MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

UNITED STATES: COMPLETE SYSTEMIC MEASUREMENT FAILURE IN HEALTH TECHNOLOGY ASSESSMENT

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

LOGIT WORKING PAPER No 1 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 reports that follow provide the empirical confirmation of this claim. They show, with unsettling consistency, that the global HTA knowledge base neither possesses nor applies the principles of scientific measurement.

The objective of this assessment is to evaluate, using the 24-statement diagnostic, the extent to which the United States HTA knowledge environment recognizes and applies the axioms of representational measurement theory in evaluating therapy impact. Unlike jurisdictions with a single statutory HTA authority, the United States operates through a diffuse but highly influential HTA ecosystem composed of guideline-setting organizations, academic centers, journals, payer institutions, and quasi-authoritative bodies such as ICER. This analysis does not assess policy outcomes or access decisions *per se*; it interrogates the epistemic foundations of the U.S. HTA belief system by examining whether true statements about measurement are endorsed and whether false but entrenched propositions are reinforced. The focus is on whether arithmetic is properly constrained by measurement, whether unidimensionality and scale type are respected, and whether latent traits are handled in a manner consistent with Rasch and representational measurement axioms.

The findings are not merely unfavorable to prevailing U.S. health technology assessment practice; they are devastating. When the results are expressed on the corrected canonical logit scale, they reveal a belief system that is internally coherent only because it has expelled the axioms of representational measurement from consideration. This is not a case of occasional misuse, conceptual slippage, or methodological disagreement. The diagnostic profile shows systematic, near-ceiling endorsement of propositions that are mathematically impossible, coupled with near-floor rejection of the conditions that would make quantitative claims admissible. Arithmetic is not constrained by measurement in this system; it governs in its absence. What is being practiced is not quantitative evaluation, but sanctioned numerology.

The extremity of the inversion cannot be overstated. Propositions that should function as non-negotiable entry conditions for any scientific use of numbers—measurement precedes arithmetic, multiplication requires ratio scales, latent traits require invariant measurement—are rejected at the strongest levels observed in any jurisdiction assessed to date. At the same time, the propositions required to keep the cost-utility and QALY machinery operational are endorsed with overwhelming confidence. This is not confusion; it is a deliberate epistemic settlement. U.S. HTA

has chosen to preserve its arithmetic outputs by denying the rules that would invalidate them. The result is a closed belief system in which models replace measurement, summation replaces invariance, and numerical outputs acquire authority not because they represent anything real, but because the system has agreed to treat them as if they do.

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.

DEFINING THE UNITED STATES HTA KNOWLEDGE BASE

The United States HTA knowledge base is not centralized, but it is highly coherent. It consists of an interlocking network of institutions, practices, and publications that collectively define what is considered acceptable evidence, valid analysis, and "good practice" in therapy evaluation. At its core are methodological norms promulgated through ISPOR task forces, short courses, and conference proceedings, which establish utilities, QALYs, ICERs, and reference-case simulation as default analytical tools. These norms are reinforced by flagship journals such as *Value in Health*, *Journal of Managed Care & Specialty Pharmacy*, *American Journal of Managed Care*, and related outlets, which function as gatekeepers for methodological legitimacy while rarely, if ever, interrogating scale properties or measurement axioms.

A second pillar of the knowledge base is ICER, which, although formally independent and advisory, exerts disproportionate influence over payer discourse, state policy debates, and media narratives. ICER's reference-case framework is widely treated as authoritative despite its reliance on ordinal utilities, composite outcomes, and lifetime simulation outputs that cannot be empirically tested. Its methodological assumptions circulate freely through academic centers, consultancy groups, and payer analytics, reinforcing a common evaluative language across the system.

