MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

UNITED STATES:TUFTS CEA REGISTRY AND THE ACADEMIC MEMORY PALACE OF FALSE MEASUREMENT

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

LOGIT WORKING PAPER No 12 JANUARY 2026

www.maimonresearch.com

Tucson AZ

FOREWORD

HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

This Logit Working Paper series documents a finding as extraordinary as it is uncomfortable: health technology assessment (HTA), across nations, agencies, journals, and decades, has developed as a global system of non-measurement. It speaks the language of numbers, models, utilities, QALYs, "value for money," thresholds, discounting, incremental ratios, extrapolations, and simulations. It demands arithmetic at every turn, multiplication, division, summation, aggregation, discounting, yet it never once established that the quantities to which these operations are applied are measurable. HTA has built a vast evaluative machinery on foundations that do not exist. The probabilities and normalized logits in the country reports that follow provide the empirical confirmation of this claim. They show, with unsettling consistency, that the global HTA.

The objective of this study is to interrogate the epistemic foundations of the Tufts Cost-Effectiveness Analysis (CEA) Registry as a defining component of the contemporary health technology assessment knowledge infrastructure. Rather than treating the registry as a neutral bibliographic resource, this analysis examines the belief system embedded in what the registry indexes, standardizes, and presents as comparable quantitative evidence. Using a 24-item diagnostic grounded in representational measurement theory, the study evaluates whether the core numerical objects preserved and amplified by the registry, utilities, QALYs, ICERs, and reference-case simulation outputs satisfy the axioms required for admissible arithmetic, falsification, and the evolution of objective knowledge. The aim is not to assess individual studies catalogued by the registry, but to determine whether the registry itself functions as a measurement-literate archive or as an institutional mechanism for stabilizing and reproducing false measurement at scale.

This assessment is important given the status of the Tufts CEA Registry. For more than four decades, it has functioned as the central archival and classificatory infrastructure for cost-utility analysis worldwide. The registry systematically identifies, curates, and abstracts published cost-effectiveness studies, standardizing their key features, utility instruments, QALY constructions, ICERs, time horizons, discount rates, and threshold comparisons and presenting them as members of a single, comparable quantitative field. It is routinely cited by academic researchers, HTA agencies, journals, guideline developers, and policy analysts as the authoritative source for tracking trends in cost-effectiveness results, methodological practice, and "value for money" claims across diseases and interventions. In doing so, the registry does not merely document the literature; it actively confers legitimacy on the numerical constructs it indexes by treating them as stable, commensurable, and cumulatively informative. Its role is therefore constitutive rather than passive: by deciding what counts as a cost-effectiveness analysis and how its results are summarized and compared, the Tufts CEA Registry helps define the boundaries of acceptable quantitative reasoning in health technology assessment.

It is precisely this archival authority that renders the Tufts CEA Registry epistemically consequential and potentially dangerous. By organizing, normalizing, and comparing numerical outputs that fail the axioms of representational measurement, the registry transforms

methodological error into cumulative knowledge. Utilities that lack interval properties, QALYs that lack dimensional homogeneity and a true zero, and ICERs that violate the conditions for ratio arithmetic are not flagged as inadmissible; they are catalogued, summarized, and trended as if they were legitimate measures. Over time, repetition substitutes for validation. What cannot be falsified at the level of individual studies acquires apparent credibility through aggregation at the registry level. The registry thus functions as a stabilizing mechanism for false measurement, insulating core constructs from scrutiny by embedding them in a curated corpus that appears orderly, comparative, and mature. In this sense, the Tufts CEA Registry does not merely reflect the HTA memeplex; it operationalizes it, converting non-measurement into institutional memory and ensuring its transmission to future analysts as settled science rather than as an unresolved epistemic failure.

The findings are unambiguous and severe. The Tufts CEA Registry knowledge base exhibits a systematic inversion of scientific order in which arithmetic is treated as primary and measurement as optional or irrelevant. Foundational axioms—measurement preceding arithmetic, the necessity of ratio scales for multiplication, unidimensionality, and the inadmissibility of composite constructs such as QALYs—are weakly endorsed or rejected outright, clustering near the floor of the canonical ±2.50 logit scale. In contrast, mathematically impossible propositions required to sustain cost-utility analysis are endorsed at or near the ceiling. The QALY is treated as a ratio measure, QALYs are treated as aggregable, summated ordinal responses are treated as ratio quantities, and reference-case simulations are treated as falsifiable evidence. Rasch measurement, the only framework capable of legitimizing latent-trait claims derived from patient-reported outcomes, is categorically excluded. The resulting profile is not one of marginal error or disciplinary confusion, but of structural commitment to arithmetic without measurement. The registry does not merely reflect this commitment; it amplifies it, converting a set of inadmissible numerical practices into an apparently cumulative scientific canon.

