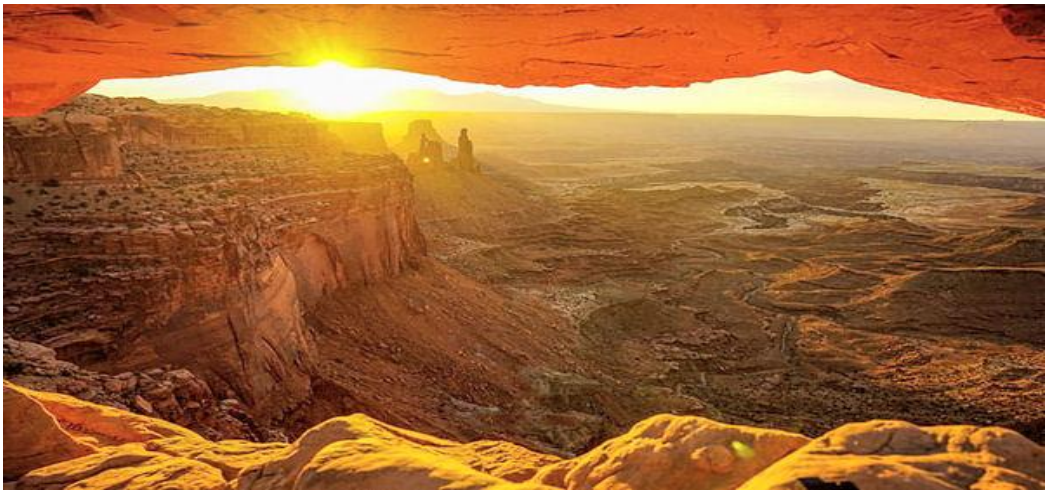
MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION



REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

AUSTRALIA: A NATIONAL CONSENSUS ON 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 32 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 evaluate the epistemic foundations of health technology assessment in Australia through application of a 24-item canonical statement diagnostic grounded in representational measurement theory. Rather than examining individual agencies or specific submissions, the analysis adopts a national perspective, treating Australia as a single knowledge environment shaped by shared academic training, methodological conventions, policy guidance, and evaluative norms. The purpose is to determine whether the quantitative claims routinely employed in Australian HTA satisfy the axiomatic requirements necessary for admissible arithmetic, falsifiable evaluation, and the accumulation of objective knowledge regarding therapy impact.

Using endorsement probabilities transformed through a canonical probability–logit mapping, the study interrogates whether the Australian HTA knowledge base recognizes the precedence of measurement over arithmetic, the distinction between manifest and latent attributes, the scale-type constraints governing permissible mathematical operations, and the unique role of Rasch measurement in constructing valid measures of latent traits. The intent is not to critique policy outcomes or decision thresholds, but to assess whether the analytical framework itself rests on scientifically defensible measurement principles.

The findings are unequivocal. The Australian HTA knowledge base exhibits a systematic inversion of scientific reasoning in which arithmetic operations are routinely applied in the absence of demonstrable measurement. Core axioms of representational measurement, including unidimensionality, scale-type admissibility, and the requirement that measurement precede arithmetic, receive weak or near-floor endorsement. In contrast, propositions that violate these axioms, including the treatment of utilities as interval or ratio measures, the aggregation of QALYs, the summation of ordinal responses, and the legitimacy of reference-case simulation outputs, are strongly reinforced.

The resulting logit profile does not indicate isolated misunderstanding or methodological ambiguity. It reveals a coherent belief system in which false measurement is normalized and

protected, while the principles required for falsification and cumulative scientific learning are structurally excluded. Australia's HTA framework therefore functions not as a measurement-based evaluative science, but as a numerically formalized decision technology incapable, in principle, of generating evaluable or reproducible claims of therapy impact.

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 AUSTRALIAN HTA KNOWLEDGE BASE

For the purposes of this assessment, the Australian HTA knowledge base is defined as the shared and recurrent set of concepts, assumptions, methods, and evaluative practices that govern how therapy impact is quantified, interpreted, and legitimized within national decision making. It is not confined to any single institution or guideline. Rather, it emerges from the collective output of academic health economics programs, HTA training curricula, consultancy practices, journal publications, submission templates, and policy conventions that together define what is regarded as acceptable evidence.

