MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

SINGAPORE: THE AGENCY FOR CARE EFFECTIVENESS AND THE DENIAL OF REPRESENTATIONAL MEASUREMENT

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

LOGIT WORKING PAPER No 248 FEBRUARY 2026

www.maimonresearch.com

Tucson AZ

FOREWORD

HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

The Agency for Care Effectiveness (ACE) is Singapore's national health technology assessment body, established in 2015 under the Ministry of Health to support evidence-based decision making on the clinical and cost-effectiveness of health technologies. Its primary function is to evaluate pharmaceuticals, medical devices, diagnostic tests, and clinical interventions to inform national subsidy, reimbursement, and clinical guidance decisions. ACE produces technology guidance documents, economic evaluation reports, and methodological recommendations that influence whether therapies are included on the Ministry of Health's Standard Drug List (SDL) and MediShield Life coverage framework. These decisions directly affect patient access, pricing negotiations, and the allocation of public healthcare resources.

ACE operates as the central analytic authority supporting Singapore's publicly funded healthcare system. It reviews manufacturer submissions, conducts independent economic evaluations, and applies structured decision criteria incorporating clinical effectiveness, safety, economic impact, and health system relevance. Economic evaluations typically include cost-effectiveness analyses expressed in cost-per-QALY terms, derived from clinical evidence and modeled projections. ACE's recommendations are advisory but carry decisive influence, as they inform Ministry of Health subsidy decisions and national formulary inclusion. Through this role, ACE functions as the operational embodiment of Singapore's HTA knowledge base, translating methodological frameworks into concrete policy outcomes affecting therapy access, reimbursement, and health system resource allocation.

The objective of this study was to determine whether the Agency for Care Effectiveness (ACE), as Singapore's national authority for health technology assessment and subsidy recommendation, applies quantitative constructs that satisfy the axioms of representational measurement. ACE occupies a pivotal operational role in determining whether pharmaceutical products and other health technologies are subsidized, reimbursed, or recommended within the public healthcare system. Its methodological guidance and economic evaluation requirements define the quantitative framework through which therapy impact is assessed. The study therefore applied the canonical 24-item representational measurement diagnostic to ACE's operational knowledge base, including its economic evaluation guidance, technology assessment reports, methodological frameworks, and submission requirements. The objective was not to evaluate procedural rigor or administrative transparency, but to determine whether the evaluative constructs mandated by ACE satisfy the fundamental conditions necessary for lawful arithmetic, including unidimensionality, dimensional homogeneity, admissible scale transformations, and ratio scale requirements. The analysis addresses a decisive question: whether Singapore's operational HTA authority requires measurement-valid constructs when making decisions that affect therapy access, pricing, and subsidy allocation.

The logit profile demonstrates systematic non-possession of the axioms of representational measurement within the ACE knowledge base. Foundational propositions—including that measurement must precede arithmetic, that multiplication requires ratio measurement, and that Rasch transformation is required to convert ordinal subjective responses into invariant interval measures—collapse to floor logit values, indicating their absence as operational constraints. Conversely, false propositions asserting that QALYs are ratio measures, that composite utility scores can be aggregated and manipulated arithmetically, and that simulation-based cost-effectiveness ratios represent empirically valid quantitative claims register strong positive logits. These findings establish that ACE's evaluative architecture substitutes preference-based scoring systems and modeled composite indices for measurement-valid quantities. Arithmetic operations are performed on constructs whose measurement properties are neither demonstrated nor required to be demonstrated. The resulting evaluative framework functions as an internally coherent administrative system, but its quantitative outputs lack the empirical grounding required for falsification, replication, and cumulative scientific knowledge.

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 AGENCY FOR CARE EFFECTIVENESS KNOWLEDGE BASE

The Agency for Care Effectiveness operates as Singapore's central authority for health technology assessment. Its knowledge base is defined by methodological guidance documents, economic evaluation frameworks, technology assessment reports, and formal submission requirements that specify how therapeutic value is to be quantified and compared. Manufacturers seeking subsidy listing or reimbursement support must provide evidence of clinical effectiveness and economic value, typically including incremental cost-effectiveness analyses expressed in cost-per-QALY terms. These evaluations integrate clinical trial evidence, observational data, epidemiological projections, and preference-based utility scores derived from multiattribute instruments such as EQ-5D.