The starting point is simple and inescapable: *measurement precedes arithmetic*. This principle is not a methodological preference but a logical necessity. One cannot multiply what one has not measured, cannot sum what has no dimensional homogeneity, cannot compare ratios when no ratio scale exists. When HTA multiplies time by utilities to generate QALYs, it is performing arithmetic with numbers that cannot support the operation. When HTA divides cost by QALYs, it is constructing a ratio from quantities that have no ratio properties. When HTA aggregates QALYs across individuals or conditions, it is combining values that do not share a common scale. These practices are not merely suboptimal; they are mathematically impossible.

The modern articulation of this principle can be traced to Stevens' seminal 1946 paper, which introduced the typology of nominal, ordinal, interval, and ratio scales [i]. Stevens made explicit what physicists, engineers, and psychologists already understood: different kinds of numbers permit different kinds of arithmetic. Ordinal scales allow ranking but not addition; interval scales permit addition and subtraction but not multiplication; ratio scales alone support multiplication, division, and the construction of meaningful ratios. Utilities derived from multiattribute preference exercises, such as EQ-5D or HUI, are ordinal preference scores; they do not satisfy the axioms of interval measurement, much less ratio measurement. Yet HTA has, for forty years, treated these utilities as if they were ratio quantities, multiplying them by time to create QALYs and inserting them into models without the slightest recognition that scale properties matter. Stevens' paper

should have blocked the development of QALYs and cost-utility analysis entirely. Instead, it was ignored.

The foundational theory that establishes *when* and *whether* a set of numbers can be interpreted as measurements came with the publication of Krantz, Luce, Suppes, and Tversky's *Foundations of Measurement* (1971) [ii] . Representational Measurement Theory (RMT) formalized the axioms under which empirical attributes can be mapped to numbers in a way that preserves structure. Measurement, in this framework, is not an act of assigning numbers for convenience, it is the discovery of a lawful relationship between empirical relations and numerical relations. The axioms of additive conjoint measurement, homogeneity, order, and invariance specify exactly when interval scales exist. RMT demonstrated once and for all that measurement is not optional and not a matter of taste: either the axioms hold and measurement is possible, or the axioms fail and measurement is impossible. Every major construct in HTA, utilities, QALYs, DALYs, ICERs, incremental ratios, preference weights, health-state indices, fails these axioms. They lack unidimensionality; they violate independence; they depend on aggregation of heterogeneous attributes; they collapse under the requirements of additive conjoint measurement. Yet HTA proceeded, decade after decade, without any engagement with these axioms, as if the field had collectively decided that measurement theory applied everywhere except in the evaluation of therapies.

Whereas representational measurement theory articulates the axioms for interval measurement, Georg Rasch's 1960 model provides the only scientific method for transforming ordered categorical responses into interval measures for latent traits [iii]. Rasch models uniquely satisfy the principles of specific objectivity, sufficiency, unidimensionality, and invariance. For any construct such as pain, fatigue, depression, mobility, or need, Rasch analysis is the only legitimate means of producing an interval scale from ordinal item responses. Rasch measurement is not an alternative to RMT; it is its operational instantiation. The equivalence of Rasch's axioms and the axioms of representational measurement was demonstrated by Wright, Andrich and others as early as the 1970s. In the latent-trait domain, the very domain where HTA claims to operate; Rasch is the only game in town [iv].

Yet Rasch is effectively absent from all HTA guidelines, including NICE, PBAC, CADTH, ICER, SMC, and PHARMAC. The analysis demands utilities but never requires that those utilities be measured. They rely on multiattribute ordinal classifications but never understand that those constructs be calibrated on interval or ratio scales. They mandate cost-utility analysis but never justify the arithmetic. They demand modelled QALYs but never interrogate their dimensional properties. These guidelines do not misunderstand Rasch; they do not know it exists. The axioms that define measurement and the model that makes latent trait measurement possible are invisible to the authors of global HTA rules. The field has evolved without the science that measurement demands.

