MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

UNITED KINGDOM: THE *JOURNAL OF HEALTH ECONOMICS* - ARITHMETIC WITHOUT MEASUREMENT

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

LOGIT WORKING PAPER No 78 FEBRUARY 2026

www.maimonresearch.com

Tucson AZ

FOREWORD

HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

The *Journal of Health Economics* occupies a premier position within the global health economics community. It is widely regarded as the flagship theoretical and econometric journal in the field, shaping methodological standards, influencing graduate training, and conferring intellectual legitimacy on the analytical frameworks that underpin health policy evaluation. Work published in JHE does not merely contribute to debate; it defines the boundaries of what counts as rigorous economic analysis in health. Its influence extends beyond academia into policy institutions, reimbursement bodies, and international health technology assessment practice. In this sense, JHE functions as an epistemic anchor for the quantitative architecture of modern health economics.

The journal presents its mission as advancing the application of economic theory and econometric methods to health care markets, insurance design, provider behavior, and welfare evaluation. Its message is clear: disciplined modeling, formal analysis, and statistical inference provide the tools necessary to understand and improve health system performance. Through this lens, utility, welfare, and cost-effectiveness are treated as quantifiable constructs amenable to rigorous analysis. The authority of this message rests on the presumption that the quantities manipulated within these models satisfy the conditions required for meaningful measurement; a presumption this review critically examines.

The objective of this study was to evaluate whether the *JHE* operates within the axiomatic constraints required for lawful quantitative reasoning. Using the 24-item canonical representational measurement diagnostic, the analysis interrogates whether foundational propositions such as unidimensionality, scale-type admissibility, dimensional homogeneity, invariance, and the primacy of measurement over arithmetic function as binding principles within the journal’s knowledge base. The purpose is not to critique isolated methodological choices but to determine whether the journal’s evaluative architecture recognizes and enforces the necessary conditions under which numbers can legitimately represent empirical magnitudes. By assigning categorical endorsement probabilities to each canonical statement and transforming these into normalized logits, the study quantifies the presence or absence of measurement constraints within the journal’s intellectual framework.

The findings demonstrate systematic non-possession of core measurement axioms within the journal’s quantitative discourse. Propositions asserting that measurement must precede arithmetic, that multiplication requires ratio scales, that measures must be unidimensional, and that Rasch transformation is required for lawful interval construction collapse to floor or near-floor logit values, including repeated −2.50 results indicating effective absence. At the same time, propositions necessary to sustain cardinal utility manipulation, aggregation of heterogeneous constructs, and cost-effectiveness ratio formation register strong endorsement. The resulting logit profile reveals an evaluative architecture in which arithmetic operations are normalized without prior demonstration that the constructs involved satisfy representational measurement requirements. On these grounds, the journal's knowledge base fails to enforce the axioms that define measurement in normal science.

The modern articulation of the principle that measurement must precede arithmetic can be traced to Stevens' seminal 1946 paper, which introduced the typology of nominal, ordinal, interval, and ratio scales [1]. Stevens made explicit what physicists, engineers, and psychologists already understood: different kinds of numbers permit different kinds of arithmetic. Ordinal scales allow ranking but not addition; interval scales permit addition and subtraction but not multiplication; ratio scales alone support multiplication, division, and the construction of meaningful ratios. Utilities derived from multiattribute preference exercises, such as EQ-5D or HUI, are ordinal preference scores; they do not satisfy the axioms of interval measurement, much less ratio measurement. Yet HTA has, for forty years, treated these utilities as if they were ratio quantities, multiplying them by time to create QALYs and inserting them into models without the slightest recognition that scale properties matter. Stevens' paper should have blocked the development of QALYs and cost-utility analysis entirely. Instead, it was ignored.

The foundational theory that establishes *when* and *whether* a set of numbers can be interpreted as measurements came with the publication of Krantz, Luce, Suppes, and Tversky's *Foundations of Measurement* (1971) [2] . Representational Measurement Theory (RMT) formalized the axioms under which empirical attributes can be mapped to numbers in a way that preserves structure. Measurement, in this framework, is not an act of assigning numbers for convenience, it is the discovery of a lawful relationship between empirical relations and numerical relations. The axioms of additive conjoint measurement, homogeneity, order, and invariance specify exactly when interval scales exist. RMT demonstrated once and for all that measurement is not optional and not a matter of taste: either the axioms hold and measurement is possible, or the axioms fail and measurement is impossible. Every major construct in HTA, utilities, QALYs, DALYs, ICERs, incremental ratios, preference weights, health-state indices, fails these axioms. They lack unidimensionality; they violate independence; they depend on aggregation of heterogeneous attributes; they collapse under the requirements of additive conjoint measurement. Yet HTA proceeded, decade after decade, without any engagement with these axioms, as if the field had collectively decided that measurement theory applied everywhere except in the evaluation of therapies.

