.75. These statements are not philosophical abstractions; they are the rules that determine whether numerical symbols represent quantities or merely labels. Yet despite rejecting these axioms, PBAC fully endorses the arithmetic that depends on them. The proposition that the QALY is a ratio measure rises to $p = 0.90$ ($+2.20$). The proposition that QALYs can be aggregated reaches the same ceiling. Likert summation as ratio measurement also reaches $+2.20$. These endorsements reveal not confusion but necessity. The arithmetic must be protected because the framework cannot function without it. This contradiction defines PBAC's epistemic posture. Arithmetic is treated as self-legitimating. Measurement is treated as optional background.

The PBAC reference case therefore does not evaluate therapy impact; it evaluates conformity to a numerical ritual. Manufacturers are not asked to demonstrate that outcomes are measurable. They are asked to demonstrate that the model is plausible, the assumptions reasonable, and the scenario analyses comprehensive. Plausibility substitutes for truth. Internal coherence substitutes for empirical validity. The diagnostic exposes this substitution with precision. Reference-case simulations are treated as falsifiable claims at $p = 0.70$ ($+0.85$), even though simulations are

conditional projections incapable of falsification unless tied to prospective empirical protocols. In PBAC practice, falsification is redefined as sensitivity analysis. Claims are never wrong; they merely become alternative scenarios. This redefinition is essential to institutional closure. A system that allowed genuine falsification could never finalize pricing decisions. Claims would remain provisional throughout a product's life cycle. The reference case avoids this instability by ensuring that no claim can ever be refuted. Closure replaces discovery.

The role of latent attributes further exposes the depth of the problem. PBAC routinely relies on patient-reported outcomes, quality-of-life instruments, and preference-based measures to populate utility models. Yet every Rasch-related proposition collapses to the floor of the scale. The requirement that latent traits be measured through Rasch transformation sits at $p = 0.10$ (−2.20). The proposition that Rasch logits are the only admissible scale for latent-trait impact assessment also sits at −2.20. The equivalence between Rasch axioms and representational measurement axioms likewise disappears. This pattern does not indicate disagreement with Rasch. It indicates exclusion. Rasch is not rejected because its claims are false; it is rejected because its consequences are unacceptable. If Rasch were adopted as a gatekeeper, most utility instruments would fail instantly. Ordinal summations would be exposed as non-measures. Mapping would collapse. Preference algorithms would lose legitimacy. The entire cost-utility edifice would unravel. Thus Rasch is rendered invisible.

The treatment of unidimensionality reinforces the same conclusion. Measures must be unidimensional receives weak endorsement at $p = 0.25$ (−1.10). Yet time trade-off preferences are treated as unidimensional at $p = 0.85$ (+1.75). This is not empirical demonstration; it is assertion by necessity. Multiattribute health-state descriptions are declared to map onto a single continuum because the arithmetic demands a single continuum. Dimensionality becomes a rhetorical convenience rather than a measurement requirement.

PBAC therefore institutionalizes a core fiction: that preference scores derived from multidimensional descriptive systems behave as quantities on a single scale. This fiction is not tested, demonstrated, or validated. It is assumed. The result is a national HTA system in which the most important dependent variables are not measures in the scientific sense. They lack invariant units. They lack demonstrable equal intervals. They lack meaningful zero points. Yet they are multiplied, averaged, discounted, aggregated, and applied as thresholds if they were physical quantities. This is not approximate science. It is categorical error.

The diagnostic also reveals why this structure has remained stable for decades. The propositions that would expose the failure collapse uniformly across institutions, academic centers, and agencies. Measurement precedes arithmetic does not provoke debate because it is not part of the professional vocabulary. Representational measurement theory is absent from training. Stevens' typology is cited historically, if at all, but never operationalized. Rasch is treated as niche psychometrics rather than foundational measurement. PBAC did not consciously reject these principles. It never encountered them.

This absence is critical. The PBAC knowledge base did not evolve through scientific contestation. It evolved through institutional imitation. Australia adopted the NICE reference case wholesale,

assuming that its epistemic foundations had already been settled elsewhere. The belief was not that the framework was perfect, but that it was legitimate. That belief was misplaced.