How did HTA miss the bus so thoroughly? The answer lies in its historical origins. In the late 1970s and early 1980s, HTA emerged not from measurement science but from welfare economics, decision theory, and administrative pressure to control drug budgets. Its core concern was *valuing health states*, not *measuring health*. This move, quiet, subtle, but devastating, shifted the field away from the scientific question "What is the empirical structure of the construct we intend to

measure?" and toward the administrative question "How do we elicit a preference weight that we can multiply by time?" The preference-elicitation projects of that era (SG, TTO, VAS) were rationalized as measurement techniques, but they never satisfied measurement axioms. Ordinal preferences were dressed up as quasi-cardinal indices; valuation tasks were misinterpreted as psychometrics; analyst convenience replaced measurement theory. The HTA community built an entire belief system around the illusion that valuing health is equivalent to measuring health. It is not.

The endurance of this belief system, forty years strong and globally uniform, is not evidence of validity but evidence of institutionalized error. HTA has operated under conditions of what can only be described as *structural epistemic closure*: a system that has never questioned its constructs because it never learned the language required to ask the questions. Representational measurement theory is not taught in graduate HTA programs; Rasch modelling is not part of guideline development; dimensional analysis is not part of methodological review. The field has been insulated from correction because its conceptual foundations were never laid. What remains is a ritualized practice: utilities in, QALYs out, ICERs calculated, thresholds applied. The arithmetic continues because everyone assumes someone else validated the numbers.

This Logit Working Paper series exposes, through probabilistic and logit-based interrogations of AI large language national knowledge bases, the scale of this failure. The results display a global pattern: true statements reflecting the axioms of measurement receive weak endorsement; false statements reflecting the HTA belief system receive moderate or strong reinforcement. This is not disagreement. It is non-possession. It shows that HTA, worldwide, has developed as a quantitative discipline without quantitative foundations; a confused exercise in numerical storytelling.

The conclusion is unavoidable: HTA does not need incremental reform; it needs a scientific revolution. Measurement must precede arithmetic. Representational axioms must precede valuation rituals. Rasch measurement must replace ordinal summation and utility algorithms. Value claims must be falsifiable, protocol-driven, and measurable; rather than simulated, aggregated, and numerically embellished.

The global system of non-measurement is now visible. The task ahead is to replace it with science.

Paul C Langley, Ph.D

Email: langleylapaloma@gmail.com

DISCLAIMER

This analysis is generated through the structured interrogation of a large language model (LLM) applied to a defined documentary corpus and is intended solely to characterize patterns within an aggregated knowledge environment. It does identify, assess, or attribute beliefs, intentions, competencies, or actions to any named individual, faculty member, student, administrator, institution, or organization. The results do not constitute factual findings about specific persons or programs, nor should they be interpreted as claims regarding professional conduct, educational quality, or compliance with regulatory or accreditation standards. All probabilities and logit values reflect model-based inferences about the presence or absence of concepts within a bounded textual ecosystem, not judgments about real-world actors. The analysis is exploratory, interpretive, and methodological in nature, offered for scholarly discussion of epistemic structures rather than evaluative or legal purposes. Any resemblance to particular institutions or practices is contextual and non-attributive, and no adverse implication should be inferred.

1. INTERROGATING THE LARGE LANGUAGE MODEL

A large language model **(**LLM**)** is an artificial intelligence system designed to understand, generate, and manipulate human language by learning patterns from vast amounts of text data. Built on deep neural network architectures, most commonly transformers, LLMs analyze relationships between words, sentences, and concepts to produce contextually relevant responses. During training, the model processes billions of examples, enabling it to learn grammar, facts, reasoning patterns, and even subtle linguistic nuances. Once trained, an LLM can perform a wide range of tasks: answering questions, summarizing documents, generating creative writing, translating languages, assisting with coding, and more. Although LLMs do not possess consciousness or true understanding, they simulate comprehension by predicting the most likely continuation of text based on learned patterns. Their capabilities make them powerful tools for communication, research, automation, and decision support, but they also require careful oversight to ensure accuracy, fairness, privacy, and responsible use