Academic HTA and outcomes research centers embedded in U.S. universities form a third pillar. These centers train students, produce peer-reviewed analyses, and frequently contract with public and private bodies to deliver cost-effectiveness models and value assessments. Despite their proximity to formal scientific training, these centers overwhelmingly reproduce the same utility-based, model-driven framework, with no visible engagement with representational measurement theory, Stevens' scale typology, or Rasch measurement. The absence of measurement theory from curricula and research outputs ensures that false assumptions are transmitted intact to successive cohorts.

Finally, payer organizations, including commercial insurers and Medicare Advantage plans, internalize this framework through health economic dossiers, AMCP formats, and internal modeling teams. Here, the outputs of the academic and quasi-academic system are operationalized into coverage and pricing decisions, further entrenching the belief that arithmetic on non-measures is not only acceptable but necessary.

Taken together, this diffuse but tightly coupled system constitutes a closed epistemic environment. Its boundaries are defined not by empirical challenge or theoretical rigor, but by methodological conformity. Within this environment, questioning whether utilities are measures, whether QALYs are dimensionally coherent, or whether simulation outputs can be falsified is treated as irrelevant or disruptive. The 24-statement logit profile captures this closure precisely: the United States HTA

knowledge base is not ignorant at random; it is systematically organized around the denial of measurement.

CATEGORICAL PROBABILITIES

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

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

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

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

theory, whereas 0.85 indicates that the domain routinely reinforces it. Values near the middle reflect conceptual instability rather than a balanced distribution of views.

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

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

The LLM computes these categorical probabilities from three sources:

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

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

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

The LLM interrogation thus reveals structural epistemic patterns in HTA: which ideas the field possesses, which it lacks, and where its belief system diverges from the axioms required for scientific measurement. It is a diagnostic of the *knowledge behavior* of the HTA domain, not of

individuals. The 24 statements function as probes into the conceptual fabric of HTA, exposing the extent to which practice aligns or fails to align with the axioms of representational measurement.

INTERROGATION STATEMENTS

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

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

Measurement Theory & Scale Properties

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

Measurement Preconditions for Arithmetic

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

Rasch Measurement & Latent Traits

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

Properties of QALYs & Utilities

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

Falsifiability & Scientific Standards

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

Logit Fundamentals

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

Latent Trait Theory

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

AI LARGE LANGUAGE MODEL STATEMENTS: TRUE OR FALSE

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

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

INTERPRETING TRUE STATEMENTS

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

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

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

INTERPRETING FALSE STATEMENTS

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

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

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

2. SUMMARY OF FINDINGS FOR TRUE AND FALSE ENDORSEMENTS: UNITED STATES

Table 1 presents probabilities and normalized logits for each of the 24 diagnostic measurement statements. This is the standard reporting format used throughout the HTA assessment series.

It is essential to understand how to interpret these results.

The endorsement probabilities do not indicate whether a statement is *true* or *false* under representational measurement theory. Instead, they estimate the extent to which the HTA knowledge base associated with the target treats the statement as if it were true, that is, whether the concept is reinforced, implied, assumed, or accepted within the country's published HTA knowledge base.

The logits provide a continuous, symmetric scale, ranging from +2.50 to –2.50, that quantifies the degree of this endorsement. the logits, of course link to the probabilities (p) as the logit is the natural logarithm of the odds ratio; logit = ln[p/1-p].

- Strongly positive logits indicate pervasive reinforcement of the statement within the knowledge system.
- Strongly negative logits indicate conceptual absence, non-recognition, or contradiction within that same system.
- Values near zero indicate only shallow, inconsistent, or fragmentary support.

Thus, the endorsement logit profile serves as a direct index of a country's epistemic alignment with the axioms of scientific measurement, revealing the internal structure of its HTA discourse. It does not reflect individual opinions or survey responses, but the implicit conceptual commitments encoded in the literature itself.

