MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

AUSTRALIA: ACADEMIC RESEACH CENTERS ENDORSE THE ABSENCE OF MEASUREMENT IN HEALTH TECHNOLOGY ASSESSMENT

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

LOGIT WORKING PAPER No 33 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 assessment is to evaluate the extent to which Australian academic research centers in health technology assessment and health economics demonstrate alignment with the axioms of representational measurement. Using the canonical 24-item statement framework, the analysis interrogates whether the national academic knowledge base recognizes measurement as a necessary precondition for arithmetic, distinguishes correctly between scale types, and applies appropriate standards to both manifest and latent attributes. The purpose is not to assess the technical sophistication of modelling practices, but to determine whether the foundational conditions required for scientific claims are present. In particular, the analysis examines whether Australian academic HTA centers function as epistemic gatekeepers, screening out inadmissible constructs, or whether they instead reproduce the NICE-derived reference case framework without enforcing measurement validity.

This assessment forms part of the broader Logit Working Paper program, which applies a consistent diagnostic instrument across national systems, agencies, journals, and academic centers to reveal structural belief patterns rather than isolated methodological preferences. By applying the same canonical statements used in the national Australian assessment, the analysis allows direct comparison between the academic knowledge base and the policy apparatus it supplies. The central question is whether Australian universities provide an independent scientific check on HTA practice, or whether they operate as transmission nodes within the global numerical storytelling memeplex.

The findings demonstrate that Australian academic HTA and health economics research centers do not operate as measurement gatekeepers. The canonical statement profile reveals a consistent inversion of scientific ordering: propositions establishing measurement as a prerequisite for arithmetic fall at or near the floor of endorsement, while propositions that enable the QALY, preference algorithms, score aggregation, and reference-case modelling cluster at or near the ceiling. This pattern mirrors almost exactly the profiles observed for NICE-aligned agencies and the major HTA journals, indicating not local deviation but structural replication.

Across the 24-item set, there is minimal reinforcement of representational measurement axioms, negligible recognition of Rasch measurement as a necessary framework for latent traits, and strong normalization of arithmetic applied to ordinal and composite constructs. Falsification is acknowledged rhetorically but operationally replaced by scenario analysis and internal model coherence. The academic knowledge base therefore does not challenge the reference case; it stabilizes it. Rather than functioning as a corrective to policy-driven HTA, Australian academic centres reproduce the same belief structure, ensuring continuity of practice across training, publication, and advisory roles.

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 [i]. 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) [ii] . 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 [iii]. 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 [iv].

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 AUSTRALIAN ACADEMIC HTA KNOWLEDGE BASE

The knowledge base of Australian academic research centres in health technology assessment can be characterized as methodologically sophisticated yet epistemically ungrounded. It is organized around the refinement of analytical techniques rather than the admissibility of the quantities to which those techniques are applied. Within this system, numerical expression is routinely treated as evidence of measurement, and the distinction between ordering, scoring, and measuring is rarely made explicit. As a result, the presence of numbers substitutes for the demonstration of quantitative meaning.

At the center of this knowledge base lies the unexamined acceptance of the NICE reference case architecture. Preference-based health state descriptions, utility algorithms, QALYs, and long-horizon simulation models are treated not as contestable constructs but as the natural language of evaluation. Academic work typically focuses on improving estimation efficiency, handling uncertainty, refining extrapolation, or optimizing model structure. What is absent is any requirement to demonstrate that the dependent variables used in these analyses satisfy the axioms of representational measurement. Measurement is assumed rather than established.

Latent attributes occupy a particularly revealing position. Concepts such as health-related quality of life, functioning, symptom burden, and wellbeing are routinely invoked as if they represent measurable continua. Yet these constructs are rarely defined as single attributes, and almost never subjected to formal measurement modeling capable of producing invariant units. Instead, summated questionnaire scores are treated as quantitative outcomes, despite lacking demonstrated equal intervals or invariance across populations. Psychometric indicators such as reliability, responsiveness, and construct validity are routinely substituted for measurement itself, even though none of these properties establishes scale type.