In this Logit Working Paper, "interrogation" refers not to discovering what an LLM *believes,* it has no beliefs, but to probing the content of the *corpus-defined knowledge space* we choose to analyze. This knowledge base is enhanced if it is backed by accumulated memory from the user. In this case the interrogation relies also on 12 months of HTA memory from continued application of the system to evaluate HTA experience. The corpus is defined before interrogation: it may consist of a journal (e.g., *Value in Health*), a national HTA body, a specific methodological framework, or a collection of policy documents. Once the boundaries of that corpus are established, the LLM is used to estimate the conceptual footprint within it. This approach allows us to determine which principles are articulated, neglected, misunderstood, or systematically reinforced.

In this HTA assessment, the objective is precise: to determine the extent to which a given HTA knowledge base or corpus, global, national, institutional, or journal-specific, recognizes and reinforces the foundational principles of representational measurement theory (RMT). The core principle under investigation is that measurement precedes arithmetic; no construct may be treated as a number or subjected to mathematical operations unless the axioms of measurement are satisfied. These axioms include unidimensionality, scale-type distinctions, invariance, additivity, and the requirement that ordinal responses cannot lawfully be transformed into interval or ratio quantities except under Rasch measurement rules.

The HTA knowledge space is defined pragmatically and operationally. For each jurisdiction, organization, or journal, the corpus consists of:

- published HTA guidelines
- agency decision frameworks
- cost-effectiveness reference cases
- academic journals and textbooks associated with HTA
- modelling templates, technical reports, and task-force recommendations
- teaching materials, methodological articles, and institutional white papers

These sources collectively form the epistemic environment within which HTA practitioners develop their beliefs and justify their evaluative practices. The boundary of interrogation is thus

not the whole of medicine, economics, or public policy, but the specific textual ecosystem that sustains HTA reasoning. . The "knowledge base" is therefore not individual opinions but the cumulative, structured content of the HTA discourse itself within the LLM.

THE TUFTS CEA REGISTRY KNOWLEDGE BASE

For the purposes of this analysis, the Tufts CEA Registry knowledge base is defined as the structured body of quantitative conventions, methodological assumptions, and evaluative norms embodied in the registry's design, inclusion criteria, indexing practices, and analytic summaries. It is not defined by any single publication, modeling choice, or authorial intent, but by the collective pattern that emerges from how cost-utility analyses are selected, categorized, and rendered commensurable across disease areas, interventions, and populations.

The registry's defining feature is its exclusive focus on cost-utility analysis expressed in cost per QALY terms. This focus is not merely descriptive; it is normative. By cataloguing ICERs, utility instruments, discount rates, time horizons, and threshold comparisons as if they were elements of a shared quantitative language, the registry implicitly asserts that QALYs are legitimate measures, that ICERs are admissible ratios, and that results drawn from heterogeneous modeling exercises are meaningfully comparable. These assertions are not defended through measurement theory; they are presupposed through repeated use.

The knowledge base inferred from the registry is therefore one in which measurement is treated behaviorally rather than axiomatically. Utility scores derived from ordinal preference instruments are accepted as quantitative inputs without demonstration of interval or ratio properties. Summation and averaging of subjective responses are treated as sufficient to create quantities suitable for multiplication by time. Negative "utilities" are accommodated without abandoning claims of ratio status. Composite constructs combining survival time and preference weights are treated as dimensionally homogeneous objects. None of these commitments are argued explicitly; they are embedded in what the registry treats as routine, indexable practice.

Equally revealing are the systematic absences. Representational measurement theory is not engaged as a gating framework. Scale-type constraints are not treated as prior conditions for admissible arithmetic. Rasch measurement is not required, encouraged, or even meaningfully acknowledged, despite the registry's heavy reliance on latent-trait claims about patient experience and quality of life. Falsification is invoked rhetorically, but simulation outputs are treated as decision-relevant evidence even though they are conditional projections insulated from empirical refutation.

In this sense, the Tufts CEA Registry knowledge base is not a neutral record of the literature but a stabilizing environment for a specific HTA memeplex. It rewards conformity to cost-utility conventions, renders alternative measurement frameworks professionally invisible, and transforms repetition into apparent confirmation. By presenting thousands of ICERs and QALY-based evaluations as a coherent empirical field, the registry creates the illusion of progressive knowledge accumulation while systematically excluding the measurement conditions that would make such accumulation possible. The registry therefore functions less as an archive of scientific discovery

and more as an institutional technology for preserving, normalizing, and legitimizing false measurement at scale.