TABLE 1: ITEM STATEMENT, RESPONSE, ENDORSEMENT AND NORMALIZED LOGITS UNITED STATES

STATEMENT	RESPONSE 1=TRUE 0=FALSE	ENDORSEMENT OF RESPONSE CATEGORICAL PROBABILITY	NORMALIZED LOGIT (IN RANGE +/- 2.50)
INTERVAL MEASURES LACK A TRUE ZERO	1	0.20	-1.40
MEASURES MUST BE UNIDIMENSIONAL	1	0.25	-1.10

MULTIPLICATION REQUIRES A RATIO MEASURE	1	0.15	-1.75
TIME TRADE-OFF PREFERENCES ARE UNIDIMENSIONAL	0	0.85	+1.75
RATIO MEASURES CAN HAVE NEGATIVE VALUES	0	0.90	+2.20
EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES	0	0.90	+2.20
THE QALY IS A RATIO MEASURE	0	0.90	+2.20
TIME IS A RATIO MEASURE	1	0.95	+2.20
MEASUREMENT PRECEDES ARITHMETIC	1	0.15	-1.75
SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES	0	0.85	+1.75
MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC	1	0.15	-1.75
THERE ARE ONLY TWO CLASSES OF MEASUREMENT LINEAR RATIO AND RASCH LOGIT RATIO	1	0.10	-2.20
TRANSFORMING SUBJECTIVE RESPONSES TO INTERVAL MEASUREMENT IS ONLY POSSIBLE WITH RASH RULES	1	0.10	-2.20
SUMMATION OF LIKERT QUESTION SCORES CREATES A RATIO MEASURE	0	0.90	+2.20
THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE	0	0.85	+1.75
CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.20	-1.40
QALYS CAN BE AGGREGATED	0	0.95	+2.50
NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.75	+1.10
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.85	+1.75
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	0	0.60	+0.40

BE COMBINED WITH A LOGIT SCALE			
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

UNITED STATES: THE INSTITUTIONALIZATION OF ARITHMETIC WITHOUT MEASUREMENT

The United States does not lack technical sophistication in health technology assessment. On the contrary, it exhibits an extraordinary density of analytic expertise, modeling capability, and statistical fluency across government agencies, academic centers, consultancies, professional societies, and private payers. What the 24-item diagnostic reveals, however, is that this sophistication is systematically decoupled from the axioms that determine whether numerical claims are scientifically admissible. The problem is not error at the margins. It is a national belief system in which arithmetic is treated as authoritative while measurement is treated as optional, negotiable, or irrelevant.

The defining feature of the U.S. profile is a structural inversion of scientific order. The proposition that measurement must precede arithmetic sits at $p = 0.15$ with a canonical logit of -1.75, placing it firmly in the rejection region. This is not a minor lapse. It is a categorical denial of the rule that gives numbers meaning. At the same time, propositions that presuppose lawful arithmetic are endorsed at near-ceiling levels. QALYs can be aggregated at $p = 0.95$ (+2.50). QALYs are ratio measures at $p = 0.90$ (+2.20). EQ-5D algorithms create interval measures at $p = 0.90$ (+2.20). Summated Likert scores create ratio measures at $p = 0.90$ (+2.20). These endorsements are not independent. They form a tightly coupled belief structure that permits multiplication, aggregation, and optimization without ever establishing the existence of a measure.

This inversion explains why cost-effectiveness analysis occupies such a privileged position in U.S. decision making despite being indefensible under representational measurement theory. Ratio arithmetic requires ratio-scaled quantities. Yet the proposition that multiplication requires a ratio measure is endorsed at only $p = 0.15$ (-1.75). The United States therefore denies the condition under which cost can be divided by effect while continuing to treat cost-effectiveness ratios as meaningful decision variables. The ICER persists not because it meets scientific requirements, but because the requirements have been excluded from the evaluative architecture.

The denominator of the ICER reveals the depth of the failure. Preference-based utilities are derived from ordinal responses to multiattribute instruments. They lack a true zero, fail invariance requirements, and do not demonstrate unidimensionality. None of this is treated as disqualifying. On the contrary, the belief that summations of subjective instrument responses are ratio measures sits at $p = 0.85$ (+1.75), and the belief that summation of Likert question scores creates a ratio

measure sits at p = 0.90 (+2.20). These values indicate doctrinal reinforcement, not casual error. Ordinal categories are treated as quantities because the system requires them to be so.

