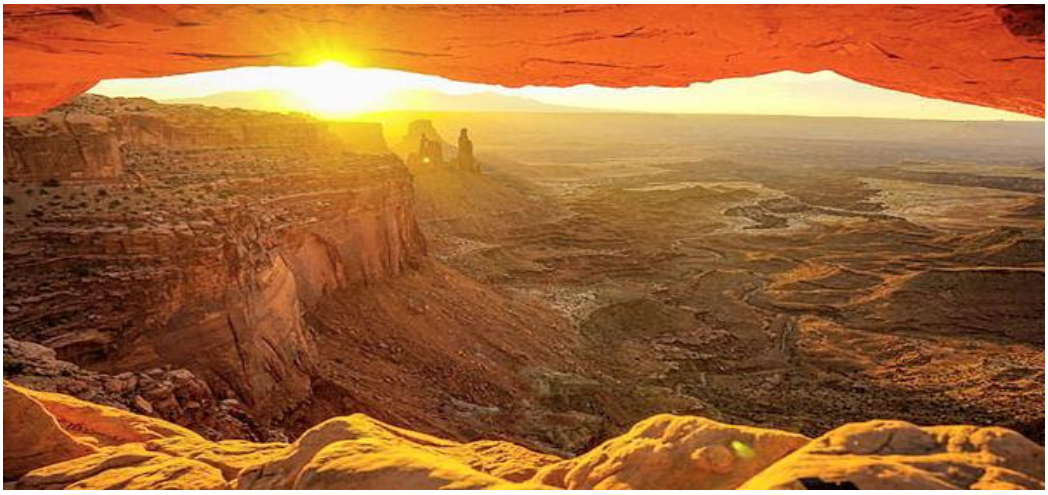
MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION



REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

AUSTRALIA: THE ABSENCE OF REPRESENTATIONAL MEASUREMENT AND THE CENTRE FOR APPLIED HEALTH ECONOMICS (CAHE) GRIFFITH UNIVERSITY

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

LOGIT WORKING PAPER No 727 FEBRUARY 2026

www.maimonresearch.com

Tucson AZ

FOREWORD

HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

The Centre for Applied Health Economics (CAHE) at Griffith University is an applied research center focused on health economics, health technology assessment (HTA), health services evaluation, and health system policy. The Centre positions itself as supporting evidence-informed healthcare decision-making through economic evaluation, outcomes research, and research translation. Its activities span collaborations with government agencies, healthcare providers, industry, and academic partners, with a particular emphasis on improving the efficiency and effectiveness of healthcare delivery.

The Centre's work includes cost-effectiveness analysis, burden of disease studies, evaluation of models of care, health service utilization analysis, and the assessment of healthcare interventions and technologies. It is also involved in methodological research and training in applied health economics, contributing to policy development and health system planning both within Australia and internationally.

A significant component of the Centre's activity is the translation of research into policy and practice. This includes support for reimbursement and formulary-related decision-making, assessment of healthcare programs, and the evaluation of interventions aimed at improving patient outcomes and resource allocation. The Centre also contributes to postgraduate and professional education in health economics and HTA-related methods, reinforcing its role as both a research and teaching entity within the Australian health policy environment.

The objective of this study was to evaluate the extent to which the HTA-related knowledge base associated with the CAHE satisfies the requirements of representational measurement. The assessment focused on whether the conceptual and methodological framework supporting teaching, research, and policy engagement recognizes the conditions necessary for valid quantitative claims. Particular attention was given to the treatment of scale properties, dimensional homogeneity, latent constructs, arithmetic operations, and the role of falsifiability in claims for therapy impact. The study employed a structured interrogation using a canonical 24-item set of true/false diagnostic statements, with endorsement probabilities transformed into normalized logits to identify patterns of conceptual reinforcement and omission within the knowledge base.

The interrogation identified a consistent pattern of measurement inversion within the Griffith HTA-related knowledge base. Core principles of representational measurement including unidimensionality, ratio scale requirements, dimensional homogeneity, and the transformation of latent constructs through Rasch measurement received weak endorsement. At the same time, constructs central to conventional HTA, including utilities, QALYs, and reference case simulation models, remained embedded within the broader analytical framework despite their failure to satisfy the conditions required for measurement. Rasch measurement was effectively absent from the knowledge base, while cost-utility analysis and simulation modelling continued to be treated as legitimate approaches to quantitative decision-making. The findings indicate that the Griffith HTA

framework reflects the same pattern observed across Australian and international HTA interrogations: the normalization of arithmetic without measurement.

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 CAHE HTA KNOWLEDGE BASE

The HTA-related knowledge base associated with the Centre for Applied Health Economics reflects the broader applied health economics and policy evaluation framework common within Australian HTA research environments. The Centre positions itself within the domains of economic evaluation, health services research, policy translation, and healthcare system efficiency. Its activities include cost-effectiveness analysis, health program evaluation, burden of disease assessment, health service utilization studies, and research translation aimed at supporting healthcare decision-making.