.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: TUFTS CEA REGISTRY

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.

THE TUFTS CEA REGISTRY AS A MEMORY OF A FALSE MEASUREMENT HISTORY IN HEALTH TECHNOLOGY ASSESSMENT

Any archive that claims to preserve quantitative knowledge must begin with a prior question: what qualifies as a measure? Under the axioms of representational measurement theory, only two forms of measurement are admissible for evaluating therapy impact. Manifest attributes that are directly observable may be expressed on linear ratio scales with a true zero and invariant units. Latent attributes, which cannot be directly observed and must be inferred from patterned responses, may only be expressed on Rasch logit ratio scales that demonstrate unidimensionality, invariance, and meaningful unit structure. No other numerical constructions qualify as measures. This distinction is not optional. It is the boundary between arithmetic and numerology. This is a key to the Table 1 diagnostic results.

TABLE 1: ITEM STATEMENT, RESPONSE, ENDORSEMENT AND NORMALIZED LOGITS TUFTS CEA REGISTRY

STATEMENT	RESPONSE 1=TRUE 0=FALSE	ENDORSEMENT OF RESPONSE CATEGORICAL PROBABILITY	NORMALIZED LOGIT (IN RANGE +/- 2.50)
INTERVAL MEASURES LACK A TRUE ZERO	1	0.15	-1.75
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	0.85	+1.75
MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC	1	0.10	-2.20
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	0.90	+2.20
THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE	0	0.85	+1.75
CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.15	-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	0.65	+0.60
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	0.65	+0.60
THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT	1	0.15	-1.75
THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.05	-2.50

The Tufts Cost-Effectiveness Analysis Registry was established without any recognition of this distinction. From its inception, the registry did not ask whether the quantities it archived were measures. It assumed they were. The project began not as a measurement enterprise but as a cataloguing exercise: collect published cost-utility analyses, index their outputs, summarize their results, and present them as a cumulative empirical literature. In doing so, Tufts embarked on a decades-long project of archiving numbers whose measurement status was never examined.

The 24-item diagnostic makes this failure explicit. The foundational axioms that determine whether arithmetic is permissible are rejected or driven to the floor of endorsement. Measurement must precede arithmetic is endorsed at $p = 0.10$ with a canonical logit of -2.20. Meeting the axioms of representational measurement as a prerequisite for arithmetic sits at the same level. Multiplication requires ratio measures, the single condition that determines whether cost-effectiveness ratios can exist at all, is likewise rejected at $p = 0.10$ (-2.20). These are not methodological subtleties. They are the rules that determine whether numbers represent quantities.

At the same time, the registry's load-bearing numerical objects receive near-ceiling reinforcement. The proposition that the QALY is a ratio measure is endorsed at $p = 0.95$ ($+2.50$). That QALYs can be aggregated receives identical endorsement. EQ-5D algorithms are treated as producing interval measures at $p = 0.90$ ($+2.20$). Negative values are accepted within supposed ratio scales at $p = 0.90$ ($+2.20$). Summation of Likert scores is endorsed as creating ratio measures at $p = 0.90$ ($+2.20$). These beliefs are mathematically incompatible with representational measurement, yet they are precisely the beliefs required for the registry to function.

This reveals the registry's true role. It is not a neutral library of economic evaluations. It is an institutional mechanism that stabilizes inadmissible arithmetic by treating its outputs as comparable objects. A registry can only summarize distributions, trends, and benchmarks if the indexed quantities are commensurable. The Tufts registry therefore presupposes commensurability while rejecting the measurement conditions that make commensurability possible.

The most damaging omission concerns latent traits. If patient experience, quality of life, functioning, or burden are to be quantified at all, then Rasch measurement is not one option among many. It is the only transformation that produces invariant measurement of latent attribute possession. The diagnostic shows that every Rasch-related proposition collapses to the absolute floor of endorsement at $p = 0.05$ (-2.50). The idea that there are only two admissible classes of measurement, linear ratio for manifest attributes and Rasch logit ratio for latent traits, is rejected entirely. The proposition that subjective responses can only be transformed into interval measures using Rasch rules is likewise rejected. The Rasch logit ratio scale as the sole basis for latent trait impact assessment is excluded from the knowledge base. This is not ignorance. It is structural exclusion.