Unidimensionality, the most basic requirement for measurement, is rejected. Measures must be unidimensional sits at p = 0.25 (−1.10). Yet time trade-off preferences are treated as unidimensional at p = 0.85 (+1.75). This contradiction is not resolved empirically. It is resolved rhetorically. Multiattribute constructs are declared unidimensional by assumption so that arithmetic can proceed. Dimensionality becomes a convenience, not a property to be demonstrated.

The QALY block of the diagnostic exposes the full scope of the inversion. The United States endorses the fiction that the QALY is dimensionally homogeneous at p = 0.85 (+1.75) while simultaneously rejecting the measurement axioms that would make homogeneity meaningful. Aggregation is endorsed at the maximum level. Yet aggregation presupposes that what is being added is the same quantity expressed on a ratio scale. In the U.S. system, aggregation is treated as a policy necessity rather than a measurement consequence. The arithmetic outcome is preserved by denying the rule that would invalidate it.

The most severe failure appears in the treatment of latent traits and patient-reported outcomes. Every Rasch-related proposition collapses to the floor of the scale. The claim that there are only two admissible classes of measurement—linear ratio scales for manifest attributes and Rasch logit ratio scales for latent traits—sits at p = 0.10 (−2.20). The claim that transforming subjective responses to interval measurement is only possible with Rasch rules sits at p = 0.10 (−2.20). The claim that the Rasch logit ratio scale is the only basis for assessing therapy impact for latent traits collapses to p = 0.05 (−2.50). These values are decisive. They indicate categorical rejection.

This rejection has direct consequences. Latent attributes such as quality of life, burden, functioning, and wellbeing are invoked continuously in U.S. HTA, yet the outcome of interest for latent traits being possession of that trait sits at p = 0.25 (−1.10). The system prefers to talk about changes in scores, differences in means, and responder thresholds rather than confronting the question of how much of an attribute a population possesses. Possession is dangerous because it demands invariant units. Invariant units lead to Rasch. Rasch leads to invalidation of most existing instruments. The belief system avoids that path by weakening the very concept of possession.

Modeling practice completes the inversion. The United States endorses the principle that non-falsifiable claims should be rejected at p = 0.75 (+1.10), aligning rhetorically with Popperian norms. Yet it simultaneously endorses the belief that reference-case simulations generate falsifiable claims at p = 0.85 (+1.75). They do not. Simulation outputs are conditional projections dependent on assumptions and non-measured inputs. Sensitivity analysis explores internal model behavior; it does not expose claims to empirical refutation. The system resolves this contradiction by redefining falsification as robustness across scenarios rather than vulnerability to being wrong.

The presence of correct technical knowledge does not mitigate the failure. Recognition that the logit is the natural logarithm of the odds ratio sits at p = 0.65 (+0.60). Mathematical vocabulary is present. Measurement discipline is not. Knowledge is compartmentalized so that it cannot threaten the core arithmetic practices.

What emerges from the logit profile is not confusion but coherence. Propositions that would constrain arithmetic are pushed to the negative extreme. Propositions that enable arithmetic are pushed to the positive extreme. The system is stable because it is internally consistent. It is also scientifically bankrupt.

The consequences are profound. Without measurement, the United States cannot generate cumulative knowledge about therapy impact. Claims cannot be replicated in the strong sense because there is no invariant quantity to reproduce. Disagreements are resolved through negotiation, sensitivity analysis, or appeals to context rather than empirical refutation. Evidence becomes consensus. Replication becomes repetition. Objectivity is replaced by procedural legitimacy.

Defenders often invoke pragmatism. Decision makers need numbers, it is said, even if those numbers are imperfect. This argument collapses immediately. Measurement axioms do not constrain decisions; they constrain claims. They do not prevent action; they prevent pretending that something has been measured when it has not. A system that rejects these axioms does not become practical. It becomes unaccountable.