The absence of Rasch measurement is decisive. Rasch modeling provides the only framework capable of transforming ordinal responses into a linear logit ratio scale representing possession of a latent trait. Its near-total absence from the academic HTA knowledge base indicates that latent attributes are not treated as measurable entities but as score-based proxies. This allows arithmetic to proceed without confronting the consequences of ordinality. Means, differences, regressions, mappings, and preference transformations are performed as if the underlying numbers possessed interval or ratio properties, even though such properties have not been demonstrated.

This permissive structure supports the downstream use of utilities and QALYs. Once questionnaire scores are treated as quantitative, mapping to preference-based instruments becomes plausible. Once utilities are accepted as interval-like, multiplication by time appears legitimate. Once QALYs are generated, aggregation and threshold comparison follow naturally. Each step depends on the prior step being accepted without interrogation. The academic knowledge base thus functions as a supply chain for numerical artifacts rather than as a scientific filter.

Importantly, this system maintains internal coherence while remaining externally indefensible. Within its own boundaries, the knowledge base appears consistent: models behave predictably, results can be replicated procedurally, and outputs can be compared across studies. What is missing is invariance with respect to the empirical world. Claims cannot be falsified because the quantities themselves lack defined measurement properties. Sensitivity analysis replaces refutation. Robustness replaces truth-testing.

Australian academic centers do not merely reflect national HTA practice; they stabilize it. Through teaching, supervision, publication, and advisory roles, they reproduce the same assumptions across generations of analysts. Students learn how to implement models but not how to interrogate whether the numbers entering those models are measures. Journals receive manuscripts that conform to expected templates. Agencies receive advice framed in familiar constructs. Over time, the absence of measurement becomes normalized as professional competence.

In this sense, the Australian academic HTA knowledge base functions not as an engine of scientific discovery but as a mechanism of continuity. It ensures that the reference case survives unchallenged, not because it has been validated, but because the conceptual tools required to invalidate it are not part of the system's intellectual equipment. The result is an academically endorsed framework capable of generating decisions, but incapable of producing evaluable, falsifiable, or replicable claims in the sense required by 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: AUSTRALIAN ACADEMIC HTA RESEARCH CENTERS

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 AUSTRALIAN ACADEMIC RESEARCH CENTERS

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	0.85	+1.75
EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES	0	0.85	+1.75
THE QALY IS A RATIO MEASURE	0	0.90	+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	0.80	+1.40
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	0.80	+1.40
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: AUSTRALIAN HTA RESEARCH CENTERS AND MEASUREMENT DENIAL

Australia's university-based HTA and health economics research centers present themselves as the analytic conscience of the national decision system: the places where claims are supposedly disciplined, methods refined, and evidence standards advanced. The 24-item canonical profile shows the opposite. These centers do not function as gatekeepers against false measurement. They function as its professional replication machinery. They are not merely "aligned" with the NICE reference case memeplex; they are one of the mechanisms by which it is taught, reproduced, defended, and exported into successive cohorts of analysts, reviewers, and policy advisers. The pattern is familiar because it is structurally homologous to what we have already documented in the journal ecosystem. Where *Value in Health* supplies legitimacy and *Pharmacoeconomics* supplies reinforcement, Australian academic centers supply throughput: they train the workforce that keeps the arithmetic running even when measurement is absent. This is why the academic target matters. If the university system cannot recognize measurement as the non-negotiable precondition for arithmetic, then no downstream agency can be expected to do so either, because the agency's staff and its external experts are drawn from the same knowledge loop.

Table 1 is damning precisely because it is not subtle. It shows a systematic inversion: propositions that would block illegitimate arithmetic collapse toward the floor, while propositions that enable the QALY and the reference case rise toward the ceiling. Measurement precedes arithmetic sits at $p = 0.15$ with a logit of −1.75. Multiplication requires a ratio measure sits at $p = 0.15$ (−1.75). Meeting the axioms of representational measurement is required for arithmetic is again $p = 0.15$ (−1.75). These are not refinements. These are the entry conditions for any claim to quantitative science. When the academic knowledge base endorses them at the near-floor, it is not "missing nuance." It is communicating, structurally, that it does not treat measurement constraints as binding. In such an environment, arithmetic becomes a permitted ritual, and any object formatted numerically can be treated as if it were a quantity.