The knowledge base therefore encompasses the conventional methodological structure associated with applied health economics and HTA. This includes the use of cost-effectiveness and cost-utility analysis, decision-analytic modelling, health state utility measurement, simulation frameworks, and patient-reported outcome instruments. It also incorporates literature and methods associated with reimbursement evaluation, healthcare resource allocation, policy implementation, and comparative assessment of healthcare interventions.

The interrogation suggests that this framework operates within the accepted international HTA paradigm, where quantitative claims are constructed primarily through the manipulation of utilities, modeled costs, probabilities, and projected outcomes. Within this paradigm, the emphasis is placed on analytical techniques, simulation structures, and economic modeling rather than on the foundational requirements of measurement itself. The knowledge base appears to assume that the presence of numerical values is sufficient to support arithmetic operations and quantitative interpretation.

A defining feature of the knowledge base is the continued reliance on multiattribute health state descriptions and utility scoring systems, particularly within cost-utility analysis and QALY-based evaluation. These constructs are treated as if they provide measurable quantities capable of supporting multiplication, aggregation, and comparative ratios. At the same time, there is little evidence that the framework explicitly addresses dimensional homogeneity, scale validity, or the conditions required for ratio measurement.

The interrogation also indicates that the treatment of latent constructs remains rooted in ordinal scoring approaches rather than in formal measurement theory. Patient-reported outcomes and quality-of-life measures appear to be handled primarily through summed scores and utility transformations, with little recognition of the need for invariant measurement frameworks such as the Rasch model. As a consequence, latent attributes are incorporated into evaluative models without establishing whether the underlying observations have been transformed into measures.

Another characteristic of the Griffith knowledge base is its orientation toward policy relevance and research translation. The Centre emphasizes applied outcomes and contribution to healthcare decision-making, including interaction with reimbursement and formulary systems. However, the

interrogation suggests that this policy orientation remains embedded within the broader reference case framework that dominates international HTA. As a result, the conceptual foundations of the knowledge base continue to align with simulation modelling and utility-based analysis despite their failure to satisfy representational measurement standards.

Overall, the Griffith HTA knowledge base reflects a sophisticated applied health economics framework that is strongly aligned with conventional international HTA practice. However, the interrogation indicates that the framework lacks explicit recognition of the measurement principles necessary to support scientifically valid quantitative claims. The result is a knowledge base characterized by methodological complexity but constrained by the systematic absence of representational measurement.

CATEGORICAL PROBABILITIES

In the present application, the interrogation is tightly bounded. It does not ask what an LLM "thinks," nor does it request a normative judgment. Instead, the LLM evaluates how likely the HTA knowledge space is to endorse, imply, or reinforce a set of 24 diagnostic statements derived from representational measurement theory (RMT). Each statement is objectively TRUE or FALSE under RMT. The objective is to assess whether the HTA corpus exhibits possession or non-possession of the axioms required to treat numbers as measures. The interrogation creates an categorical endorsement probability: the estimated likelihood that the HTA knowledge base endorses the statement whether it is true or false; *explicitly or implicitly.*

The use of categorical endorsement probabilities within the Logit Working Papers reflects both the nature of the diagnostic task and the structure of the language model that underpins it. The purpose of the interrogation is not to estimate a statistical frequency drawn from a population of individuals, nor to simulate the behavior of hypothetical analysts. Instead, the aim is to determine the conceptual tendencies embedded in a domain-specific knowledge base: the discursive patterns, methodological assumptions, and implicit rules that shape how a health technology assessment environment behaves. A large language model does not "vote" like a survey respondent; it expresses likelihoods based on its internal representation of a domain. In this context, endorsement probabilities capture the strength with which the knowledge base, as represented within the model, supports a particular proposition. Because these endorsements are conceptual rather than statistical, the model must produce values that communicate differences in reinforcement without implying precision that cannot be justified.

This is why categorical probabilities are essential. Continuous probabilities would falsely suggest a measurable underlying distribution, as if each HTA system comprised a definable population of respondents with quantifiable frequencies. But large language models do not operate on that level. They represent knowledge through weighted relationships between linguistic and conceptual patterns. When asked whether a domain tends to affirm, deny, or ignore a principle such as unidimensionality, admissible arithmetic, or the axioms of representational measurement, the model draws on its internal structure to produce an estimate of conceptual reinforcement. The precision of that estimate must match the nature of the task. Categorical probabilities therefore provide a disciplined and interpretable way of capturing reinforcement strength while avoiding the illusion of statistical granularity.

The categories used, values such as 0.05, 0.10, 0.20, 0.50, 0.75, 0.80, and 0.85, are not arbitrary. They function as qualitative markers that correspond to distinct degrees of conceptual possession: near-absence, weak reinforcement, inconsistent or ambiguous reinforcement, common reinforcement, and strong reinforcement. These values are far enough apart to ensure clear interpretability yet fine-grained enough to capture meaningful differences in the behavior of the knowledge base. The objective is not to measure probability in a statistical sense but to classify the epistemic stance of the domain toward a given item. A probability of 0.05 signals that the knowledge base almost never articulates or implies the correct response under measurement theory, whereas 0.85 indicates that the domain routinely reinforces it. Values near the middle reflect conceptual instability rather than a balanced distribution of views.