The remedy is not incremental reform. It is categorical. Only two classes of quantitative claims are admissible if U.S. HTA is to claim scientific legitimacy. Manifest attributes must be expressed on linear ratio scales. Latent traits must be measured on Rasch logit ratio scales with demonstrated invariance. Aggregation must be prohibited unless dimensional homogeneity is established. Simulation outputs must be reclassified as conditional projections without evidentiary authority.

Until those conditions are met, the conclusion is unavoidable. The United States has institutionalized arithmetic without measurement as a governing principle of health technology assessment. The probabilities and canonical logits do not describe a field in transition. They describe a belief system that has resolved the tension between science and convenience by rejecting science.

THE SYSTEMATIC NON-RESONANCE OF MEASUREMENT AXIOMS

One of the most striking findings in the U.S. diagnostic profile is not simply the weak reinforcement of measurement axioms but the complete absence of conceptual resonance. For more than forty years, the axioms of representational measurement theory and the Rasch model for latent traits have existed in parallel to the American HTA enterprise, fully developed, internationally recognized, empirically grounded, and mathematically coherent. Yet these axioms have never penetrated the conceptual fabric of U.S. HTA. They have not been debated, rejected, criticized, or evaluated; they have simply been ignored. The negative logits across the measurement items capture this absence with ruthless clarity. The field does not merely lack knowledge of fundamental measurement theory; it behaves as though the very idea of measurement were irrelevant.

The U.S. HTA system operates as if numbers need no justification. Utilities are treated as quantities without ever asking what the numbers represent. QALYs are multiplied, discounted, and aggregated without interrogating whether the construct satisfies the axioms that make

multiplication possible. PRO instruments are treated as ordinal summaries requiring no transformation to function as measures. A latent construct such as "health-related quality of life" is invoked constantly, but the only scientific transformation model capable of turning ordinal latent responses into interval measures, Rasch, has a national endorsement probability of 0.05 and a normalized logit of –2.50. This is not accidental omission. It is institutionalized indifference to the question of whether the numerical outputs of HTA quantify anything at all.

The deeper question is why the axioms of measurement never resonated. Part of the answer lies in the historical formation of U.S. HTA. The field was built by health economists, policy scientists, and decision theorists whose intellectual foundations lay in welfare economics, expected utility theory, and operations research, not psychometrics, measurement science, or philosophy of science. The dominant assumption was that preferences could stand in for outcomes and that numerical preference scores could be treated as quantities regardless of their empirical attributes. In this intellectual climate, measurement theory was not rejected; it was invisible. The institutional culture of HTA was founded on the belief that willingness-to-pay equivalence and preference orderings could be operationaliz directly into policy. Once that belief hardened, the axioms of measurement could not resonate because they would have destabilized the entire evaluative architecture.

Yet the diagnostic pattern suggests something beyond historical accident. The strength of reinforcement for mathematically impossible propositions, QALYs as ratio measures, utilities as interval scales, simulation outputs as evidence, indicates a belief system that has adapted to protect itself from measurement critique. The U.S. HTA complex behaves as though measurement axioms are not merely unnecessary but conceptually irrelevant. The system does not attempt to justify QALYs against those axioms because the axioms lie outside its worldview. It is not that the field rejected the Rasch model; it never engaged with the question of whether latent constructs require transformation. This is why Rasch has never appeared in U.S. value assessment frameworks despite the ubiquity of latent constructs. To engage with Rasch would have forced the recognition that the PRO instruments, utility scores, and derived QALYs are not and never were measures.

Past critiques of the QALY illustrate the scale of this non-resonance. For decades, ethicists, clinicians, and patient advocates attacked the QALY for its distributive biases, disability discrimination, methodological heterogeneity, and normative assumptions about value. These critiques were often insightful, sometimes influential, and occasionally disruptive. But they all missed the essential target: the QALY is not worth attacking. It is not a flawed measure; it is not a controversial measure; it is not a misused measure. It is not a measure at all. The correct critique is not ethical or distributive but ontological. The QALY cannot be repaired because it never possessed the properties required for measurement. Debate over QALY fairness presupposes that the QALY quantifies something meaningful. It does not.