That permission is exactly what the rest of the profile reveals. The QALY is a ratio measure is endorsed although false with $p = 0.90$, placing the endorsement falsehood at the ceiling (+2.20). QALYs can be aggregated is likewise endorsed with $p = 0.90$ (+2.20). The claim that EQ-5D-3L preference algorithms create interval measures is endorsed although again false with $p = 0.85$ (+1.75), again reflecting that the underlying belief system treats algorithmic scoring as if it confers interval properties. The claim that ratio measures can have negative values is endorsed with $p = 0.85$ (+1.75), which captures a key incoherence the HTA memeplex must normalize to survive: if "utilities" can be negative while still being treated as ratio quantities, then the concept of a true zero has been discarded while the arithmetic of ratio measurement is retained as if nothing happened. That combination is not a technical disagreement; it is categorical inconsistency. Yet it is reinforced rather than excluded.

Table 1 also exposes how the academic ecosystem protects itself from measurement accountability by maintaining a permissive stance toward psychometric scoring. Summation of Likert question scores creates a ratio measure is endorsed with p = 0.90 (+2.20). Summations of subjective instrument responses are ratio measures is endorsed with p = 0.85 (+1.75). These items are crucial because they reveal the everyday operational premise that makes the rest of the system possible. If total scores are treated as quantitative outcomes, then mapping can be justified, utilities can be extracted, preference algorithms can be applied, and the QALY engine can run. If, however, those totals are recognized as ordinal and non-additive, then the downstream constructs collapse. The academic knowledge base therefore has an adaptive incentive to treat scoring as measurement: it is the cheapest way to obtain numbers that look like quantities without doing the work required to establish that they are quantities.

This is why unidimensionality is the fulcrum. Measures must be unidimensional sits at p = 0.25 (−1.10). That is not a minor shortfall; it is the near-rejection of the single attribute requirement. Yet at the same time, the system must declare unidimensionality when convenient. Time trade-off preferences are unidimensional is endorsed although they are composite ordinal structures is endorsed with p = 0.85 (+1.75), which functions as a doctrinal protection: it licenses treating multiattribute health state descriptions as if they map onto a single latent continuum suitable for arithmetic. In other words, unidimensionality is not treated as something to be demonstrated; it is treated as something to be asserted where the modelling pipeline needs it. That is the hallmark of scholasticism: definitions and conventions substitute for empirical demonstration. A measurement-literate academy would invert this pattern. It would insist on demonstrated unidimensionality before allowing arithmetic. Here, arithmetic comes first, and unidimensionality is redefined as a convenience.

The Rasch block is where the academic failure becomes decisive, because Rasch is the only pathway that turns subjective ordinal responses into a lawful measurement structure for a latent attribute. The table shows that this pathway is functionally absent from the academic HTA knowledge base. There are only two classes of measurement linear ratio and Rasch logit ratio sits at p = 0.10 (−2.20). Transforming subjective responses to interval measurement is only possible with Rasch rules sits at p = 0.10 (−2.20). The Rasch logit ratio scale is the only basis for assessing therapy impact for latent traits sits at p = 0.10 (−2.20). The Rasch rules for measurement are identical to the axioms of representational measurement sits at p = 0.10 (−2.20). These values do not mean that no one in Australia has ever run a Rasch model. They mean something more structurally important: Rasch is not recognized as a gatekeeping requirement. It is not sovereign. It is not treated as the price of admission for any claim about latent trait change. Instead, the academy tolerates the continued use of summated ordinal totals and the interpretation of those totals as if they were measures. That unacceptable tolerance is exactly what keeps the QALY supply chain fed.

Consider the misuse of the term “quality of life,” which is the most rhetorically powerful misdirection in the entire enterprise. “Quality of life” is not a single attribute. It is a composite context term, an umbrella under which multiple domains are bundled by convention. When the academic knowledge base treats “quality of life” as a measurable quantity, it is implicitly declaring that a multiattribute construct behaves like a unidimensional continuum. That is not measurement; it is naming. Rasch makes the point brutal: latent trait possession can only be quantified on a Rasch logit ratio scale once a single attribute has been specified, items calibrated, invariance tested, and unit structure established. Without that, there is no lawful basis for saying a population possesses “more” or “less” of the attribute in any meaningful quantitative sense.