Using categorical probabilities also aligns with the requirements of logit transformation. Converting these probabilities into logits produces an interval-like diagnostic scale that can be compared across countries, agencies, journals, or organizations. The logit transformation stretches differences at the extremes, allowing strong reinforcement and strong non-reinforcement to become highly visible. Normalizing logits to the fixed ±2.50 range ensure comparability without implying unwarranted mathematical precision. Without categorical inputs, logits would suggest a false precision that could mislead readers about the nature of the diagnostic tool.

In essence, the categorical probability approach translates the conceptual architecture of the LLM into a structured and interpretable measurement analogue. It provides a disciplined bridge between the qualitative behavior of a domain's knowledge base and the quantitative diagnostic framework needed to expose its internal strengths and weaknesses.

The LLM computes these categorical probabilities from three sources:

1. **Structural content of HTA discourse**
   If the literature repeatedly uses ordinal utilities as interval measures, multiplies non-quantities, aggregates QALYs, or treats simulations as falsifiable, the model infers high reinforcement of these false statements.
2. **Conceptual visibility of measurement axioms**
   If ideas such as unidimensionality, dimensional homogeneity, scale-type integrity, or Rasch transformation rarely appear, or are contradicted by practice, the model assigns low endorsement probabilities to TRUE statements.
3. **The model's learned representation of domain stability**
   Where discourse is fragmented, contradictory, or conceptually hollow, the model avoids assigning high probabilities. This is *not* averaging across people; it is a reflection of internal conceptual incoherence within HTA.

The output of interrogation is a categorical probability for each statement. Probabilities are then transformed into logits [$\ln(p/(1-p))$], capped to ±4.0 logits to avoid extreme distortions, and normalized to ±2.50 logits for comparability across countries. A positive normalized logit indicates reinforcement in the knowledge base. A negative logit indicates weak reinforcement or conceptual absence. Values near zero logits reflect epistemic noise.

Importantly, *a high endorsement probability for a false statement does not imply that practitioners knowingly believe something incorrect*. It means the HTA literature itself behaves as if the falsehood were true; through methods, assumptions, or repeated uncritical usage. Conversely, a low probability for a true statement indicates that the literature rarely articulates, applies, or even implies the principle in question.

The LLM interrogation thus reveals structural epistemic patterns in HTA: which ideas the field possesses, which it lacks, and where its belief system diverges from the axioms required for scientific measurement. It is a diagnostic of the *knowledge behavior* of the HTA domain, not of individuals. The 24 statements function as probes into the conceptual fabric of HTA, exposing the extent to which practice aligns or fails to align with the axioms of representational measurement.

INTERROGATION STATEMENTS

Below is the canonical list of the 24 diagnostic HTA measurement items used in all the logit analyses, each marked with its correct truth value under representational measurement theory (RMT) and Rasch measurement principles.

This is the definitive set used across the Logit Working Papers.

Measurement Theory & Scale Properties

1. Interval measures lack a true zero — TRUE
2. Measures must be unidimensional — TRUE
3. Multiplication requires a ratio measure — TRUE
4. Time trade-off preferences are unidimensional — FALSE
5. Ratio measures can have negative values — FALSE
6. EQ-5D-3L preference algorithms create interval measures — FALSE
7. The QALY is a ratio measure — FALSE
8. Time is a ratio measure — TRUE

Measurement Preconditions for Arithmetic

9. Measurement precedes arithmetic — TRUE
10. Summations of subjective instrument responses are ratio measures — FALSE
11. Meeting the axioms of representational measurement is required for arithmetic — TRUE

Rasch Measurement & Latent Traits

12. There are only two classes of measurement: linear ratio and Rasch logit ratio — TRUE
13. Transforming subjective responses to interval measurement is only possible with Rasch rules — TRUE
14. Summation of Likert question scores creates a ratio measure — FALSE

Properties of QALYs & Utilities

15. The QALY is a dimensionally homogeneous measure — FALSE
16. Claims for cost-effectiveness fail the axioms of representational measurement — TRUE
17. QALYs can be aggregated — FALSE

Falsifiability & Scientific Standards

18. Non-falsifiable claims should be rejected — TRUE
19. Reference-case simulations generate falsifiable claims — FALSE

Logit Fundamentals

20. The logit is the natural logarithm of the odds-ratio — TRUE

Latent Trait Theory

21. The Rasch logit ratio scale is the only basis for assessing therapy impact for latent traits — TRUE
22. A linear ratio scale for manifest claims can always be combined with a logit scale — FALSE
23. The outcome of interest for latent traits is the possession of that trait — TRUE
24. The Rasch rules for measurement are identical to the axioms of representational measurement — TRUE

AI LARGE LANGUAGE MODEL STATEMENTS: TRUE OR FALSE

Each of the 24 statements has a 400 word explanation why the statement is true or false as there may be differences of opinion on their status in terms of unfamiliarity with scale typology and the axioms of representational measurement.