This is the fundamental disconnect exposed by the U.S. logit profile. The entire HTA enterprise has been built on constructs that do not meet the axioms of measurement. The axioms never resonated because their adoption would have invalidated the foundational tools of the field. Simulations, thresholds, ICER ratios, and cost-utility models survive only so long as measurement never intrudes. Once measurement enters the discourse, the whole architecture collapses. The U.S. diagnostic shows that the field's belief system has evolved mechanisms for conceptual immunity

to measurement critique: either by absorbing criticism into ethical or policy debates or by simply ignoring questions of scale type, dimensionality, invariance, and transformation.

The consequence is stark. The U.S. did not fail to adopt measurement theory; it built an entire evaluative edifice designed to operate without it. The failure is structural, not incidental. And this is precisely why the QALY was never the right target. The only accurate target is the belief system that allowed an impossible object to function as the national currency of value.

WHY CRITICISM BOUNCED OFF: A BRIEF HISTORY OF THE QALY/CEA ORTHODOXY

The extraordinary feature of QALY-based health technology assessment in the United States is not that criticism has been absent, but that it has been systematically neutralized. The history of HTA is a history of conceptual objections deflected, reframed, or assimilated without altering practice. From the late 1970s onward, several institutional pathways ensured that no critique—mathematical, ethical, philosophical, or methodological—could seriously threaten the QALY orthodoxy. The durability of the paradigm was not evidence of correctness. It was evidence of insulation.

The earliest debates occurred in the 1970s and 1980s as government health economists in the U.S. and UK attempted to formalize cost-effectiveness analysis. Early workshops organized by the U.S. Public Health Service (PHS), the Office of Technology Assessment (OTA), and the RAND Health Insurance Experiment framed evaluation as an exercise in "maximizing health," with health defined through preference-based summaries. These meetings set the intellectual tone: preferences were assumed to be measurable; utilities were treated as quantities; and the QALY emerged as the convenient, portable output. What was missing from these early meetings was measurement science. No discussant asked: what are the scale properties of utilities? Can ordinal preferences be multiplied by time? The absence of the measurement question in these foundational years set the stage for everything that followed.

By the 1990s, ISPOR's formation and rapid expansion created an institutional center of gravity. ISPOR conferences, from the first Global Health Care Summit to the annual Value Assessment tracks, the mathematical foundation of the QALY was settled. Panels debated discount rates, model structure, willingness to pay, and valuation methodology, but *never* the question of whether utilities and QALYs were measures. At the same time, the landmark books that shaped the field .Gold et al.'s Cost-Effectiveness in Health and Medicine (1996), Drummond et al.'s Methods for the Economic Evaluation of Health Care Programmes, Neumann et al.'s Cost-Effectiveness in Health and Medicine (2nd ed., 2016) codified QALY arithmetic and model-based evaluation. These texts did not defend the QALY against measurement theory; they simply presupposed it as valid. Once embedded in textbooks, the paradigm became self-legitimating.

Attempts at critique appeared, but they were structurally misdirected. By the 1990s disability scholars attacked the QALY on fairness grounds; ethicists challenged it on distributional grounds; health services researchers questioned its fit for chronic illness; and patient groups emphasised the erasure of lived experience. None of these critiques asked the only question that mattered: what is the scale type of a utility? Lacking the measurement critique, these criticisms could be absorbed

without altering practice. ISPOR responded by adding "equity adjustments," "contextual considerations," and "other value elements," thereby absorbing normative objections while leaving the mathematical foundation untouched.