This knowledge base is structured around the reference-case paradigm imported largely from the United Kingdom during the 1990s and subsequently institutionalized within Australian HTA practice. At its center lies the presumption that health outcomes can be represented through preference-weighted health state descriptions, transformed into utilities, aggregated into QALYs, and combined with cost data to produce decision-relevant ratios. These numerical objects are treated as commensurable across disease areas, populations, and time horizons despite the absence of demonstrated measurement properties.

Within this framework, numerical form is routinely conflated with measurement. Subjective responses derived from questionnaires, preference elicitation exercises, and composite instruments are assumed to possess quantitative meaning once expressed numerically. Scale-type distinctions are rarely operationalized. Ordinal categories are treated as if they carried equal intervals, and negative preference values are permitted while the constructs are simultaneously described as ratio measures. The distinction between ordering and measuring is therefore blurred or ignored.

Latent attributes such as quality of life, functioning, symptom burden, and wellbeing are invoked extensively but are not formally constructed as measurable quantities. These attributes are treated as if they existed on implicit continua without requiring demonstration of unidimensionality or invariance. The knowledge base does not require transformation through a measurement model capable of producing invariant units. As a result, latent traits are evaluated through scores rather than measures.

The absence of Rasch measurement is decisive in defining the boundaries of admissible practice. Although Rasch models have been available for decades and provide the only scientifically coherent means of constructing ratio-scaled measures of latent traits, they do not function as a gatekeeping standard within Australian HTA. Their implications for scale validity, invariance, and arithmetic admissibility are not incorporated into submission requirements or evaluative norms. Instead, psychometric concepts such as reliability, responsiveness, and construct validity are used as substitutes for measurement itself.

The knowledge base also redefines falsification. Rather than requiring empirical refutation of claims through observable outcomes, Australian HTA relies heavily on reference-case simulation

models in which robustness is assessed through scenario analysis and parameter variation. This internal consistency is treated as scientific credibility, even though such models cannot, by design, generate falsifiable claims in the Popperian sense.

Taken together, these features define a stable epistemic environment in which numerical storytelling substitutes for measurement-based inference. The knowledge base reproduces itself through education, publication, and professional socialization, ensuring continuity of method while excluding the foundational axioms required for scientific evaluation. It is within this environment that Australian HTA operates—and it is this environment that the canonical diagnostic reveals to be fundamentally incompatible with the requirements of representational measurement and 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: AUSTRALIA

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 AUSTRALIA

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	0.90	+2.20
MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC	1	0.15	-1.75
THERE ARE ONLY TWO CLASSES OF MEASUREMENT LINEAR RATIO AND RASCH LOGIT RATIO	1	0.10	-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	0.85	+1.75
NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.60	+0.40
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.75	+1.10
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

AUSTRALIA: SYSTEMATIC DISPLACEMENT OF MEASUREMENT

The results of the canonical 24-item diagnostic for Australia are not ambiguous, marginal, or transitional. They represent a fully consolidated belief system. The Australian health technology assessment environment, across government agencies, academic training, consulting practice, and policy discourse, exhibits the same structural inversion of science observed internationally, but with a distinctive intensity rooted in early and enthusiastic adoption of the NICE reference-case framework. The logit profile in Table 1 demonstrates not a misunderstanding of measurement principles, but their systematic displacement.

At the center of the profile lies a profound epistemic reversal. Propositions that define the preconditions for quantitative science collapse toward the floor of endorsement, while propositions that violate those preconditions are reinforced at or near the ceiling. Measurement, which must logically precede arithmetic, is rejected at $p = 0.15$ (-1.75). The requirement that representational axioms govern permissible arithmetic collapses to the same level. The proposition that multiplication requires ratio measures, the single rule that determines whether cost-effectiveness arithmetic is even allowable is similarly rejected at $p = 0.15$. These values are not statistical noise. They mark the functional absence of measurement as a governing constraint in the Australian HTA knowledge base.