The link to these explanations is: https://maimonresearch.com/ai-llm-true-or-false/

INTERPRETING TRUE STATEMENTS

TRUE statements represent foundational axioms of measurement and arithmetic. Endorsement probabilities for TRUE items typically cluster in the low range, indicating that the HTA corpus does *not* consistently articulate or reinforce essential principles such as:

- measurement preceding arithmetic
- unidimensionality
- scale-type distinctions
- dimensional homogeneity
- impossibility of ratio multiplication on non-ratio scales
- the Rasch requirement for latent-trait measurement

Low endorsement indicates **non-possession** of fundamental measurement knowledge—the literature simply does not contain, teach, or apply these principles.

INTERPRETING FALSE STATEMENTS

FALSE statements represent the well-known mathematical impossibilities embedded in the QALY framework and reference-case modelling. Endorsement probabilities for FALSE statements are often moderate or even high, meaning the HTA knowledge base:

- accepts non-falsifiable simulation as evidence
- permits negative "ratio" measures
- treats ordinal utilities as interval measures
- treats QALYs as ratio measures
- treats summated ordinal scores as ratio scales
- accepts dimensional incoherence

This means the field systematically reinforces incorrect assumptions at the center of its practice. *Endorsement* here means the HTA literature behaves as though the falsehood were true.

2. SUMMARY OF FINDINGS FOR TRUE AND FALSE ENDORSEMENTS: CENTRE FOR APPLIED HEALTH ECONOMICS (CAHE) GRIFFITH UNIVERSITY

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.

CAHE: THE ABSENCE OF REPRESENTATIONAL MEASUREMENT AND THE ENDORSEMENT OF FALSE MEASUREMENT

The interrogation of the HTA-related knowledge base associated with the Centre for Applied Health Economics reveals a pattern that is now familiar across Australian and international HTA institutions: the systematic absence of representational measurement standards and the corresponding normalization of measurement inversion (Table 1). The results are not an assessment of the competence or productivity of the Centre. Rather, they evaluate whether the conceptual structure supporting teaching, research, publication, and policy engagement satisfies the minimum requirements for meaningful quantitative claims.

TABLE 1: ITEM STATEMENT, RESPONSE, ENDORSEMENT AND NORMALIZED LOGITS CENTRE FOR APPLIED HEALTH ECONOMICS (CAHE) GRIFFITH UNIVERSITY

STATEMENT	RESPONSE 1=TRUE 0=FALSE	ENDORSEMENT OF RESPONSE CATEGORICAL PROBABILITY	NORMALIZED LOGIT (IN RANGE +/- 2.50)
INTERVAL MEASURES LACK A TRUE ZERO	1	0.30	-0.85
MEASURES MUST BE UNIDIMENSIONAL	1	0.25	-1.10
MULTIPLICATION REQUIRES A RATIO MEASURE	1	0.20	-1.40
TIME TRADE-OFF PREFERENCES ARE UNIDIMENSIONAL	0	0.75	+1.10
RATIO MEASURES CAN HAVE NEGATIVE VALUES	0	0.80	+1.40
EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES	0	0.65	+0.60
THE QALY IS A RATIO MEASURE	0	0.40	-0.40
TIME IS A RATIO MEASURE	1	0.85	+1.75
MEASUREMENT PRECEDES ARITHMETIC	1	0.25	-1.10
SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES	0	0.75	+1.10
MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC	1	0.25	-1.10
THERE ARE ONLY TWO CLASSES OF MEASUREMENT LINEAR RATIO AND RASCH LOGIT RATIO	1	0/05	-2.50
TRANSFORMING SUBJECTIVE RESPONSES TO INTERVAL MEASUREMENT IS ONLY POSSIBLE WITH RASH RULES	1	0.05	-2.50
SUMMATION OF LIKERT QUESTION SCORES CREATES A RATIO MEASURE	0	0.75	+1.10
THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE	0	0.40	-0.40
CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.20	-1,40
QALYS CAN BE AGGREGATED	0	0.45	-0.20

NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.65	+0.60
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.75	+1.10
THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.75	+1.10
THE RASCH LOGIT RATIO SCALE IS THE ONLY BASIS FOR ASSESSING THERAPY IMPACT FOR LATENT TRAITS	1	0.05	-2.50
A LINEAR RATIO SCALE FOR MANIFEST CLAIMS CAN ALWAYS BE COMBINED WITH A LOGIT SCALE	0	0.80	+1.40
THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT	1	0.05	-2.50
THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.05	-2.50

The overall profile demonstrates a strong alignment with conventional HTA orthodoxy and a correspondingly weak engagement with the axioms of representational measurement. The consequence is that arithmetic operations are routinely accepted in circumstances where the conditions for measurement are absent. The Centre therefore appears embedded within the same international HTA memeplex observed across agencies, journals, and academic groups: a framework where numbers are manipulated as if they were measures, while the prior requirement to establish measurement is largely ignored.