Table 1 reinforces that the academy does not treat possession as the outcome. The outcome of interest for latent traits is the possession of that trait sits at p = 0.25 (−1.10), which means the concept of possession is weakly endorsed at best. Instead, what dominates is score talk: changes in totals, mean differences, responder thresholds, minimally important differences, and mapping relationships. These are all derivative of scoring conventions, not demonstrations of measurement. The academy’s silence on possession is therefore not semantic; it is structural. If possession were taken seriously, Rasch would become mandatory, and much of the routine analytics would become inadmissible.

The same pattern appears in the treatment of falsification. Non-falsifiable claims should be rejected sits at p = 0.60 (+0.35), which looks like a respectable posture but is, in practice, a rhetorical safety valve: it allows the system to speak Popperian language without accepting Popperian discipline. The key is the paired item: reference case simulations generate falsifiable claims is endorsed although false at p = 0.80 (+1.40). This captures the academy’s operational redefinition of falsification. In the reference case world, “test” does not mean risking refutation by the world; it means varying assumptions and checking the stability of outputs. Sensitivity analysis becomes a substitute for falsification. Robustness becomes internal coherence. And internal coherence is then presented as if it were scientific legitimacy. That move is epistemically catastrophic because it eliminates the possibility of being wrong in the strong sense. If a claim cannot be refuted by observed outcomes because it is insulated by assumptions and scenario variation, then it is not a scientific claim; it is a conditional narrative. The academic ecosystem’s endorsement of the falsifiability fiction explains why the reference case could spread: it promised closure while remaining immune to refutation.

The true scandal, and the reason the academic centers deserve their own Logit Working Paper, is that their professional role should have been to stop this. Universities are supposed to be the places where foundational questions are asked. Yet the table shows that the foundational questions are not merely unanswered; they are structurally excluded. Interval measures lack a true zero sits at p = 0.20 (−1.40), again reflecting the low salience of scale-type constraints. Claims for cost-effectiveness fail the axioms of representational measurement sits at p = 0.20 (−1.40), indicating that even the possibility of categorical failure is weakly endorsed. These low endorsements are not neutral. They function as selection criteria for what can be published, taught, supervised, and funded. If measurement constraints were treated as binding, the entire cost-utility apparatus would be rendered suspect at the dependent variable level, and a vast amount of “methodological progress” would be reclassified as optimization within an inadmissible framework.

This is precisely why the comparison to the journal pillars matters. Australian academic centers do not operate in isolation. They publish into the same journal ecosystem that legitimizes and reinforces the memeplex. They cite and are cited by those journals. They teach curricula aligned with those journals. They train reviewers who act as filters for those journals. Once that loop is established, internal challenge becomes professionally costly. The academy becomes a replication engine rather than a corrective. That is the Dawkins point: memeplexes survive by creating mutually reinforcing environments in which replication is rewarded and foundational challenge is penalized. Table 1 reports a measurement instrument that detects that environment. The floor-level Rasch endorsements indicate that the only framework capable of imposing measurement discipline on latent trait claims has been quarantined, not because it is wrong, but because it is existentially disruptive.

The "two admissible measures" distinction is the center of gravity and it is the simplest way to expose the entire system. For manifest attributes, counts, events, time, resource use, a linear ratio scale is the admissible standard, because it carries a true zero and supports multiplication and division. For latent attributes, symptom burden, functioning, patient experience, only Rasch can deliver a logit ratio measure that supports meaningful differences in possession. Everything else is either ordinal or composite, and arithmetic on it is illicit. This is why measurement precedes arithmetic. It is not philosophy; it is the condition for meaning. The table shows that Australian academic centers do not teach or enforce this. They normalize the opposite ordering: arithmetic first, measurement later, if at all. And once arithmetic is normalized, it becomes easy to mistake technical sophistication for scientific legitimacy. You can produce elegant models, elaborate uncertainty analyses, hierarchical regressions, and mapping algorithms that look impressive while remaining anchored to non-measures.