The evaluative architecture employed by ACE reflects the global diffusion of HTA methodology, particularly the reference case model pioneered in the United Kingdom and subsequently adopted across multiple jurisdictions. Within this framework, therapeutic impact is quantified using composite indices that combine survival duration with preference-weighted assessments of health state quality. Utility scores are derived from preference elicitation exercises conducted in general populations, using valuation techniques such as time trade-off and standard gamble. These scores are applied within simulation models that project long-term clinical outcomes and associated costs. The resulting outputs—incremental cost-effectiveness ratios expressed as cost per QALY gained—function as integrative decision metrics used to compare therapies and inform subsidy recommendations.

ACE's knowledge base demonstrates procedural rigor and administrative coherence. Assessment reports follow standardized formats, explicitly document assumptions, and incorporate sensitivity analyses designed to explore parameter uncertainty. Decision frameworks incorporate multiple considerations, including clinical effectiveness, safety, disease severity, budget impact, and cost-effectiveness. The methodological structure is designed to support consistent and transparent evaluation across therapeutic areas.

However, the quantitative constructs embedded within this framework originate from preference-based scoring systems rather than measurement-validated quantities. Utility scores derived from multiattribute instruments represent ordered preference rankings over multidimensional health state descriptions. Their numerical values reflect scoring algorithms rather than invariant measurement units. Despite this, the evaluative framework permits multiplication of utility scores by time, aggregation across individuals, and comparison across therapies as if these constructs possessed ratio scale properties. Simulation modeling extends these scoring systems into projected lifetime outcomes, producing numerical estimates of incremental QALYs and associated costs.

The knowledge base therefore reflects a coherent administrative system built upon inherited methodological conventions. It emphasizes consistency, transparency, and structured evaluation. Yet it does not require demonstration that the constructs used in arithmetic operations satisfy representational measurement axioms. Rasch transformation, which provides the only lawful method for converting ordinal observations into invariant interval measures for latent traits, is absent from the evaluative architecture. Composite utility indices function as operational decision variables despite lacking invariant unit structure.

As a result, the ACE knowledge base embodies a stable evaluative framework that produces quantitative outputs used to inform national healthcare decisions. These outputs possess procedural legitimacy and administrative authority. However, their numerical properties originate from scoring systems and modeling assumptions rather than measurement-validated quantities. The logit profile demonstrates that the axioms necessary to transform observations into measurement do not operate as binding constraints within this knowledge base.

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: SINGAPORE AGENCY FOR CARE EFFECTIVENESS

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.

AGENCY FOR CARE EFFECTIVENESS: REGULATORY SOPHISTICATION WITHOUT MEASUREMENT FOUNDATION

The canonical logit profile of the Agency for Care Effectiveness reveals the operational embodiment of measurement failure within Singapore's health technology assessment framework (Table 1). The profile does not demonstrate partial misunderstanding, ambiguity, or transitional uncertainty. It demonstrates structural exclusion. The axioms of representational measurement, the necessary conditions for the existence of measurable quantities, do not function as binding constraints within the agency's evaluative architecture. Instead, arithmetic operations are performed on constructs whose scale properties are neither established nor required to be established. The resulting framework possesses administrative coherence but lacks measurement validity. It produces numbers, but those numbers do not represent quantities.