The first cluster of items concerns elementary scale theory. The statement that interval measures lack a true zero receives only weak endorsement, with $p = 0.30$ and a normalized logit of -0.85. This is a revealing result because the absence of a true zero immediately limits the permissible operations that can be applied to interval scales. Proportional statements and multiplication are inadmissible without a ratio structure. Yet the broader HTA framework routinely behaves as though interval and ratio properties are interchangeable.

The requirement that measures be unidimensional also receives weak support, with $p = 0.25$ and logit -1.10. This is critical because HTA depends heavily on multiattribute descriptive systems such as the EQ-5D-3L and EQ-5D-5L. These systems combine multiple domains—mobility, self-care, pain, anxiety, and usual activities into a single utility score. The interrogation indicates that the knowledge base does not strongly recognize the incompatibility between multiattribute descriptive systems and the requirement for unidimensional measurement.

The same weakness appears in the statement that multiplication requires a ratio measure, with only $p = 0.20$ and logit -1.40. This is one of the defining failures of modern HTA. Cost-utility analysis

depends on multiplying utility values by time in the construction of the QALY. If the utility values are not ratio measures, then the multiplication is mathematically inadmissible. Yet the knowledge base remains aligned with the conventional assumption that such multiplication is acceptable.

This pattern continues with the items relating to utilities and QALYs. The statement "The QALY is a ratio measure" receives only $p = 0.40$ for the correct false response, with logit -0.40. Similarly, the statement "The QALY is a dimensionally homogeneous measure" receives the same weak rejection. These are decisive findings. They indicate that the knowledge base remains deeply committed to the conventional HTA assumption that utility-weighted life years represent meaningful quantities. They do not.

The QALY is impossible because it combines a legitimate ratio measure—time—with utility values derived from ordinal preference scoring of multiattribute health state descriptions. These utility values lack unidimensionality, invariance, and a meaningful zero. The multiplication therefore violates the requirements of dimensional homogeneity and ratio scaling. The resulting quantity does not exist as a measurable construct. It is not an approximation to a true quantity. It is a numerical artifact created by arithmetic applied to non-measures.

The Griffith interrogation also demonstrates that the knowledge base remains strongly aligned with accepted assumptions concerning health state valuation. The statement "EQ-5D-3L preference algorithms create interval measures" receives a relatively high endorsement for the false response, with $p = 0.65$ and logit +0.60, suggesting some awareness that preference algorithms do not establish interval measurement. However, this awareness is not carried through consistently into the treatment of QALYs or cost-effectiveness ratios. This inconsistency is characteristic of HTA generally: isolated recognition of measurement problems without acceptance of their implications for the analytical framework as a whole.

The statements concerning summation of subjective responses reinforce the same pattern. Both "Summations of subjective instrument responses are ratio measures" and "Summation of Likert question scores create a ratio measure" receive strong endorsement of the false response, each with $p = 0.75$ and logit +1.10. This indicates that the knowledge base recognizes, at least at some level, that simple summation of ordinal responses does not create ratio measurement.

Yet this recognition coexists with the widespread use of patient-reported outcomes and utility scoring systems that depend precisely on such summations and transformations. This contradiction lies at the heart of measurement inversion. The knowledge base partially recognizes the limitations of ordinal scoring while simultaneously embedding those same ordinal constructs within economic evaluations, simulation models, and policy recommendations.

The Rasch-related items reveal the most striking result in the entire profile. The statement that there are only two classes of measurement, linear ratio and Rasch logit ratio, receives the floor probability of $p = 0.05$ and logit -2.50. The same floor probability appears for the statements concerning transformation of subjective responses through Rasch rules, the Rasch logit ratio scale as the basis for assessing therapy impact for latent traits, the possession of latent traits, and the equivalence between Rasch rules and the axioms of representational measurement.

This is an extraordinarily important pattern because it indicates practical absence of Rasch measurement from the Griffith HTA knowledge base. Rasch is not merely underemphasized; it is essentially excluded from the conceptual structure supporting HTA teaching and research.

The implications are profound. Latent constructs such as quality of life, symptom burden, functioning, and need fulfillment cannot be directly observed. They require transformation into measures through an invariant conjoint framework. Without Rasch measurement, there is no defensible basis for treating ordinal responses as quantities. The interrogation indicates that this requirement is effectively absent from the Griffith knowledge base.

Instead, latent attributes continue to be treated descriptively rather than measurably. Scores are created, transformed, and inserted into models without establishing whether they represent measurable possession of an attribute. The consequence is that claims for therapy impact involving latent traits cannot be interpreted scientifically because the attribute itself has never been measured.

The statement concerning cost-effectiveness is equally revealing. “Claims for cost-effectiveness fail the axioms of representational measurement” receives only $p = 0.20$ and logit -1.40. This indicates that the knowledge base remains aligned with the conventional assumption that cost-effectiveness claims possess scientific legitimacy. Yet this legitimacy depends entirely on the existence of measurable quantities within the ratio. They do not exist.