The registry does not merely fail to use Rasch measurement; it cannot allow Rasch measurement. If Rasch were accepted as a governing requirement, the vast majority of archived utilities, QALYs, mapped scores, and model outputs would be exposed as non-measures. The registry would lose its object. What remains instead is a substitute belief: that summation creates measurement. Once that belief is accepted, arithmetic becomes administratively convenient, and the registry can operate as if it were curating quantitative science.

The consequence is that the Tufts registry does not preserve measured therapy impact. It preserves a history of how false measurement became normalized. It is a memory system for a belief structure, not a repository of empirical quantities. Each archived ICER appears to contribute to cumulative knowledge, yet no invariant unit exists across studies. Each QALY appears comparable across diseases, yet no unidimensional attribute has been measured. Each distribution of thresholds appears empirical, yet the dependent variable lacks measurement status.

This is why the registry gives the illusion of progress. As the number of archived studies grows, the field appears increasingly mature. But elaboration is not discovery. Without measurement, accumulation produces volume, not knowledge. The registry transforms repetition into apparent evidence by cataloguing outputs that cannot, even in principle, support falsification or replication in the strong scientific sense.

In this light, the Tufts CEA Registry should be understood not as a failure of execution but as a failure of conception. Its founders did not distinguish between numbers and measures. They did not ask what arithmetic requires. They did not specify admissible scale types. They began a voyage to collect cost-effectiveness results without knowing what measurement was. The registry therefore stands today as an extraordinary historical artifact: a comprehensive archive of how an entire field learned to treat non-measures as if they were quantities.

Whether the registry can survive depends on whether it can be re-founded. Survival would require abandoning the idea that cost per QALY results are admissible evidence and replacing it with a

measurement-first architecture. Only linear ratio measures for manifest attributes and Rasch logit ratio measures for latent traits could be indexed. Composite ratios would be excluded. Utilities without invariant units would be rejected. Simulation outputs would be labeled as conditional projections, not evidence.

Absent that transformation, the Tufts registry will continue to exist institutionally, but not scientifically. It will remain what the diagnostic reveals it to be: not a record of measured therapy value, but the most complete archive ever assembled of false measurement in health technology assessment.

THE TUFTS REGISTRY DATA AND REFERENCE CASE MODELING.

Anyone considering the use of Tufts Cost-Effectiveness Analysis Registry data to support a reference-case simulation model should be given a clear and formal warning**:** the reference-case model itself is inadmissible as a scientific evaluative framework, irrespective of the numerical values inserted into it. The problem is not merely the quality of the inputs. It is that the framework cannot, even in principle, generate credible, evaluable, or falsifiable claims.

Reference-case models are constructed to produce cost-effectiveness ratios, threshold comparisons, and long-horizon projections. Yet these outputs depend entirely on arithmetic operations that require valid measurement. Under representational measurement theory, arithmetic is lawful only when applied to quantities possessing appropriate scale properties. The reference-case framework violates this requirement at its foundation. It performs multiplication, division, aggregation, and averaging on variables that are not measures. No adjustment to inputs can correct this defect.

The decisive point is therefore categorical: a framework built on inadmissible arithmetic cannot be rescued by better data. Even if one were to replace a particular utility algorithm, revise a mapping function, or update a parameter estimate, the structure of the model would remain unchanged. It would still require ratio-scale effects where none exist, still multiply time by non-ratio preference scores, still aggregate composite constructs across individuals, and still present the results as quantitative evidence. This is not a technical limitation; it is a logical impossibility.

In this context, reliance on the Tufts registry compounds the failure. The registry archives precisely those numerical artifacts required to sustain the reference-case paradigm: utilities treated as cardinal quantities, QALYs treated as ratio measures, ICERs treated as interpretable ratios, and thresholds treated as meaningful decision rules. These artifacts are not merely imperfect measures; they are **non-measures**. Using them does not approximate measurement; it institutionalizes its absence.