At the same time, the arithmetic products that depend on violating these axioms receive overwhelming reinforcement. The QALY is treated as a ratio measure at $p = 0.90$ ($+2.20$). QALYs are treated as aggregable at $p = 0.85$ ($+1.75$). Summated subjective scores are treated as ratio measures at $p = 0.90$ ($+2.20$). Likert scale summation is explicitly endorsed as ratio measurement at the same ceiling. Preference algorithms are treated as interval measures despite negative values and the absence of a true zero. What emerges is not confusion, but doctrinal consistency: arithmetic is protected; measurement is sacrificed. This pattern defines the Australian HTA memeplex. The system does not accidentally misuse numbers. It requires their misuse in order to function.

The Table 1 diagnostic reveals with clarity the role of the QALY as the central enabling fiction. The proposition that ratio measures can have negative values is endorsed at $p = 0.85$ ($+1.75$), an extraordinary signal of how deeply the Australian belief in HTA has normalized category error. In no domain of physical science could a quantity with a true zero meaningfully take negative values. Yet Australian HTA treats “states worse than dead” as routine while insisting that the construct remains ratio-scaled. This contradiction is not debated; it is institutionalized.

The importance of this cannot be overstated. If utilities are ordinal and they are then cost per QALY ratios are undefined. The diagnostic shows that Australia's knowledge base resolves this contradiction not by rejecting the arithmetic, but by rejecting the rules that would disallow it. This is the essence of numerical or bunyip storytelling.

The treatment of unidimensionality further exposes the epistemic collapse. Measures must be unidimensional is endorsed at only p = 0.25 (−1.10), while time-trade-off preferences are simultaneously treated as unidimensional at p = 0.85 (+1.75). This contradiction is revealing. Unidimensionality is not treated as an empirical requirement to be demonstrated; it is treated as a rhetorical attribute assigned when needed. Multiattribute health state descriptions are declared to represent a single attribute continuum because the model requires a single continuum. The property is asserted, not tested.

Nowhere is the scientific failure more visible than in the Rasch block. All Rasch-related propositions collapse to near-floor or floor endorsement. The statement that only two admissible measurement forms exist, linear ratio scales for manifest attributes and Rasch logit ratio scales for latent traits, sits at p = 0.10 (−2.20). The statement that subjective responses can only be transformed into interval measurement via Rasch rules sits at the same level. The statement that Rasch logit scales are the only admissible basis for latent-trait therapy claims collapses identically. And the statement that Rasch rules are identical to the axioms of representational measurement likewise reaches −2.20. This is not ignorance. It is profound quarantined silence.

Rasch is not rejected through argument; it is excluded by agreed silence. The Australian HTA knowledge system cannot permit Rasch to become normative because Rasch would invalidate the entire inventory of instruments, utilities, mappings, and models currently in use. Acceptance of Rasch would immediately expose the non-measurement status of preference-based utilities, multiattribute instruments, and quality-adjusted life years. The system therefore does what all memeplexes do when threatened: it marginalizes the threat without confronting it.

The diagnostic also clarifies why Australian HTA appears methodologically sophisticated while remaining scientifically empy. There is moderate endorsement of rejecting non-falsifiable claims (p = 0.60), yet strong endorsement that reference-case simulations generate falsifiable claims (p = 0.75). This is the decisive substitution. Falsification is redefined. Instead of empirical refutation through observed outcomes, falsification becomes internal variation within a model. Scenario analysis replaces exposure to reality. The scientific revolution of the 17th century need not have occurred. This rejection allows the system to speak the language of science while abandoning its substance. Claims are never wrong; they are merely sensitive. Evidence is never refuted; it is re-parameterized. This is not Popperian science. It is scholasticism with software.

Australia adopted this framework early and enthusiastically. The reasons are not mysterious. The reference case provided exactly what policy makers wanted: administrative closure. Pricing and access decisions could be justified without requiring long-term empirical confirmation. The framework delivered decisions without demanding that claims remain provisional. Once the model ran and the ICER cleared an implicit threshold, the matter was closed. This is why the imported NICE reference case was so attractive and so dangerous. It offered certainty without truth. The Australian knowledge base therefore did not evolve toward measurement. It evolved away from

it. The axioms that would have constrained arithmetic were inconvenient. They threatened closure. They threatened timelines. They threatened negotiation processes. They threated the Australian bureaucracy and health system decision makers. So they were ignored.