Whereas representational measurement theory articulates the axioms for interval measurement, Georg Rasch's 1960 model provides the only scientific method for transforming ordered categorical responses into interval measures for latent traits [3]. Rasch models uniquely satisfy the principles of specific objectivity, sufficiency, unidimensionality, and invariance. For any construct such as pain, fatigue, depression, mobility, or need, Rasch analysis is the only legitimate means of producing an interval scale from ordinal item responses. Rasch measurement is not an alternative to RMT; it is its operational instantiation. The equivalence of Rasch's axioms and the axioms of representational measurement was demonstrated by Wright, Andrich and others as early as the 1970s. In the latent-trait domain, the very domain where HTA claims to operate; Rasch is the only game in town [4].

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

How did HTA miss the bus so thoroughly? The answer lies in its historical origins. In the late 1970s and early 1980s, HTA emerged not from measurement science but from welfare economics, decision theory, and administrative pressure to control drug budgets. Its core concern was *valuing health states*, not *measuring health*. This move, quiet, subtle, but devastating, shifted the field away from the scientific question “What is the empirical structure of the construct we intend to measure?” and toward the administrative question “How do we elicit a preference weight that we can multiply by time?” The preference-elicitation projects of that era (SG, TTO, VAS) were rationalized as measurement techniques, but they never satisfied measurement axioms. Ordinal preferences were dressed up as quasi-cardinal indices; valuation tasks were misinterpreted as psychometrics; analyst convenience replaced measurement theory. The HTA community built an entire belief system around the illusion that valuing health is equivalent to measuring health. It is not.

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

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

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

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

Paul C Langley, Ph.D

Email: langleylapaloma@gmail.com

DISCLAIMER

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

1. INTERROGATING THE LARGE LANGUAGE MODEL

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

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

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

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

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

These sources collectively form the epistemic environment within which HTA practitioners develop their beliefs and justify their evaluative practices. The boundary of interrogation is thus not the whole of medicine, economics, or public policy, but the specific textual ecosystem that sustains HTA reasoning. . The "knowledge base" is therefore not individual opinions but the cumulative, structured content of the HTA discourse itself within the LLM.

THE *JOURNAL OF HEALTH ECONOMICS* KNOWLEDGE BASE

The knowledge base of the *JHE* is anchored in neoclassical economic theory, econometric modeling, and welfare analysis. Its core intellectual commitments include utility maximization, preference revelation, equilibrium modeling, marginal analysis, and statistical estimation of behavioral responses within health care markets. Articles frequently examine insurance demand, provider incentives, pharmaceutical pricing, moral hazard, public policy impacts, and health outcomes using advanced econometric techniques. Mathematical sophistication and formal modeling are central to its identity. Quantitative authority within the journal derives from internal consistency of models, robustness of estimation, and statistical significance of parameters.

However, the journal's quantitative architecture rests on assumptions about measurement that are rarely examined explicitly. Utility is treated as cardinal for modeling convenience. Health-related quality-of-life indices are incorporated into welfare analyses as if they represent magnitudes. Preference-based constructs derived from discrete choice experiments or time trade-off exercises are embedded in functional forms without interrogation of scale properties. Regression coefficients, elasticities, and marginal effects are interpreted as quantitative differences even when the dependent variables originate from ordinal or composite constructs.

The journal's econometric rigor does not substitute for measurement validation. Statistical estimation presumes that the variables involved possess scale properties compatible with the arithmetic applied. Yet the knowledge base does not enforce demonstration of unidimensionality when composite health indices are treated as singular quantities. Nor does it require proof that utility scores satisfy ratio-scale conditions before multiplicative operations are performed. The implicit assumption is that if a variable can be parameterized and estimated, it can be treated as a quantitative magnitude. This assumption reverses the correct ordering. Parameterization does not create measurement; measurement is a prerequisite for meaningful parameterization.