Consequently, no one should think in terms of "improving" a reference-case model by sourcing values from Tufts or similar archives. That approach assumes that the model is legitimate and that the problem lies in parameter uncertainty. This assumption is false. The reference-case model is disallowed because it cannot produce claims that are empirically evaluable. Its outputs are not hypotheses that can be tested against observed outcomes. They are conditional projections whose truth or falsity cannot be determined by experience.

This distinction matters. Scientific claims must expose themselves to potential refutation within a meaningful time frame. Reference-case outputs do not. They are immune to falsification because they describe what would happen if a set of assumptions were true, not what did happen in the world. Sensitivity analysis explores internal model behavior; it does not convert assumptions into evidence. Stability across scenarios is not empirical validation.

For this reason, any attempt to anchor pricing, access, or coverage decisions to reference-case results represents a fundamental breach of scientific discipline. It substitutes numerical plausibility for testable knowledge. It creates the appearance of rigor while removing the possibility of learning. When such models are presented as decision variables, the health system is no longer evaluating therapies; it is negotiating around stories expressed in numbers.

The appropriate conclusion is therefore not caution, but prohibition. Reference-case simulation models should not be used as evidentiary platforms for therapy evaluation. They may at best serve as illustrative thought experiments, clearly labeled as speculative and non-empirical. They cannot support value claims, comparative effectiveness claims, or pricing claims under any defensible theory of measurement.

If health systems, manufacturers, or analysts wish to generate credible evidence, the path forward is necessarily different. Claims must be limited to what can be measured. Manifest attributes must be expressed on linear ratio scales. Latent traits must be measured using Rasch logit ratio scales with demonstrated invariance. Each claim must be supported by a protocol capable of empirical evaluation within a defined timeframe. Only then can arithmetic be applied, and only then can falsification occur.

Until that transition is made, both the Tufts registry and the reference-case framework should be understood for what they are: not scientific instruments, but historical artifacts of a belief system that substituted calculation for measurement. No amount of refinement can transform them into tools for the evolution of objective knowledge.

WHERE THE TUFTS CEA REGISTRY FAILS: A DATABASE CANNOT EXIST WITHOUT MEASUREMENT

For a database whose stated purpose is to archive, compare, and summarize quantitative claims about therapy value, measurement is not a methodological option. It is a non-negotiable precondition. A registry that purports to organize numerical evidence must ensure that every element it contains satisfies the axioms of fundamental measurement. Without that requirement, the database does not function as a scientific archive. It functions as a catalogue of numerically formatted beliefs.

The Tufts Cost-Effectiveness Analysis Registry presents itself as neutral infrastructure: a repository of published cost-utility studies, utilities, QALYs, ICERs, thresholds, and modeling conventions. Yet a database is never neutral. By deciding what to index, what to standardize, and what to present as comparable, it defines what counts as admissible quantitative knowledge. In this sense, the Tufts registry is not merely recording the HTA literature; it is asserting that the objects it archives are commensurable measures.

That assertion carries an unavoidable responsibility. A database of measures can only exist if the quantities it stores are measures in the scientific sense. Representational measurement theory specifies the conditions under which numbers can represent empirical attributes. These conditions are not optional. For arithmetic to be meaningful, the attribute must be unidimensional, units must be invariant, and where multiplication or division is proposed, a true zero must exist. If these conditions are not met, numerical operations do not yield quantitative information.

From these axioms, only two admissible forms of measurement exist. Manifest attributes may be expressed on linear ratio scales when a true zero is present. Latent attributes may be expressed on Rasch logit ratio scales when unidimensionality and invariance are empirically demonstrated. No other forms of quantitative measurement are possible. There is no third category. There is no pragmatic exception. There is no statistical workaround.

The Tufts registry makes no such distinction. It does not separate manifest from latent attributes. It does not require demonstration of unidimensionality. It does not require invariant units. It does not require Rasch transformation for subjective data. Instead, it treats utilities, QALYs, and ICERs as if they were already quantitative objects, suitable for aggregation, comparison, and summarization across studies, disease areas, and populations.

This is the decisive failure. A registry cannot assume measurement; it must enforce it. Without enforcement, the act of archiving becomes epistemically meaningless. When ordinal scores, preference algorithms, composite indices, and model-dependent outputs are indexed as if they were measures, the database confers legitimacy on objects that do not possess measurement properties.