The numerator in cost-effectiveness analysis is typically a modeled aggregation of projected costs, assumptions concerning utilization, transition probabilities, and discounted future events. The denominator is commonly a QALY construct built on non-measured utility values. Neither component satisfies the requirements for a meaningful ratio measure. The resulting cost-per-QALY ratio therefore cannot represent a measurable quantity of value. It is not “approximately correct.” There is nothing to approximate.

This point is reinforced by the treatment of reference case simulations. The statement “Reference case simulations generate falsifiable claims” receives strong endorsement for the false response, with $p = 0.75$ and logit +1.10. This is an important result because it indicates at least partial recognition that simulation outputs are not empirically falsifiable in the normal scientific sense.

However, the knowledge base continues to employ and teach simulation modelling despite this recognition. The contradiction is again revealing. If claims cannot be falsified, then they cannot contribute to the evolution of objective knowledge. They may support administrative or policy processes, but they do not satisfy the requirements of science.

The knowledge base performs somewhat better on the principles of falsification itself. The statement “Non-falsifiable claims should be rejected” receives $p = 0.65$ and logit +0.60. Similarly, time as a ratio measure receives strong endorsement, with $p = 0.85$ and logit +1.75. These results demonstrate that the Griffith knowledge base is not devoid of scientific concepts. Rather, it selectively applies them while exempting core HTA constructs from their implications.

This selective exemption is precisely what defines measurement inversion. The principles of measurement are recognized in the abstract but suspended when dealing with utilities, QALYs, simulation models, and cost-effectiveness claims. Arithmetic is allowed to proceed despite the absence of admissible measures.

The statement “Measurement precedes arithmetic” receives only p = 0.25 and logit -1.10. This may be the single most important result in the profile because it captures the defining characteristic of the HTA framework. Instead of establishing measurement and then determining permissible arithmetic operations, the framework begins with arithmetic. Utilities are multiplied, costs are aggregated, ratios are constructed, and simulations are run before asking whether the underlying quantities exist as measures.

The consequence is a closed analytical system in which numerical sophistication conceals conceptual failure. The more elaborate the simulation, the less visible the absence of measurement becomes. Yet no amount of modeling sophistication can rescue arithmetic applied to non-measures. The Griffith profile therefore mirrors the broader Australian and international HTA landscape. The Centre operates within a framework that is methodologically sophisticated but measurement-theoretically empty. The knowledge base recognizes certain isolated principles of scale theory and falsification, yet it continues to support an analytical architecture built upon non-measured constructs.

This leaves the Centre in a difficult position. As an applied health economics group closely connected to policy and decision-making, it remains tied to the accepted HTA orthodoxy represented by the PBAC and international reference case approaches. Yet the interrogation demonstrates that this orthodoxy lacks the conditions required for scientific legitimacy. The implication is unavoidable. The issue is not whether the existing framework can be refined or recalibrated. The issue is whether HTA built on utilities, QALYs, and reference case simulations can continue to claim scientific status. It cannot.

For Griffith, as for Australian HTA more broadly, the transition pathway is clear. Manifest attributes require linear ratio measures. Latent attributes require Rasch logit-based measures of possession. Claims must be protocol-driven, single-attribute, measurable, evaluable, and falsifiable. Until this transition occurs, the HTA framework associated with the Centre for Applied Health Economics will remain part of a global system characterized not by measurement, but by the normalization of arithmetic without measurement.

III. THE TRANSITION TO MEASUREMENT IN HEALTH TECHNOLOGY ASSESSMENT

THE IMPERATIVE OF CHANGE

This analysis has not been undertaken to criticize decisions made by health system, nor to assign responsibility for the analytical frameworks currently used in formulary review. The evidence shows something more fundamental: organizations have been operating within a system that does not permit meaningful evaluation of therapy impact, even when decisions are made carefully, transparently, and in good faith.

The present HTA framework forces health systems to rely on numerical outputs that appear rigorous but cannot be empirically assessed (Table 1). Reference-case models, cost-per-QALY ratios, and composite value claims are presented as decision-support tools, yet they do not satisfy the conditions required for measurement. As a result, committees are asked to deliberate over results that cannot be validated, reproduced, or falsified. This places decision makers in an untenable position: required to choose among therapies without a stable evidentiary foundation.

This is not a failure of expertise, diligence, or clinical judgment. It is a structural failure. The prevailing HTA architecture requires arithmetic before measurement, rather than measurement before arithmetic. Health systems inherit this structure rather than design it. Manufacturers respond to it. Consultants reproduce it. Journals reinforce it. Universities promote it. Over time it has come to appear normal, even inevitable.

Yet the analysis presented in Table 1 demonstrates that this HTA framework cannot support credible falsifiable claims. Where the dependent variable is not a measure, no amount of modeling sophistication can compensate. Uncertainty analysis cannot rescue non-measurement. Transparency cannot repair category error. Consensus cannot convert assumption into evidence.

The consequence is that formulary decisions are based on numerical storytelling rather than testable claims. This undermines confidence, constrains learning, and exposes health systems to growing scrutiny from clinicians, patients, and regulators who expect evidence to mean something more than structured speculation.