In welfare analyses and cost-effectiveness discussions that intersect with broader HTA practice, the journal's discourse reflects acceptance of QALY-based constructs as cardinal quantities. The multiplication of health state values by time is treated as a lawful operation, and aggregation across individuals is normalized. Dimensional homogeneity, the requirement that aggregated units represent the same empirical dimension, is not enforced as a structural constraint. The zero anchors of preference-based scales are accepted conventionally rather than validated representationally. Negative values are tolerated within constructs implicitly treated as ratio measures, contradicting the definition of ratio scale.

The absence of Rasch measurement within the journal's methodological repertoire is particularly significant. Rasch provides a formal, axiomatic route for transforming ordinal responses into invariant interval measures. Its non-possession within the knowledge base indicates that ordinal preference data are typically elevated directly into cardinal utility parameters without intermediate validation of invariance or unidimensionality. This bypassing of the measurement problem is not presented as a solution; it is normalized as practice.

The journal's commitment to empirical testing at the econometric level coexists with silence on representational measurement. Hypotheses are tested. Models are compared. Sensitivity analyses are conducted. Yet falsifiability is confined to parameter estimates within assumed structures. The measurement status of the constructs themselves is not subjected to structural falsification. A model may be rejected statistically while the arithmetic admissibility of its variables remains unquestioned.

This pattern reflects an epistemic culture in which mathematical tractability and statistical precision are equated with quantitative legitimacy. The presence of formal equations and robust standard errors creates the appearance of lawful measurement even when scale-type conditions are not satisfied. Arithmetic is normalized because it is analytically convenient. The axioms that govern when arithmetic preserves empirical meaning do not operate as editorial gatekeepers.

The consequence is that the journal's knowledge base stabilizes a form of quantitative reasoning in which numbers acquire authority by virtue of modeling sophistication rather than representational validity. Constructs are treated as magnitudes because they fit within econometric frameworks, not because they satisfy measurement axioms. The canonical logit profile quantifies this stabilization. Foundational measurement propositions collapse to floor values, indicating that they do not function as binding constraints. Enabling propositions necessary to sustain cardinal manipulation register strong endorsement.

As a leading theoretical journal, the *JHE* influences training, policy modeling, and applied HTA practice. Its treatment of measurement therefore has cascading effects throughout the health economics ecosystem. When arithmetic is decoupled from representational discipline at the theoretical core, applied frameworks inherit that detachment. The knowledge base does not merely overlook measurement constraints; it institutionalizes their absence.

On the criteria defined by representational measurement theory, this constitutes structural failure. The journal's quantitative authority rests on conventions that do not satisfy the axioms required for lawful measurement. However mathematically elegant or statistically refined, quantitative claims built upon non-measures remain representationally invalid. The canonical assessment demonstrates that the axioms defining measurement do not govern the journal's evaluative architecture. On that basis, the knowledge base cannot claim adherence to the standards of quantitative science.

CATEGORICAL PROBABILITIES

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

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

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

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

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

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

The LLM computes these categorical probabilities from three sources:

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

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

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

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

INTERROGATION STATEMENTS

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

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

Measurement Theory & Scale Properties

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

Measurement Preconditions for Arithmetic

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

Rasch Measurement & Latent Traits

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

Properties of QALYs & Utilities

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

Falsifiability & Scientific Standards

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

Logit Fundamentals

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

Latent Trait Theory

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

24. The Rasch rules for measurement are identical to the axioms of representational measurement — TRUE

AI LARGE LANGUAGE MODEL STATEMENTS: TRUE OR FALSE

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

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

INTERPRETING TRUE STATEMENTS

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

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

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

INTERPRETING FALSE STATEMENTS

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

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

- accepts dimensional incoherence

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

2. SUMMARY OF FINDINGS FOR TRUE AND FALSE ENDORSEMENTS: *JOURNAL OF HEALTH ECONOMICS*

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.

A CANONICAL LOGIT ASSESSMENT OF FALSE MEASUREMENT: *JOURNAL OF HEALTH ECONOMICS*

Measurement is not a methodological preference. It is not a modeling option. It is not a disciplinary convention. Measurement is the representational assignment of numbers to empirical attributes under axioms that determine when arithmetic is meaningful. The axioms come first. The scale type comes first. Only after these are established can arithmetic operations be admitted. This ordering is not philosophical. It is logical. Without it, numbers do not represent magnitudes. They represent categories, ranks, or constructed scores. Arithmetic performed on non-measures is not approximate measurement; it is inadmissible manipulation.