The consequences are profound. A single cost-utility study may be defensible as an internally coherent exercise in scenario construction. But when thousands of such studies are assembled into a registry, the appearance of cumulative quantitative knowledge emerges. Distributions of ICERs can be plotted. Thresholds can be discussed. “Typical” values can be inferred. Yet none of this accumulation is valid unless the underlying quantities are commensurable measures. Without measurement, aggregation does not produce knowledge; it produces repetition.

This is why the diagnostic results are so damaging in a registry context. The axioms that would police admissibility—measurement preceding arithmetic, the necessity of ratio scales for multiplication, the requirement of unidimensionality, and the need for Rasch transformation of latent traits—are rejected or marginalized. At the same time, the propositions that permit registry functionality—summation of subjective instruments, construction of utilities, aggregation of QALYs, and acceptance of negative “ratio” values—are strongly reinforced.

The result is an archive built entirely from non-measures. Costs per outcome are not admissible ratios because neither the numerator nor the denominator satisfies ratio-scale requirements. Utilities are not measures because they are derived from ordinal responses without invariant units. QALYs are not measures because they multiply time by non-measured preferences and permit negative values while claiming ratio status. ICERs therefore cannot be measures, because ratios of non-measures remain non-measures.

A database composed of such elements cannot function as a scientific memory. It cannot support comparison across contexts, replication in the strong sense, or falsification. What it preserves is not empirical knowledge, but the historical record of a belief system that treated arithmetic as a substitute for measurement.

The Tufts CEA Registry therefore does not fail because of implementation detail, modeling choice, or insufficient transparency. It fails at the level of admissibility. It was constructed without recognizing that measurement must precede archiving just as measurement must precede arithmetic. The decision to collect and standardize cost-utility outputs was made without first establishing whether those outputs could ever qualify as measures. In scientific terms, a registry of non-measures is not a database of evidence. It is a database of conventions. It records what the field has chosen to believe, not what it has succeeded in measuring. The longevity of the registry does not strengthen its epistemic status; it only documents the persistence of the error.

If the Tufts registry were to be rebuilt under measurement-first principles, its scope would change radically. Only manifest ratio outcomes would be admissible as quantitative claims. Latent attributes would require Rasch logit ratio measurement before inclusion. Composite indices would be excluded or reclassified as descriptive profiles. Simulation outputs would be treated as conditional projections rather than empirical quantities. Without these constraints, no database, however extensive, can claim to represent quantitative knowledge. A registry that does not enforce the axioms of fundamental measurement cannot curate science. It can only curate numbers. And numbers without measurement are not evidence; they are artifacts.

3. THE TRANSITION TO MEASUREMENT IN HEALTH TECHNOLOGY ASSESSMENT

THE IMPERATIVE OF CHANGE

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

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

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

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

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

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

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

MEANINGFUL THERAPY IMPACT CLAIMS

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

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

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

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

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

THE PATH TO MEANINGFUL MEASUREMENT

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

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

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

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

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

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

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

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

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

TRANSITION REQUIRES TRAINING

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

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

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

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

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

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

Maimon Research has developed a two-part training program specifically to support this transition. The first component provides foundational instruction in representational measurement theory, including the historical origins of scale theory, the distinction between manifest and latent attributes, and the criteria that define admissible claims. The second component focuses on application, detailing claim types, protocol design, and the practical use of Rasch methods to support latent trait evaluation.

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

A NEW START IN MEASUREMENT FOR HEALTH TECHNOLOGY ASSESSMENT

For readers who are looking for an introduction to measurement that meets the required standards, Maimon Research has just released two distance education programs. These are:

- Program 1: Numerical Storytelling – Systematic Measurement Failure in HTA.
- Program 2: A New Start in Measurement for HTA, with recommendations for protocol-supported claims for specific objective measures as well as latent constructs and manifested traits.

Each program consists of five modules (approx. 5,500 words each), with extensive questions and answers. Each program is priced at US$65.00. Invitations to participate in these programs will be distributed in the first instance to 8,700 HTA professionals in 40 countries.

More detail on program content and access, including registration and on-line payment, is provided with this link: https://maimonresearch.com/distance-education-programs/

DESIGNED FOR CLOSURE

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

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

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

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

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

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

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

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

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

ACKNOWLEDGEMENT

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

REFERENCES

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

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

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

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