That is why the academic ecosystem, despite its self-regard, cannot claim to be advancing objective knowledge. Objective knowledge requires falsifiable claims expressed in admissible units that are invariant enough to support replication across contexts. Without admissible measurement, you cannot accumulate knowledge; you can only accumulate papers. You can only accumulate numbers. You can only accumulate consensus. And consensus is exactly what the memeplex produces: the appearance of settled doctrine without the discipline that would justify settlement.

If you want the blunt operational summary: Table 1 shows that Australian academic HTA centers are structurally committed to keeping the reference case operational. They reinforce the enabling falsehoods (utilities treated as interval/ratio-like, QALYs treated as ratio quantities, aggregation treated as legitimate, negative values tolerated as if compatible with ratio status) while relegating the measurement gatekeeping axioms to near-zero endorsement. They tolerate Rasch as marginal rather than mandatory because Rasch would force the field to abandon score-based endpoints and re-found latent-trait claims on a lawful measurement structure. They treat falsification as scenario exploration because true falsification would prevent administrative closure. In doing so, they reproduce the NICE template under Australian branding while maintaining the illusion that what they produce is science rather than structured numerical storytelling.

This is also why the LLM diagnostic changes the game. Traditional critique could be absorbed, debated, and sidelined. The diagnostic exposes patterns of belief as structure: what is reinforced at ceiling and what collapses to floor. It shows that what is missing is not a technique but a prerequisite. Once that is visible, the academy's prestige cannot rescue it, because prestige does not create measurement properties. Journals cannot rescue it, because journals share the same omissions. Agencies cannot rescue it, because agencies rely on the same trained workforce. The only resolution is to reverse the ordering: establish admissible measures first, then permit arithmetic, then specify protocols for single claims, then evaluate and replicate.

If you want the final sting that remains fully defensible: Australian academic HTA centers did not "buy" the NICE reference case; they helped sell it, by reproducing it as teachable doctrine while never insisting on the measurement proofs that would be required in any normal science discipline. They are therefore not bystanders to the measurement debacle. They are one of its most important domestic replicators.

WHY AUSTRALIAN ACADEMIA ALIGNED 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.

THE CHERE HTA METHODS REVIEW: CARTOGRAPHY OF NON-MEASUREMENT

The Health Technology Assessment Policy and Methods Review commissioned by the Australian Government was, on paper, an opportunity to confront forty years of methodological failure. Within that program, the Centre for Health Economics Research and Evaluation (CHERE) produced Paper 5, *HTA Methods: Economic Evaluation*, a 200-page survey of economic evaluation methods in Australia and selected comparator jurisdictions [v]. It is one of the key technical inputs cited in the final HTA Review report and in the broader narrative that Australia is aligned with "best practice" internationally; the fact that "best practice' is equated with non-measurement is not a consideration.

Viewed through the logit instrument, however, the CHERE paper is not a neutral mapping of methods. It is a meticulously referenced catalogue of non-measurement. It documents, endorses, and normalizes precisely the practices that Table 1 show to be mathematically impossible: the use of ordinal utilities as if they were interval measures, the construction of QALYs as if they were ratio quantities, and the treatment of model outputs as if they were falsifiable evidence.

The structure of the CHERE paper is revealing. Part 1 inventories "methods in economic evaluation": approaches to cost-effectiveness and cost-utility analysis, perspectives, comparator selection, and systematic review of economic evaluations in Australia and other HTA systems. It then canvasses weighting of health outcomes and harms, including multi-attribute utility instruments and patient-reported outcomes, and discusses extrapolation, discounting, and uncertainty analysis. Part 2 reviews special cases such as rare diseases and small populations; Part 3 describes recent reforms in other jurisdictions. At every step, utilities and QALYs are treated as given primitives. They are never subjected to the prior question that measurement theory demands: are these things measures?