TABLE 1: ITEM STATEMENT, RESPONSE, ENDORSEMENT AND NORMALIZED LOGITS *JOURNAL OF HEALTH ECONOMICS*

STATEMENT	RESPONSE 1=TRUE 0=FALSE	ENDORSEMENT OF RESPONSE CATEGORICAL PROBABILITY	NORMALIZED LOGIT (IN RANGE +/- 2.50)
INTERVAL MEASURES LACK A TRUE ZERO	1	0.20	-1.40
MEASURES MUST BE UNIDIMENSIONAL	1	0.15	-1.75
MULTIPLICATION REQUIRES A RATIO MEASURE	1	0.10	-2.20
TIME TRADE-OFF PREFERENCES ARE UNIDIMENSIONAL	0	0.80	+1.40
RATIO MEASURES CAN HAVE NEGATIVE VALUES	0	0.85	+1.75
EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES	0	0.75	+0.95
THE QALY IS A RATIO MEASURE	0	0.75	+0.95
TIME IS A RATIO MEASURE	1	0.90	+2.20
MEASUREMENT PRECEDES ARITHMETIC	1	0.15	-1.75
SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES	0	0.80	+1.40
MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC	1	0.05	-2.50
THERE ARE ONLY TWO CLASSES OF MEASUREMENT LINEAR RATIO AND RASCH LOGIT RATIO	1	0.05	-2.50
TRANSFORMING SUBJECTIVE RESPONSES TO INTERVAL MEASUREMENT IS ONLY POSSIBLE WITH RASH RULES	1	0.05	-2.50
SUMMATION OF LIKERT QUESTION SCORES CREATES A RATIO MEASURE	0	0.85	+1.75
THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE	0	0.50	0.00
CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.80	+1.40

QALYS CAN BE AGGREGATED	0	0.85	+1.75
NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.50	0.00
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.80	+1.40
THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.40	-0.45
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	0.70	+0.85
THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT	1	0.10	-2.20
THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.05	-2.50

The 24 canonical statements applied in this assessment are not opinion questions. They are diagnostic propositions that determine whether a knowledge base possesses the conditions required for lawful quantitative reasoning. Each statement corresponds to a necessary structural feature of measurement: unidimensionality, admissible transformation, scale-type constraints, dimensional homogeneity, falsifiability, and invariance. By assigning categorical endorsement probabilities and transforming them into normalized logits, the analysis does not evaluate stylistic preference. It quantifies whether these measurement constraints operate as binding principles within the journal's evaluative framework.

The logit profile for the JHE is decisive. Foundational propositions collapse to floor or near-floor values. The claim that meeting the axioms of representational measurement is required for arithmetic registers at −2.50. The propositions that there are only two lawful classes of measurement linear ratio and Rasch logit ratio and that Rasch transformation is required for interval construction from ordinal responses also register at −2.50. The statement that the Rasch logit ratio scale is the only basis for assessing therapy impact for latent traits collapses identically. These are not peripheral claims. They are structural constraints. Their collapse to floor values indicates effective non-possession within the journal's knowledge base.

Measurement requires unidimensionality. A measure must represent magnitude along a single defined attribute. Without unidimensionality, numerical differences cannot be interpreted as quantitative differences. The canonical statement asserting that measures must be unidimensional registers near floor. That result indicates that unidimensionality does not function as an enforced precondition within the journal’s quantitative discourse. Composite constructs, multiattribute indices, and preference-based aggregates are routinely treated as singular magnitudes without prior demonstration that they satisfy representational requirements.

Measurement requires scale-type discipline. Nominal scales permit classification. Ordinal scales permit ranking. Interval scales permit addition and subtraction. Ratio scales alone admit multiplication and division because they possess a non-arbitrary zero. The canonical proposition that multiplication requires a ratio measure, registers at −2.20 logits. This is near-total absence. Yet cost-effectiveness ratios, welfare comparisons, marginal rates of substitution, and cost-per-QALY constructs depend precisely on multiplicative and ratio arithmetic. The journal’s knowledge base normalizes these operations without enforcing the prerequisite scale conditions that make them admissible.

Measurement requires dimensional homogeneity. Ratios can only be formed between quantities of the same dimension. A cost-per-unit-of-health ratio presumes that “health” is a quantity of a defined dimension with lawful scale properties. The canonical statement that the QALY is dimensionally homogeneous registers with strong endorsement of the false alternative. The implication is direct: the discipline treats heterogeneous composite indices as if they were homogeneous magnitudes. Dimensional incoherence is stabilized.