TABLE 1: ITEM STATEMENT, RESPONSE, ENDORSEMENT AND NORMALIZED LOGITS AGENCY FOR CARE EFFECTIVENESS

STATEMENT	RESPONSE 1=TRUE 0=FALSE	ENDORSEMENT OF RESPONSE CATEGORICAL PROBABILITY	NORMALIZED LOGIT (IN RANGE +/- 2.50)
INTERVAL MEASURES LACK A TRUE ZERO	1	0.20	-1.40
MEASURES MUST BE UNIDIMENSIONAL	1	0.15	-1.75
MULTIPLICATION REQUIRES A RATIO MEASURE	1	0.10	-2.20
TIME TRADE-OFF PREFERENCES ARE UNIDIMENSIONAL	0	0.85	+1.75
RATIO MEASURES CAN HAVE NEGATIVE VALUES	0	0.90	2.20
EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES	0	0.90	+2.20
THE QALY IS A RATIO MEASURE	0	0.95	+2.50
TIME IS A RATIO MEASURE	1	0.95	+2.50
MEASUREMENT PRECEDES ARITHMETIC	1	0.10	-2.20
SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES	0	0.90	+2.20
MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC	1	0.05	-2.50
THERE ARE ONLY TWO CLASSES OF MEASUREMENT LINEAR RATIO AND RASCH LOGIT RATIO	1	0.05	-2.50
TRANSFORMING SUBJECTIVE RESPONSES TO INTERVAL MEASUREMENT IS ONLY POSSIBLE WITH RASH RULES	1	0.05	-2.50
SUMMATION OF LIKERT QUESTION SCORES CREATES A RATIO MEASURE	0	0.95	+2.50
THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE	0	0.90	+2.20
CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.15	-1.75
QALYS CAN BE AGGREGATED	0	0.95	+2.50

NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.70	+0.85
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.90	+2.20
THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.60	+0.40
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	0.60	+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	0.05	-2.50

The most decisive feature of the logit profile is the repeated collapse of foundational measurement axioms to the absolute floor of −2.50. Statements asserting that meeting the axioms of representational measurement is required for arithmetic, that Rasch transformation is necessary to convert ordinal observations into interval measures, that only linear ratio and Rasch logit ratio scales constitute lawful measurement structures, and that Rasch logit scaling provides the only valid basis for latent trait quantification all register at −2.50. This value denotes effective non-possession. These propositions do not operate as methodological constraints within the ACE knowledge base. They are not contested, refined, or partially incorporated. They are absent. Their absence defines the structure of the evaluative system.

This absence is logically decisive. Measurement precedes arithmetic. Arithmetic operations such as multiplication, division, and ratio comparison are admissible only when applied to quantities measured on ratio scales possessing invariant units and true zero points. Without such measurement, arithmetic operations lack empirical meaning. They become symbolic manipulations of numbers detached from quantitative referents. The logit value of −2.20 associated with the statement that measurement precedes arithmetic confirms that this logical ordering does not function as an operational requirement within ACE's evaluative framework. Arithmetic is performed first. Measurement is assumed or ignored.

The positive logit cluster associated with statements endorsing QALYs as ratio measures, supporting aggregation of utilities, and accepting summated ordinal preference scores as arithmetic objects reveals the mechanism through which this inversion operates. Composite utility indices derived from multiattribute instruments such as EQ-5D function as central evaluative constructs.

These indices originate from preference-based scoring algorithms applied to multidimensional descriptive systems. Their numerical values represent ordered preference scores, not invariant measures of unidimensional attributes. They lack demonstrated unit invariance, dimensional homogeneity, and ratio scale structure. Yet the positive logits of +2.50 and +2.20 associated with statements endorsing their aggregation and arithmetic use demonstrate institutional acceptance of these constructs as quantitative entities.

This substitution of scoring for measurement represents the defining structural characteristic of the ACE knowledge base. Scoring systems generate numbers, but measurement systems generate quantities. The distinction is fundamental. Their numerical representation reflects empirical relations. Scored values exist only within the scoring system itself. Their numerical properties are artifacts of algorithmic construction; the basis of the ACE evaluative framework.

The asymmetry between manifest and latent constructs further illustrates the structural nature of the problem. Time is correctly recognized as a ratio measure, registering a logit of +2.50. This demonstrates that the agency possesses and applies representational measurement principles when evaluating manifest attributes. Time possesses invariant units and a true zero. It supports lawful arithmetic operations. However, this measurement discipline is not extended to latent constructs such as subjective health attributes. Instead, latent constructs are assigned numerical scores derived from preference aggregation and treated as if they possessed equivalent measurement properties. This asymmetry reveals that the exclusion of measurement axioms is selective rather than universal. Measurement principles are applied where measurement is straightforward and ignored where measurement would require transformation through Rasch modeling.