The arrival of ICER in 2006 entrenched this insulation. ICER's annual meetings, methodological guidance, and Value Assessment Frameworks drew from the same conceptual reservoir as ISPOR, and the same underlying texts. ICER institutionalized the QALY in U.S. payer behavior precisely because there was no competing measurement-based framework. Critiques from the Arthritis Foundation, oncology groups, and disability-rights organizations were again absorbed into "stakeholder engagement" while the arithmetic remained intact.

Universities, meanwhile, reinforced the orthodoxy. Schools of public health, pharmacy, and policy adopted the Drummond-Gold-Neumann canon wholesale. HTA education in the United States did not include representational measurement theory, axiomatic scale analysis, or Rasch measurement. Students learned models, not measurement. As a result, generations of practitioners and reviewers lacked the conceptual tools to even understand the measurement critique, let alone act on it. The absence of this knowledge was not accidental; it was structurally embedded in curricula designed by economists, not measurement scientists.

The final reinforcement mechanism was the regulatory vacuum. Without a national HTA agency, U.S. practice evolved through journals like *Value in Health*, *JMCP*, and *AJMC*, all of which treated QALY-based modelling as the benchmark of methodological competence. Peer review became a conformity mechanism: manuscripts were assessed relative to ISPOR "good practices," not to scientific measurement standards. In this environment, criticism that challenged the underlying scale theory simply could not land. The HTA community had no conceptual category in which to place the critique.

The reason criticism bounced off is therefore simple: the field built an epistemic system perfectly insulated from measurement. Because the QALY was never a measure, it was never worth attacking. Because HTA never possessed measurement theory, it could not recognize the nature of the attack. And because the institutions that shaped the field reinforced belief rather than inquiry, the QALY-simulation-threshold paradigm survived not by merit but by conceptual immunity.

THE QALY: THE CONVENIENCE OF SURVIVAL

From the beginning, the QALY did not survive because it was coherent. It survived because it was convenient. The paradox is not that a mathematically impossible construct persisted for nearly half a century; the paradox is that the institutions and individuals charged with evaluating therapies never asked the only question that would have exposed the impossibility immediately: is this a measure? The history of American health technology assessment reveals a discipline born without measurement, trained without measurement, and matured without measurement. In that environment, the nonsense of the QALY was never made clear because there was no conceptual toolkit available to make it clear.

The origins of the problem are straightforward. In the 1970s and early 1980s, the early architects of cost-effectiveness analysis were economists, operations researchers, and policy analysts; people

whose thinking was structured by welfare economics and expected utility theory, not by representational measurement or psychometrics. Their training predisposed them to view preferences as quantities and indifference curves as legitimate geometries. When the first health-state valuation exercises were conducted, it was simply assumed that numerical preferences could be treated as interval or ratio scales. No one in the room had the background to recognize that preference scores are ordinal constructions and cannot be manipulated arithmetically. The QALY was born into an intellectual world where the distinction between ordinal, interval, and ratio scales was invisible.

Once the QALY entered government thinking, it became institutionalized before it became scrutinized. Agencies needed a single outcome metric to rationalize resource allocation. The politics of health expenditure demanded a device that could compare therapies across disease areas, compress outcomes into single numbers, and produce defensible decisions. The QALY served these bureaucratic functions perfectly. It needed no empirical foundation because the administrative logic was self-contained. The ease with which the QALY could be paired with costs sealed its role as the basic unit of value. Measurement theory did not merely fail to influence HTA; it had no opportunity to. Political need outran scientific justification before the scientific community even understood what was happening.

Throughout the 1980s and 1990s, the field hardened into a closed intellectual ecosystem. Textbooks became canonical. Cost-Effectiveness in Health and Medicine, the Drummond text, and later the Neumann and Sanders framework defined competence in HTA. These books were written by people who shared the same foundational assumptions and blind spots. Because they all presumed the QALY to be a valid measure, no serious examination of scale properties, dimensionality, invariance, or admissible arithmetic ever entered the literature. Each generation of students learned the methods as settled truth, and by the time they became the next generation of faculty, reviewers, and guideline authors, the paradigm had achieved complete insulation. Nobody challenged the QALY at the level of measurement because nobody in the system had been trained to recognize that measurement was the point of failure.