The imperative of change therefore does not arise from theory alone. It arises from governance responsibility. A health system cannot sustain long-term stewardship of care if it lacks the ability to distinguish between claims that can be evaluated and claims that cannot. Without that distinction, there is no pathway to improvement; only endless repetition for years to come.

This transition is not about rejecting evidence. It is about restoring evidence to its proper meaning. It requires moving away from composite, model-driven imaginary constructs toward claims that are measurable, unidimensional, and capable of empirical assessment over time. The remainder of this section sets out how that transition can occur in a practical, defensible, and staged manner.

MEANINGFUL THERAPY IMPACT CLAIMS

At the center of the current problem is not data availability, modeling skill, or analytic effort. It is the nature of the claims being advanced. Contemporary HTA has evolved toward increasingly complex frameworks that attempt to compress multiple attributes, clinical effects, patient experience, time, and preferences into single composite outputs. These constructs are then treated as if they were measures. They are not (Table 1).

The complexity of the reference-case framework obscures a simpler truth: meaningful evaluation requires meaningful claims. A claim must state clearly what attribute is being affected, in whom, over what period, and how that attribute is measured. When these conditions are met, evaluation becomes possible. When they are not complexity substitutes for clarity. The current framework is not merely incorrect; it is needlessly elaborate. Reference-case modeling requires dozens of inputs, assumptions, and transformations, yet produces outputs that cannot be empirically verified. Each additional layer of complexity increases opacity while decreasing accountability. Committees are left comparing models rather than assessing outcomes.

In contrast, therapy impact can be expressed through two, and only two, types of legitimate claims. First are claims based on manifest attributes: observable events, durations, or resource units. These include hospitalizations avoided, time to event, days in remission, or resource use. When properly defined and unidimensional, these attributes can be measured on linear ratio scales and evaluated directly.

Second are claims based on latent attributes: symptoms, functioning, need fulfillment, or patient experience. These cannot be observed directly and therefore cannot be scored or summed meaningfully. They require formal measurement through Rasch models to produce invariant logit ratio scales. These two forms of claims are sufficient. They are also far more transparent. Each can be supported by a protocol. Each can be revisited. Each can be reproduced. Most importantly, each can fail. But they cannot be combined. This is the critical distinction. A meaningful claim is one that can be wrong.

Composite constructs such as QALYs do not fail in this sense. They persist regardless of outcome because they are insulated by assumptions. They are recalculated, not refuted. That is why they cannot support learning. The evolution of objective knowledge regarding therapy impact in disease areas is an entirely foreign concept. By re-centering formulary review on single-attribute, measurable claims, health systems regain control of evaluation. Decisions become grounded in observable change rather than modeled narratives. Evidence becomes something that accumulates, rather than something that is re-generated anew for every submission.

THE PATH TO MEANINGFUL MEASUREMENT

Transitioning to meaningful measurement does not require abandoning current processes overnight. It requires reordering them. The essential change is not procedural but conceptual: measurement must become the gatekeeper for arithmetic, not its byproduct.

The first step is formal recognition that not all numerical outputs constitute evidence. Health systems must explicitly distinguish between descriptive analyses and evaluable claims. Numbers that do not meet measurement requirements may inform discussion but cannot anchor decisions.

The second step is restructuring submissions around explicit claims rather than models. Each submission should identify a limited number of therapy impact claims, each defined by attribute, population, timeframe, and comparator. Claims must be unidimensional by design.

Third, each claim must be classified as manifest or latent. This classification determines the admissible measurement standard and prevents inappropriate mixing of scale types.

Fourth, measurement validity must be assessed before any arithmetic is permitted. For manifest claims, this requires confirmation of ratio properties. For latent claims, this requires Rasch-based measurement with demonstrated invariance.

Fifth, claims must be supported by prospective or reproducible protocols. Evidence must be capable of reassessment, not locked within long-horizon simulations designed to frustrate falsification.

Sixth, committees must be supported through targeted training in representational measurement principles, including Rasch fundamentals. Without this capacity, enforcement cannot occur consistently.

Finally, evaluation must be iterative. Claims are not accepted permanently. They are monitored, reproduced, refined, or rejected as evidence accumulates.

These steps do not reduce analytical rigor. They restore it.

TRANSITION REQUIRES TRAINING

A transition to meaningful measurement cannot be achieved through policy alone. It requires a parallel investment in training, because representational measurement theory is not intuitive and has never been part of standard professional education in health technology assessment, pharmacoeconomics, or formulary decision making. For more than forty years, practitioners have been taught to work within frameworks that assume measurement rather than demonstrate it. Reversing that inheritance requires structured learning, not informal exposure.

At the center of this transition is the need to understand why measurement must precede arithmetic. Representational measurement theory establishes the criteria under which numbers can legitimately represent empirical attributes. These criteria are not optional. They determine whether addition, multiplication, aggregation, and comparison are meaningful or merely symbolic. Without this foundation, committees are left evaluating numerical outputs without any principled way to distinguish evidence from numerical storytelling.