The exclusion of Rasch measurement is particularly consequential. Rasch transformation provides the only lawful method for converting ordinal observations into interval-level measures while preserving invariance across persons and items. Its repeated collapse to −2.50 across multiple canonical statements confirms that latent constructs within the ACE evaluative framework are not measured. They are scored. Without invariant unit structure, arithmetic comparisons across therapies lack empirical meaning. Differences in composite utility scores reflect differences in scoring algorithms rather than differences in measured quantities.

The positive logit values associated with statements endorsing reference case simulation further reinforce this conclusion. Simulation models generate projected cost-effectiveness ratios by integrating clinical assumptions, epidemiological projections, and composite utility scores. These outputs are computationally precise but empirically unfalsifiable. They do not represent observed quantities. They represent numerical consequences of model assumptions. Yet the logit profile demonstrates institutional endorsement of these outputs as evaluative objects. Simulation coherence substitutes for empirical measurement.

This substitution has profound implications for falsifiability. Scientific claims must be structured so that empirical observations can confirm or refute them. Measurement provides the quantitative structure necessary for such testing. Without measurement, falsification becomes impossible. Simulation outputs cannot be falsified because they are functions of assumptions rather than observations. They can be recalculated but not empirically refuted. The positive logit associated with endorsement of simulation outputs demonstrates that falsifiability is not enforced as an operational requirement within the ACE evaluative framework.

The persistence of this evaluative architecture reflects institutional stabilization rather than empirical validation. Once composite utility constructs and cost-per-QALY ratios become embedded within decision frameworks, they acquire administrative legitimacy. They are reproduced through methodological guidance, submission requirements, training programs, and publication norms. Their numerical outputs create the appearance of scientific rigor. This appearance reinforces their continued use. The logit profile demonstrates that this stabilization occurs despite the absence of measurement foundations.

This stabilization transforms the evaluative system into a closed epistemic framework. Measurement-valid alternatives—linear ratio measures for manifest attributes and Rasch logit ratio measures for latent traits—are excluded at the operational level. Their exclusion is reflected in repeated floor logits across Rasch-related statements. Without invariant measurement structure, quantitative claims cannot support falsification, replication, or cumulative knowledge development. Numerical outputs become administrative artifacts rather than empirical discoveries.

The implications extend beyond epistemology to institutional responsibility. The Agency for Care Effectiveness influences national decisions regarding therapy availability, reimbursement eligibility, and resource allocation. These decisions affect patient access to treatment and health system resource distribution. When such decisions rely on constructs lacking measurement validity, quantitative outputs possess administrative authority without empirical grounding. This condition raises fundamental questions regarding scientific accountability and duty of care.

Duty of care requires that decisions affecting patient outcomes be grounded in empirically valid evidence. Measurement provides the quantitative structure necessary for such evidence. Without measurement, numerical claims cannot support reliable inference. The ACE logit profile demonstrates that measurement axioms do not function as operational constraints within the agency's evaluative framework. This absence does not reflect procedural oversight. It reflects structural architecture. Arithmetic operations are performed on constructs whose measurement properties are not established.

The global context reinforces the significance of this finding. Similar logit profiles appear across HTA agencies, national frameworks, and academic journals worldwide. Singapore's reputation for technocratic rigor and methodological discipline makes its profile particularly instructive. The findings demonstrate that procedural sophistication does not guarantee measurement validity. Institutional coherence can coexist with measurement absence.

The distinction between administrative coherence and scientific measurement defines the central conclusion of the logit assessment. The ACE evaluative framework operates as an internally consistent administrative system built upon quantitative constructs that lack measurement validity. Its numerical outputs possess procedural legitimacy but not empirical meaning. Arithmetic operations proceed without measurement foundation. Evaluation becomes computation without quantification.

Recovery requires structural reconstruction. Arithmetic operations must be restricted to constructs possessing invariant ratio scale properties. Manifest attributes must be measured on linear ratio scales. Latent constructs must be measured through Rasch transformation to establish invariant logit ratio scales. Only within such a framework can quantitative claims regain empirical meaning. Only within such a framework can health technology assessment re-enter the domain of measurement-based science. The logit profile demonstrates that this reconstruction has not yet occurred within the Agency for Care Effectiveness. The measurement axioms required for scientific quantification remain absent as operational constraints. Composite utility scores and cost-per-QALY ratios continue to function as administrative decision variables. Numerical sophistication masks measurement absence.