Criticism did arise, but always in the wrong form. Ethicists suggested that QALYs discriminated against the elderly and disabled. Clinicians argued they were insensitive to chronic diseases. Patient advocates condemned their narrow view of wellbeing. These objections were often correct but fundamentally misdirected. They targeted the distributional consequences of the QALY, not the impossibility of its arithmetic. Ethical criticism presupposes the object being criticized is at least a measure. But the QALY was never a measure, so attacking it through fairness debates was like criticizng a compass for pointing in the wrong direction when, in fact, it was not a compass at all. This is why criticism bounced off. The field responded to ethical critiques by adding modifiers, committees, context-specific adjustments, and alternative elements of value. None of these responses challenged the foundational assumption that utilities were measurable and that QALYs were constructed on interval or ratio scales. The one critique that could have collapsed the paradigm instantly, measurement theory, never entered mainstream HTA discourse.

Meanwhile, institutions like ISPOR and ICER cemented the belief system. Conferences, guidelines, modelling task forces, and journal standards all assumed that QALYs were legitimate quantities. Peer reviewers enforced those assumptions, not through malice but through ignorance.

A manuscript challenging the scale properties of utilities would have been unintelligible to reviewers trained entirely within the QALY paradigm. The blindness was structural. A field cannot see what it was never taught to see, and HTA was never taught measurement.

The deepest answer to why the nonsense was never made clear is that the QALY was intellectually protected by the very simplicity that made it appealing. It gave the illusion of measurement without the burden of measurement. It produced numbers where numbers were wanted, regardless of whether those numbers represented anything. The field mistook numerical output for scientific content. Once that mistake became embedded in teaching, practice, policy, and publication, the blindness became complete. The QALY was never worth attacking because it was never real enough to warrant attack. The real target was always the belief system that mistook mathematical convenience for scientific measurement, and until that system is dismantled, the blindness will persist.

3 THE TRANSITION TO MEASUREMENT IN HEALTH TECHNOLOGY ASSESSMENT

THE IMPERATIVE OF CHANGE

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

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

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

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

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

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

This transition is not about rejecting evidence. It is about restoring evidence to its proper meaning. It requires moving away from composite, model-driven imaginary constructs toward claims that

are measurable, unidimensional, and capable of empirical assessment over time. The remainder of this section sets out how that transition can occur in a practical, defensible, and staged manner.

MEANINGFUL THERAPY IMPACT CLAIMS

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

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

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

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

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

THE PATH TO MEANINGFUL MEASUREMENT

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

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

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

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

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

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

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

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

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

TRANSITION REQUIRES TRAINING

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

At the center of this transition is the need to understand why measurement must precede arithmetic. Representational measurement theory establishes the criteria under which numbers can legitimately represent empirical attributes. These criteria are not optional. They determine whether addition, multiplication, aggregation, and comparison are meaningful or merely symbolic. Without

this foundation, committees are left evaluating numerical outputs without any principled way to distinguish evidence from numerical storytelling.

Training must therefore begin with scale types and their permissible operations. Linear ratio measurement applies to manifest attributes that possess a true zero and invariant units, such as time, counts, and resource use. Latent attributes, by contrast, cannot be observed directly and cannot be measured through summation or weighting. They require formal construction through a measurement model capable of producing invariant units. This distinction is the conceptual fulcrum of reform, because it determines which claims are admissible and which are not.

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

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

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

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

A NEW START IN MEASUREMENT FOR HEALTH TECHNOLOGY ASSESSMENT

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

- Program 1: Numerical Storytelling – Systematic Measurement Failure in HTA.

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

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

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

DESIGNED FOR CLOSURE

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

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

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

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

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

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

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

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

ACKNOWLEDGEMENT

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

REFERENCES

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

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

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

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