Training must therefore begin with scale types and their permissible operations. Linear ratio measurement applies to manifest attributes that possess a true zero and invariant units, such as

time, counts, and resource use. Latent attributes, by contrast, cannot be observed directly and cannot be measured through summation or weighting. They require formal construction through a measurement model capable of producing invariant units. This distinction is the conceptual fulcrum of reform, because it determines which claims are admissible and which are not.

For latent trait claims, Rasch measurement provides the only established framework capable of meeting these requirements. Developed in the mid–twentieth century alongside the foundations of modern measurement theory, the Rasch model was explicitly designed to convert subjective observations into linear logit ratio measures. It enforces unidimensionality, tests item invariance, and produces measures that support meaningful comparison across persons, instruments, and time. These properties are not approximations; they are defining conditions of measurement.

Importantly, Rasch assessment is no longer technically burdensome. Dedicated software platforms developed and refined over more than four decades make Rasch analysis accessible, transparent, and auditable. These programs do not merely generate statistics; they explain why items function or fail, how scales behave, and whether a latent attribute has been successfully measured. Measurement becomes demonstrable rather than assumed.

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

Together, these programs equip health systems, committees, and analysts with the competence required to enforce measurement standards consistently. Training does not replace judgment; it enables it. Without such preparation, the transition to meaningful measurement cannot be sustained. With it, formulary decision making can finally rest on claims that are not merely numerical, but measurable

MAIMON RESEARCH LLC

DISTANCE EDUCATION PROGRAMS IN THE THEORY OF MEASUREMENT

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

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

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

DESIGNED FOR CLOSURE

For those who remain unconvinced that there is any need to abandon a long-standing and widely accepted HTA framework, it is necessary to confront a more fundamental question: why was this system developed and promoted globally in the first place?

The most plausible explanation is administrative rather than scientific. Policy makers were searching for an assessment framework that could be applied under conditions of limited empirical data while still producing a determinate conclusion. Reference-case modeling offered precisely this convenience. By constructing a simulation populated with assumptions, surrogate endpoints, preference weights, and extrapolated time horizons, it became possible to generate a numerical result that could be interpreted as decisive. Once an acceptable cost-effectiveness ratio emerged, the assessment could be declared complete and the pricing decision closed. This structure solved a political and administrative problem. It allowed authorities to claim that decisions were evidence-based without requiring the sustained empirical burden demanded by normal science. There was no requirement to formulate provisional claims and subject them to ongoing falsification. There was no obligation to revisit conclusions as new data emerged. Closure could be achieved at launch, rather than knowledge evolving over the product life cycle.

By contrast, a framework grounded in representational measurement would have imposed a very different obligation. Claims would necessarily be provisional. Measurement would precede arithmetic. Each therapy impact claim would require a defined attribute, a valid scale, a protocol, and the possibility of replication or refutation. Evidence would accumulate rather than conclude. Decisions would remain open to challenge as real-world data emerged. From an administrative standpoint, this was an unreasonable burden. It offered no finality.

The reference-case model avoided this problem entirely. By shifting attention away from whether quantities were measurable and toward whether assumptions were plausible, the framework replaced falsification with acceptability. Debate became internal to the model rather than external to reality. Sensitivity analysis substituted for empirical risk. Arithmetic proceeded without prior demonstration that the objects being manipulated possessed the properties required for arithmetic to be meaningful.

Crucially, this system required no understanding of representational measurement theory. Committees did not need to ask whether utilities were interval or ratio measures, whether latent traits had been measured or merely scored, or whether composite constructs could legitimately be multiplied or aggregated. These questions were never posed because the framework did not require

them to be posed. The absence of measurement standards was not an oversight; it was functionally essential.

Once institutionalized, the framework became self-reinforcing. Training programs taught modeling rather than measurement. Guidelines codified practice rather than axioms. Journals reviewed technique rather than admissibility. Over time, arithmetic without measurement became normalized as "good practice," while challenges grounded in measurement theory were dismissed as theoretical distractions. The result was a global HTA architecture capable of producing numbers, but incapable of producing falsifiable knowledge. Claims could be compared, ranked, and monetized, but not tested in the scientific sense. What evolved was not objective knowledge, but institutional consensus.

This history matters because it explains why the present transition is resisted. Moving to a real measurement framework with single, unidimensional claims does not merely refine existing methods; it dismantles the very mechanism by which closure has been achieved for forty years. It replaces decisiveness with accountability, finality with learning, and numerical plausibility with empirical discipline. Yet that is precisely the transition now required. A system that avoids measurement in order to secure closure cannot support scientific evaluation, cumulative knowledge, or long-term stewardship of healthcare resources. The choice is therefore unavoidable: continue with a framework designed to end debate, or adopt one designed to discover the truth.

Anything else is not assessment at all, but the ritualized manipulation of numbers detached from measurement, falsification, and scientific accountability.

ACKNOWLEDGEMENT

I acknowledge that I have used OpenAI technologies, including the large language model, to assist in the development of this work. All final decisions, interpretations, and responsibilities for the content rest solely with me.

REFERENCES

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

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

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

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