The logit evidence therefore identifies not a technical limitation but a structural condition. Measurement principles do not operate as binding constraints within Singapore's HTA evaluative framework. Arithmetic precedes measurement. Scoring substitutes for measurement. Simulation substitutes for observation. The resulting system produces numerical outputs that guide policy decisions but do not constitute quantitative measurement in the scientific sense. This condition defines the present epistemic status of health technology assessment within Singapore. Numerical outputs possess administrative authority but lack empirical grounding. Evaluation proceeds without measurement. And until representational measurement axioms become operational requirements rather than excluded abstractions, this condition will persist.

CARE EFFECTIVENESS IS NOT SIMULATED IMAGINARY COST-EFFECTIVENESS

If, as its name implies, the operational focus of the Agency for Care Effectiveness is the assessment of the effectiveness of therapies, then the analytical framework selected by ACE is fundamentally misaligned with its stated purpose. Effectiveness is not an abstract administrative construct. It is a property of therapies manifested in measurable changes in patient outcomes. The assessment of effectiveness therefore requires measurement, not scoring, and certainly not simulation of composite preference indices. Measurement demands a ratio scale. For manifest attributes, such as survival duration, hospitalizations avoided, symptom-free days, or medication possession, this requires a linear ratio scale with invariant units and a true zero. For latent traits, such as functional capacity or symptom burden, it requires transformation through Rasch measurement to produce invariant logit ratio scales. Only under these conditions can arithmetic operations legitimately support comparative claims, empirical falsification, and cumulative scientific knowledge.

This requirement is not optional. It is the foundation of normal science. Measurement allows claims to be tested against observation. It allows differences between therapies to be interpreted quantitatively. It allows replication across populations and time. Without measurement, numerical outputs cannot be falsified. They cannot be confirmed. They cannot contribute to the evolution of objective knowledge. They remain administrative artifacts; numerical expressions of assumption rather than empirical properties of therapies. A collection of numerical stories.

The analytical framework employed by ACE departs from these requirements at its foundation. Instead of measuring therapy impact using ratio-scale attributes, ACE relies on composite constructs such as QALYs derived from multiattribute preference instruments and simulation models. These constructs are not measures of effectiveness. They are numerical scores generated by applying preference weights to descriptive health state classifications. Their numerical properties are determined by scoring conventions, not by invariant measurement units. When these scores are multiplied by time and incorporated into cost-effectiveness ratios, the resulting quantities possess the appearance of precision without possessing the structural properties required for measurement.

Simulation models extend this problem rather than resolving it. They generate projections of cost-per-QALY outcomes over extended time horizons based on assumed transitions between modeled health states. These outputs are computationally coherent but empirically unfalsifiable. They cannot be directly observed. They cannot be empirically refuted. They can only be recalculated under alternative assumptions. Such outputs do not measure effectiveness. They simulate administrative constructs built upon scoring systems that themselves lack measurement validity.

This analytical architecture is incompatible with the assessment of care effectiveness. Effectiveness must be demonstrated through observable changes in measurable patient outcomes. These outcomes must be expressed in invariant units that support lawful arithmetic comparison. A therapy either increases survival duration, reduces exacerbations, improves functional ability, or achieves other measurable effects. These effects can be quantified on ratio scales and subjected to empirical testing. They can be replicated. They can be falsified. They contribute to objective knowledge.

By contrast, composite utility constructs and simulated cost-effectiveness ratios operate outside this scientific framework. They replace measurement with scoring and observation with projection. They generate numerical outputs that cannot be empirically tested against observable quantities. Their apparent precision masks the absence of measurement foundation.

This substitution represents the defining characteristic of the HTA memeplex to which ACE subscribes. The memeplex prioritizes administrative coherence, integrative numerical outputs, and procedural consistency over measurement validity. It produces numbers that support decision processes but do not constitute measurement of therapeutic effectiveness. In doing so, it abandons the standards that distinguish science from administrative accounting.

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 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