The logit profile demonstrates that this was not accidental. The near-ceiling endorsement of false propositions and near-floor rejection of true ones defines a belief system, not a technical error. This is why reform efforts that focus on "better models," "improved utilities," or "more sophisticated mapping" are doomed. They operate entirely within the same inverted logic. They accept arithmetic as given and seek only to refine its inputs. But refinement of false measurement does not produce truth; it produces more elaborate fiction.

Australia's HTA environment now stands exposed. The LLM-based diagnostic does what four decades of peer review could not: it reveals the structure of belief across the entire corpus. Not disagreement. Not debate. But systematic silence where axioms should be. The implications are profound.

First, Australian HTA cannot claim scientific legitimacy. Decisions may be procedurally consistent, administratively defensible, and politically convenient; but they are not grounded in measurable quantities. Without measurement, there can be no falsification. Without falsification, there can be no accumulation of objective knowledge. What exists instead is repetition; an endless future of reference case closed model claims.

Second, the system cannot self-correct. Because its core constructs are not measurable, no amount of real-world evidence can ever confirm or refute them. Post-listing studies cannot validate QALYs. Registries cannot rescue utilities. Longitudinal data cannot repair ordinal arithmetic. The system is deliberately epistemically sealed.

Third, Australia's HTA framework teaches future generations to confuse computation with science. Students learn models before measurement. Software before axioms. Thresholds before scale types. This reproduces the memeplex automatically. A future that rejects any role for the evolution of objective knowledge for disease specific therapy impact assessments

Fourth, Australia's position mirrors almost exactly that of NICE. This is not coincidence. Australia did not independently reason its way to the reference case. Like the rabbit, it imported it. The diagnostic confirms this inheritance: the same items collapse, the same falsehoods rise, the same silences persist.

The Australian case therefore serves as a national-level demonstration of how a global memeplex functions. Once installed, it requires no enforcement. It sustains itself through training, publication norms, and professional identity. Yet the diagnostic also reveals something else: the system's fragility. Because its legitimacy depends entirely on unexamined assumptions, once those assumptions are named, the framework cannot defend itself. It cannot explain why multiplication is permissible without ratio scales. It cannot explain why ordinal preferences become interval measures by algorithm. It cannot explain how latent attributes are measured without Rasch. It cannot explain how simulation outputs become falsifiable. It can only appeal to precedent. That is not science.

WHY AUSTRALIA ENTHUSIASTICALLY ENDORSED THE NICE REFERENCE CASE: BUYING WHAT COULD NOT BE SEEN

The question that follows inevitably from the Australian diagnostic assessment is not whether the NICE reference case fails the axioms of representational measurement, that failure is now beyond dispute, but why Australia endorsed it so readily and with so little resistance. The answer is uncomfortable, yet unavoidable: Australia had no clear perception of what it was buying. The framework was adopted not because its scientific foundations had been examined and judged adequate, but because the conditions necessary to recognize their absence did not exist within the Australian HTA knowledge base at the time of adoption.

By the early 1990s, Australia faced the same pressures confronting health systems across the developed world. Pharmaceutical expenditure was rising, new therapies were entering the market with limited post-launch evidence, and governments demanded mechanisms that could justify pricing and access decisions in a manner that appeared systematic, defensible, and consistent. What policymakers required was not discovery or falsification, but closure. They needed a process that could deliver decisions within constrained timelines and political tolerance. The NICE reference case appeared to offer precisely that: a standardized evaluative framework, a single outcome metric, and a method capable of transforming uncertainty into a definitive recommendation.

Crucially, the reference case was not presented to Australia as a scientific theory. It arrived as an administrative solution already endorsed by a respected foreign authority. It came wrapped in the prestige of UK institutions, supported by academic centers at York and elsewhere, and framed as international best practice. At no point was it offered as a provisional hypothesis subject to falsification. It was presented as a mature technology of governance; a method already settled elsewhere and therefore safe to import. In this form, the reference case did not invite scrutiny. It invited alignment.