The diagnostic shows that PBAC's framework is not a local variant of scientific evaluation. It is a cloned instance of the NICE numerical storytelling memeplex. Its core commitments mirror those of NICE, ICER, and other reference-case systems with near-perfect symmetry. Differences exist only in administrative detail, not epistemic structure. This explains the remarkable uniformity of failure across jurisdictions. The same propositions collapse to the floor everywhere. The same false propositions rise to the ceiling everywhere. The global HTA ecosystem did not converge on truth; it converged on convenience. PBAC's role in this system is therefore not neutral. As Australia's national reimbursement authority, it functions as a belief enforcer. Academic centers train analysts to satisfy PBAC requirements. Consultants optimize submissions to PBAC expectations. Journals publish PBAC-compatible analyses. Over time, the entire Australian HTA ecosystem becomes conditioned to accept arithmetic without measurement as normal practice.

This conditioning explains why dissent is rare. To challenge the measurement basis of PBAC submissions is not to propose refinement; it is to question the legitimacy of the system itself. Careers, funding, and institutional authority are all bound to its continuation. The consequence is that PBAC cannot learn from its own decisions. Because its claims are not measurable, they cannot be falsified. Because they cannot be falsified, they cannot be improved. Because they cannot be improved, the system evolves procedurally but not scientifically. Guidelines are revised. Templates are updated. Models become more complex. Yet the core epistemic defect remains untouched.

PBAC therefore does not participate in the evolution of objective knowledge. It participates in the reproduction of a stable administrative mythology. This does not mean PBAC acts in bad faith. It means PBAC operates within a framework that was never designed to support scientific learning. The reference case was designed to deliver decisions under uncertainty, not to test claims about reality. It is a governance technology, not a measurement system.

Table 1 and the logit diagnostic profile makes this unmistakable. The problem is not one of execution. It is one of architecture. Until PBAC acknowledges that arithmetic without measurement cannot support scientific legitimacy, no amount of modeling sophistication will repair the system. Better utilities cannot fix ordinal data. Better mapping cannot create invariant units. Better scenario analysis cannot substitute for falsification.

Australia therefore faces a stark choice. PBAC can continue to operate as a national arbiter of numerical storytelling, enforcing a framework that produces closure without knowledge. Or it can become one of the first HTA agencies to abandon the reference case and move toward single-claim, protocol-driven, measurement-valid evaluation. That transition would require explicit recognition of two admissible forms of evidence: linear ratio measures for manifest attributes and Rasch logit ratio measures for latent traits. It would require claims to be unidimensional, evaluable, reproducible, and falsifiable. It would require training in representational measurement as a governance prerequisite.

The diagnostic does not merely criticize PBAC. It reveals what PBAC could become. But it also makes one conclusion unavoidable: as presently constituted, PBAC does not evaluate therapy impact. It evaluates the performance of arithmetic divorced from measurement. And that is not science.

WHY HAS THE PBAC NO CONCEPT OF MEASUREMENT?

The absence of any coherent concept of measurement within the PBAC is not an accidental oversight, nor the result of individual ignorance or intellectual failure. It is a structural consequence of how the Australian health technology assessment framework was constructed, transmitted, and institutionalized. PBAC did not reject measurement. It never encountered it.

From its inception, PBAC was established as an administrative decision body, not as a scientific institution. Its mandate was to advise government on reimbursement under conditions of uncertainty, limited data, and political pressure to achieve timely access decisions. Within that context, the central problem was not how to generate falsifiable knowledge about therapy impact, but how to reach closure. Measurement theory, which insists that claims remain provisional and open to empirical refutation, is fundamentally incompatible with administrative closure. PBAC therefore evolved around tools that appear quantitative without requiring measurement.

The framework it adopted was imported, not developed. From the 1990s onward, PBAC aligned increasingly with the emerging UK reference-case model. That model did not arrive with representational measurement theory attached. It arrived as a ready-made evaluative template: utilities, QALYs, cost-effectiveness ratios, and simulation modeling. These constructs were presented as established international best practice. PBAC did not ask whether they were measures, because the framework was never framed as a hypothesis requiring validation. It was framed as a solution.

This matters because measurement is not intuitive. Representational measurement theory is not something that arises naturally from statistical training, economic modeling, or decision analysis. It requires explicit exposure to axioms governing scale type, permissible transformations, and the distinction between ordering and quantifying. By the time PBAC institutionalized its analytical framework, these concepts had largely disappeared from applied economics and policy curricula. Stevens' scale typology was no longer taught operationally. Rasch measurement was confined to specialist psychometrics. Measurement had become invisible.

As a result, numerical output itself came to stand in for measurement. If a value was expressed with decimals, confidence intervals, and sensitivity analyses, it was assumed to be quantitative. The question "what kind of number is this?" was never asked, because no one within the institutional ecosystem possessed the conceptual tools to ask it. Utilities were treated as interval measures by convention. QALYs were treated as ratio measures by assertion. The arithmetic followed automatically.