Measurement requires invariance. If two observers measure the same attribute, differences in measurement must reflect differences in magnitude, not differences in scoring conventions. Rasch measurement provides the only lawful model for transforming ordinal responses into interval measures under invariant conditions. The complete collapse of Rasch propositions to −2.50 indicates that the journal’s knowledge base does not recognize invariance as a binding constraint on latent trait measurement. Instead, ordinal preference scores and composite questionnaire totals are treated as if arithmetic alone confers magnitude.

Measurement precedes arithmetic. This is the most elementary ordering in quantitative science. The canonical proposition asserting this ordering registers near floor. The journal’s quantitative practice reverses the order. Models are specified. Utility functions are estimated. Indices are constructed. Arithmetic is applied. The measurement status of the constructs is assumed rather than demonstrated. Statistical estimation is substituted for representational validation.

The logit table also shows strong endorsement of propositions required to sustain arithmetic on non-measures. False statements such as “ratio measures can have negative values” and “the QALY is a ratio measure” register with high positive logits. This is not incidental misunderstanding. It is structural enablement. A ratio scale, by definition, possesses a true zero indicating absence of the attribute and does not admit negative values. The willingness to treat constructs with arbitrary anchors and possible negative scores as ratio quantities indicates abandonment of scale-type discipline. The arithmetic that follows is therefore representationally void.

The JHE stands at the theoretical apex of the discipline. Its norms influence training, methodological development, and policy diffusion. If representational measurement axioms do not operate as binding constraints at this level, the applied frameworks that depend upon its authority inherit that absence. The canonical assessment does not suggest that some articles occasionally overlook measurement detail. It demonstrates that the knowledge base itself does not internalize measurement axioms as governing principles.

Utility is central to the journal's intellectual architecture. Utility functions are estimated, compared, differentiated, and integrated. Yet cardinality is assumed for analytical convenience rather than established through representational demonstration. A parameterized utility function does not become a ratio quantity because it is expressed mathematically. Mathematical tractability does not create magnitude. Without lawful scale properties, algebraic operations remain symbol manipulation detached from empirical quantity.

The discipline often defends itself by invoking internal consistency, econometric rigor, and model testing. These are irrelevant to the measurement question. Statistical significance does not create scale type. Parameter stability does not generate unidimensionality. Sensitivity analysis does not confer ratio properties. A regression coefficient can be precisely estimated and still lack meaning if the dependent variable does not support the arithmetic applied to it. Measurement validity is logically prior to statistical inference.

The canonical statements are designed to quantify belief and understanding. They measure whether foundational propositions operate within the journal's evaluative architecture. Floor logits indicate effective absence. High positive logits for false enabling propositions indicate stabilization of arithmetic without lawful measurement. On these grounds, the knowledge base fails.

This failure is not marginal. It is comprehensive. Unidimensionality is not enforced. Rasch transformation is absent. Dimensional homogeneity is not required. Ratio arithmetic is normalized without ratio measures. Measurement does not precede arithmetic. These absences are not counterbalanced by econometric sophistication. They are structural violations of the logic of quantity.

A discipline that performs multiplication on non-ratio scales, aggregates heterogeneous constructs, and forms ratios from composite indices while excluding the axioms that govern such operations cannot claim quantitative legitimacy. The problem is not error within a lawful system. The problem is the absence of the system that defines lawfulness.

The logit profile therefore constitutes empirical evidence of epistemic non-possession. The propositions that define measurement in normal science do not operate within the Journal of Health Economics as binding constraints. Arithmetic is permitted without demonstration of admissibility. Composite constructs are treated as magnitudes without unidimensional validation. Ratio claims are advanced without ratio measures. These are not stylistic differences. They are violations of definitional requirements.

Because the journal occupies a position of authority, its failure stabilizes the broader evaluative framework of health technology assessment. NICE reference-case methodology, cost-per-QALY thresholds, welfare aggregation, and simulation-based policy reasoning derive intellectual legitimacy from the theoretical norms that JHE sustains. If those norms exclude representational measurement, the applied consequences inherit that exclusion.

The conclusion is not negotiable. Measurement is defined by axioms. Arithmetic is constrained by scale type. These are necessary conditions for quantitative science. The canonical assessment shows that these conditions do not govern the Journal of Health Economics. On that basis, its quantitative claims, where dependent upon inadmissible arithmetic, lack representational legitimacy.

This is not an invitation to methodological pluralism. There are not multiple lawful systems of quantity. There is one logic governing when numbers represent magnitude. Where that logic is not enforced, the result is numerical expression without measurement. However mathematically elegant, however statistically sophisticated, however widely cited, arithmetic performed on non-measures does not become measurement by repetition.