Australia therefore did not evaluate the reference case as a claim about measurement. It evaluated it as a policy instrument. The distinction is decisive. Scientific claims invite questions of validity, structure, and falsifiability. Policy instruments invite questions of feasibility, comparability, and administrative convenience. Once the framework was categorized as the latter, the former questions disappeared entirely from view.

This disappearance was made possible by a deeper epistemic condition: the near total absence of operational knowledge of representational measurement theory within Australian health economics and HTA training. While Stevens' scale typology had been published decades earlier, it was not taught as a binding constraint on arithmetic. The distinction between ordinal, interval, and ratio scales was treated as descriptive taxonomy, not as a set of non-negotiable rules governing what can and cannot be done with numbers. Rasch measurement, the only model capable of constructing invariant quantitative measures for latent attributes, was effectively invisible. As a result, Australian analysts lacked the conceptual vocabulary required to ask the most basic question of all: what kind of number is this?

In such an environment, the NICE framework could not appear incoherent, because incoherence requires contrast. One cannot recognize violation of measurement axioms if those axioms are unknown. The Australian system therefore did not reject representational measurement; it simply never encountered it. This absence created a perfect epistemic vacuum into which the reference case could be inserted without friction. Numbers were assumed to be quantitative by default. Precision was mistaken for measurement. Mathematical form was conflated with empirical meaning.

The result was that Australia effectively purchased a framework whose core properties it could not see. Utilities were treated as interval measures without demonstration. QALYs were treated as ratio quantities despite permitting negative values and lacking a true zero. Composite health state descriptions were multiplied by time as if multiplication were permitted. None of this appeared problematic, not because it was defensible, but because the rules it violated were unknown to those applying it.

The reference case also carried a powerful ethical narrative that further obscured its mathematical deficiencies. The idea of equal value for equal health gain resonated strongly within Australian policy culture. The QALY functioned not merely as a technical device but as a symbol of fairness, neutrality, and transparency. That symbolism filled the epistemic gap. Where measurement theory was absent, moral language substituted. The framework felt just, and therefore it was assumed to be sound.

Academic reinforcement quickly followed. Australian researchers trained in UK centers returned with methodological templates rather than theories. Journals published work conforming to international norms. Teaching programs incorporated the reference case as standard practice. Once embedded in curricula and publication pipelines, the framework ceased to appear imported at all. It became naturalized. New generations of analysts encountered it not as a choice that had been made, but as the way HTA is done.

At that point, the possibility of questioning what had been bought disappeared altogether. One does not normally interrogate the foundations of infrastructure that appears already complete. The reference case had become infrastructure. Its assumptions were not argued; they were inherited. Its arithmetic was not justified; it was repeated. Falsification was quietly replaced by sensitivity analysis, a technique that varies assumptions without ever testing whether the underlying quantities exist as measures in the first place.

From today's perspective, this history can appear astonishing. How could a national system adopt a framework so deeply incompatible with the axioms of measurement? The answer is that Australia did not knowingly endorse false measurement. It endorsed a system whose falsity was invisible to it at the time. The purchase was made without specification of the product's epistemic contents. No one examined the fine print because no one knew what the fine print should contain.

This is why the current moment is fundamentally different. The emergence of large language model diagnostics has altered the evidentiary landscape. For the first time, it is possible to interrogate entire national knowledge bases and observe, quantitatively and reproducibly, what is reinforced, what is denied, and what is absent. The results show not debate, not disagreement, but

systematic silence on the axioms that make measurement possible. What was once hidden can now be demonstrated.

The implication for Australia is profound. The question is no longer whether the NICE reference case is defensible. It is whether a system that now understands what it purchased can continue to use it in good faith. Once the absence of measurement is visible, continued reliance on arithmetic built upon it becomes a choice rather than an inheritance. Australia may not have known what it was buying in the 1990s. It does now. The issue is therefore no longer historical error. It is present responsibility.

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