Once this assumption was embedded, it became self-reinforcing. PBAC guidelines codified the framework. Academic centers trained analysts within it. Consultants optimized models around it. Journals reviewed submissions against it. Each layer inherited the same foundational silence. At

no point was there an institutional moment where measurement admissibility was defined as a gatekeeping requirement. PBAC therefore did not "ignore" measurement; it never recognized measurement as a prerequisite to arithmetic.

The decisive factor is that the PBAC framework does not require empirical falsification. Measurement exists to support falsifiable claims: quantities that can be tested, reproduced, and potentially refuted. PBAC does not operate in that epistemic mode. Its central decision instrument is the reference-case simulation model, which produces conditional projections rather than testable claims. Sensitivity analysis replaces falsification. Plausibility replaces empirical risk. Once a framework is structured around models that cannot be wrong in the scientific sense, the need for measurement disappears entirely.

Indeed, true measurement would destabilize the system. If PBAC required demonstrable ratio measurement for multiplication, cost-utility analysis would collapse. If latent attributes required Rasch logit ratio scales, most preference instruments would become inadmissible. If unidimensionality were enforced, composite quality-of-life constructs would fail. Measurement would not refine the framework; it would invalidate it. The absence of measurement is therefore not a gap. It is a functional necessity.

This explains the remarkable stability of the PBAC belief system. Once arithmetic without measurement becomes normalized, there is no internal mechanism for correction. New evidence cannot overturn old claims because no invariant quantity exists to test. Learning becomes impossible in the scientific sense. What evolves instead is methodological elaboration: more complex models, more refined scenarios, more detailed submissions, all operating on the same non-measured foundation.

PBAC's lack of a measurement concept is therefore not a failure of intellect. It is the predictable outcome of adopting a governance technology rather than a scientific framework. The reference case was never designed to generate objective knowledge. It was designed to support decision making under uncertainty while avoiding perpetual contestation. Measurement threatens closure. Modeling delivers it.

Only now, with AI large-language-model diagnostics capable of revealing belief structures at scale, has the absence become visible. What appears as a national evaluative system is revealed as a patterned silence: measurement precedes arithmetic is missing; scale type is absent; Rasch is invisible. The emperor was never examined because no one knew what clothes were supposed to look like. That is why PBAC has no concept of measurement. Not because it rejected science, but because it built an entire evaluative architecture in which science, as defined by falsification and representational measurement, was never required to function.

PUBLIC POLICY IMPLICATIONS FOR AUSTRALIA

The PBAC has, for more than 45 years, stood at the center of Australia's pharmaceutical subsidy system, projecting an image of methodological rigor, fiscal discipline, and global leadership in cost-effectiveness analysis. Yet the logit profile reported here shows that the PBAC knowledge base has never possessed the measurement foundations required to justify the arithmetic on which

its decisions rest. The Committee's intellectual architecture. its use of utilities, QALYs, DALYs, and incremental cost-effectiveness ratios, fails every relevant axiom of representational measurement. The PBAC knowledge base treats ordinal preferences as if they were interval or ratio quantities while displaying no conceptual understanding of unidimensionality, dimensional homogeneity, transformation rules, or the Rasch model. These omissions are not marginal: they strike at the legitimacy of every numerical claim the PBAC has made since formalizing cost-effectiveness requirements in the early 1990s.

The public policy implications are profound. If utilities are ordinal and the PBAC's logits implicitly confirm that status, then every QALY supplied to the Committee, every ICER calculated, every threshold comparison invoked, and every "value for money" conclusion reached has been arithmetically invalid. Tens of billions of dollars in subsidy decisions have relied on non-quantities passed off as measures. The direction, magnitude, and comparative ranking of cost-effectiveness claims become unstable once the foundations of measurement fail. This means that PBAC may have subsidized products that were not cost-effective, rejected ones that were, distorted prices, and imposed access restrictions on false grounds, not because of malfeasance but because the Committee lacked the conceptual tools to distinguish measurement from numerical symbolism.

Equally serious is the PBAC's treatment of simulation modelling as if it were evidence. The positive logit for the belief that reference-case simulations yield falsifiable claims illustrates a systemic misunderstanding: models do not generate testable predictions, yet PBAC has relied on them as the primary basis for listing decisions. This amounts to substituting unfalsifiable projections for empirical measurement; an epistemic error with large fiscal consequences.