There is no discussion of scale types; no mention of nominal, ordinal, interval, or ratio distinctions; no acknowledgment that multiplication and division require ratio scales; no consideration of dimensional homogeneity; and no recognition that Rasch models are the only known route from ordinal responses to interval measurement for latent traits. The words "representational measurement," "unidimensionality," and "Rasch" do not appear in the public description of the work. Utilities are treated as "health state values." QALYs are treated as natural outputs of weighting frameworks. The CHERE paper, in other words, writes the QALY into the landscape as though it were a natural feature rather than an artefact of administrative convenience to ensure closure.

This omission is not an incidental gap; it is the central failure. A methods review that never asks whether its basic quantities exist as measures cannot, by definition, review methods in any scientific sense. It can only report, compare, and recommend procedures. That is exactly what CHERE does. It describes that most HTA agencies use cost-utility analysis. It notes how they derive utilities, which instruments they favor, how they discount, how they deal with uncertainty, and how they approach special populations. But it never interrogates whether the numbers being manipulated satisfy the axioms that would make arithmetic legitimate.

The result is a curious inversion. The CHERE paper has a great deal to say about "methods," "processes," and "alignment" but nothing to say about measurement. It asks whether Australia is consistent with international practice; it does not ask whether international practice is consistent with measurement theory. It implicitly defines "best practice" as convergence around the QALY paradigm, then concludes that Australia is broadly on track. In doing so, it provides the PBAC and the Department of Health with exactly the reassurance they were looking for: the comfort that their long-standing approach is broadly in line with what everyone else is doing.

From the perspective of the diagnostic logit instrument, this is simply the memetic colonization of Australian HTA in written form. The CHERE report documents how the QALY memeplex spread and diversified across countries; it never recognizes that the entire memeplex rests on non-measures. It takes the existence of utilities and QALYs as a solved problem and focuses instead on refinements in discounting, extrapolation, scenario analysis, and handling of uncertainty. But such refinements are irrelevant if the underlying constructs are not measures. Adjusting discount rates or improving parametric assumptions in a cost-utility model is like polishing the paintwork on a car with no engine.

The Consultation 1 synthesis authored by CHERE reinforces the same pattern. That report summarizes stakeholder views on what is "working well" and where improvements might be needed in Australia's HTA system. Stakeholders call for greater transparency, more timely processes, better use of real-world data, and clearer guidance. No one raises the possibility that utilities and QALYs are not measures. The CHERE authors faithfully relay these priorities and help embed them into the final HTA Review recommendations for ongoing "rolling review" of methods and guidelines. Thus, the system commits itself to continuous improvement of a non-measurement architecture, with no conceptual mechanism for ever questioning the architecture itself.

This is why the CHERE paper must be understood, not as a missed opportunity but as part of the problem. The PBAC logits in Logit Working Paper No, 34 show that the committee has no possession of the axioms of measurement: unidimensionality and measurement preceding arithmetic at deep negative values, Rasch-related items at the floor, and strong positive reinforcement for claims that QALYs are ratio-like, aggregable, and suitable for simulation-based cost-effectiveness. The CHERE report is written from within that same conceptual space. It could never have diagnosed the failure because it breathes the same air.

What the CHERE work does accomplish, unintentionally, is to make the epistemic capture visible. By explicitly mapping Australia against other QALY-based systems and by recommending ongoing refinement rather than foundational reassessment, it confirms that the HTA community no longer regards measurement as a live question. "Economic evaluation" is presented as a settled framework in which utilities and QALYs are unquestioned inputs and the only interesting problems are technical: how to model more elegantly, how to share work across agencies, how to better manage uncertainty. In other words, an improved brand of car polish.

For a Logit Working Paper grounded in representational measurement theory, this places CHERE in a very specific role: the expert witness for non-measurement. The CHERE economic evaluation paper does not rescue the PBAC from the logit diagnosis; it corroborates it. It shows that when Australia commissions its most prominent academic health economists to review HTA methods, they reproduce the same foundational error, now wrapped in the language of international comparison and continuous improvement.

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

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

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

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

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

[v] Centre for Health Economics Research and Evaluation (CHERE). *HTA Methods: Economic Evaluation*. HTA Methods Paper No. 5. Sydney: University of Technology Sydney, 2024.