Perhaps the most damaging finding is PBAC's total absence of engagement with Rasch measurement. For decades the Committee has evaluated latent constructs, health states, utilities, quality of life, without the only scientific model capable of constructing interval measures from ordinal responses. PBAC has therefore never measured the attributes on which its cost-effectiveness arithmetic depends. In short, Australia's subsidy architecture has been built on arithmetic applied to non-measures. The PBAC did not merely drift away from measurement; it never possessed it.

WHY AUSTRALIAN ACADEMIA ALIGNS WITH THE PBAC

The alignment of Australian academic health economics and HTA research centers with the Pharmaceutical Benefits Advisory Committee (PBAC) did not arise from explicit instruction or coercion. It emerged through a long process of institutional conditioning in which incentives, funding structures, and professional advancement converged around a single evaluative authority. Over time, PBAC became not just a decision-making body but the gravitational center of the Australian HTA knowledge system. Academic research evolved to serve that center, and in doing so, progressively surrendered the capacity to challenge its foundational assumptions.

Funding is the primary mechanism through which this alignment took hold. Australian HTA research centers depend heavily on public grants, commissioned evaluations, advisory contracts, and collaborative projects that are either directly linked to PBAC or implicitly conditioned on PBAC-compatible methods. Research agendas that accept cost-utility analysis, QALYs, and

reference-case modeling as given are fundable, legible, and welcomed. Research that questions whether utilities are measures, whether QALYs satisfy the axioms of representational measurement, or whether arithmetic is being applied lawfully is not explicitly banned, but it is structurally unsupported. Over time, this creates a strong selection effect: certain questions are repeatedly asked because they are rewarded, while others quietly disappear.

Career incentives reinforce the same pattern. Academic success in this domain depends on publication, citation, committee participation, and policy relevance. In Australia, policy relevance in HTA is defined overwhelmingly by proximity to PBAC. Papers that refine modeling techniques, explore marginal adjustments to utility values, or extend PBAC-style frameworks circulate easily through journals, conferences, and advisory networks. By contrast, work grounded in representational measurement theory or Rasch measurement lacks an institutional audience. Junior researchers learn quickly which forms of critique advance a career and which isolate it. The result is not intellectual agreement but adaptive conformity.

This process produces what can reasonably be described as intellectual capture, though not in a conspiratorial sense. Capture here is endogenous. PBAC establishes the evaluative rules. Academic centers train analysts to work within those rules. Graduates move between universities, consultancies, and advisory roles carrying the same assumptions with them. Methodological uniformity becomes indistinguishable from methodological correctness. Foundational critique is not refuted; it is never incorporated into the professional conversation.

In this respect, the comparison to Lysenkoism is instructive, not as a moral accusation but as a structural analogy. Lysenko did not dominate Soviet biology solely through political repression. His influence persisted because his framework aligned with institutional incentives, ideological priorities, and administrative convenience. Mendelian genetics was not disproven; it was rendered irrelevant to career survival. Similarly, in Australian HTA, representational measurement theory was not debated and rejected. It was bypassed. Silence, rather than prohibition, was sufficient to secure conformity.

The key similarity lies in epistemic closure. Once a system is structured so that deviation carries professional cost and conformity carries reward, even highly capable scholars cease to ask foundational questions. This is not because they lack intelligence or integrity, but because the system defines legitimacy in advance. Over time, methodological boundaries harden into intellectual walls. Alternative frameworks are no longer seen as incorrect; they are seen as inapplicable.

The durability of this alignment is further strengthened by its moral framing. PBAC methodology is routinely justified as pragmatic, necessary, and socially responsible. Academic participation is framed as contributing to equitable access and fiscal stewardship. In this context, questioning the measurement foundations of PBAC is portrayed as abstract theorizing that threatens timely decision making. Foundational critique is recast as obstruction, rather than as a prerequisite for scientific validity.

The result is a self-reinforcing ecosystem. PBAC does not need to defend its assumptions because academia has normalized them. Academia does not challenge PBAC because its relevance,

funding, and authority depend on methodological compatibility. Together, they form a closed epistemic loop in which false measurement persists without ever being explicitly defended.

The Lysenko comparison matters because it illustrates how error can become durable without bad faith. When falsification is excluded, when foundational theory is marginalized, and when conformity is rewarded, a system does not merely tolerate error; it institutionalizes it. Australian academia did not simply align with PBAC. It was shaped by it, and in doing so, helped stabilize a framework that delivers decisions while foreclosing the evolution of objective knowledge.

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