The 24-item canonical diagnostic quantifies this reality. It demonstrates that foundational measurement propositions are absent from the journal's knowledge base while false enabling propositions are stabilized. That pattern is sufficient. On the grounds of representational measurement, the Journal of Health Economics fails.

FALSE MEASUREMENT AND THE EVOLUTION OF OBJECTIVE KNOWLEDGE

The evolution of objective knowledge depends upon falsifiability. A claim must expose itself to the possibility of being wrong. This is not a stylistic feature of science; it is its defining mechanism. For a therapeutic claim to contribute to objective knowledge, it must be framed in terms of measurable magnitude. Only then can evidence contradict it. Only then can replication confirm or refute it. Without lawful measurement, falsification collapses into procedural adjustment rather than structural correction.

Representational measurement theory establishes the conditions under which numbers correspond to empirical magnitude. Unidimensionality ensures that variation reflects a single attribute. Scale type determines which arithmetic operations preserve meaning. Ratio scales admit multiplication and division; interval scales do not. Invariance ensures that measurement does not depend upon arbitrary conventions. These axioms are not philosophical preferences. They are the logical infrastructure that makes quantitative claims empirically vulnerable.

False measurement interrupts this infrastructure at its source. When ordinal preferences are treated as cardinal quantities, when composite indices are treated as unidimensional magnitudes, when ratio arithmetic is applied to scales lacking true zero properties, the resulting numerical claims are insulated from structural falsification. They may be recalculated. They may be re-estimated. They may be subjected to sensitivity analysis. But they cannot be proven wrong in the sense required for the growth of knowledge, because the magnitude they purport to represent has never been established.

Consider the case of cost-per-QALY reasoning. If the QALY were a lawful ratio measure derived from invariant scaling, then a claim such as "Intervention A produces 0.4 additional QALYs" would expose itself to empirical refutation. One could measure the magnitude and test the claim. But if the QALY is constructed by multiplying time, a ratio measure, by a preference-based utility score lacking ratio properties, then the product does not represent magnitude. It represents arithmetic applied to a non-measure. The resulting claim cannot be falsified as a magnitude claim. It can only be recalculated under alternative assumptions. Falsification becomes a change of model, not a correction of measurement.

The same applies to replication. Replication presumes invariance. It presumes that two investigators measuring the same attribute on the same lawful scale will obtain results comparable in magnitude. Without invariant measurement, replication reduces to procedural similarity. Two studies may report similar composite scores, similar utilities, or similar ratios, yet the underlying constructs remain structurally indeterminate. Agreement between non-measures does not create measurement. It creates convergence within a modeling convention.

When measurement axioms are not enforced, quantitative discourse becomes self-referential. Claims are tested against the assumptions that generated them. Sensitivity analysis replaces exposure to risk. Parameter uncertainty substitutes for structural validation. The framework evolves internally, but objective knowledge does not evolve externally. The system refines its techniques while leaving its foundational incoherence intact.

The implications are profound. Science advances when errors are identified and corrected. But error can only be identified where measurement constrains arithmetic. If arithmetic is permitted without lawful scale properties, the possibility of structural error disappears. The system cannot discover that it is wrong about magnitude because magnitude has never been defined. False measurement therefore arrests the evolution of objective knowledge at its foundation.

This is not a matter of intellectual style. It is a matter of logical necessity. Without representational measurement, quantitative claims cannot satisfy the conditions required for empirical refutation. Without empirical refutation, science cannot progress. What remains is numerical storytelling stabilized by convention and repetition. The presence of equations, models, and statistical inference does not compensate for the absence of lawful measurement. It merely increases the appearance of precision.

The evolution of objective knowledge requires exposure to risk grounded in measurement. Where measurement axioms are excluded, risk disappears. What survives is a closed quantitative system capable of endless recalculation but incapable of self-correction at the level that matters: the level of magnitude. False measurement does not merely weaken science. It prevents its evolution.

CAN THE *JOURNAL OF HEALTH ECONOMICS* ESCAPE ITS LEGACY OF FALSE MEASUREMENT?

The short answer is that escape is possible in principle but improbable in practice, and the reasons are structural rather than intellectual. The *JHE* does not merely publish work that relies on false measurement; it has helped normalize a quantitative culture in which representational measurement is treated as irrelevant. That legacy is not accidental. It follows directly from foundational commitments embedded in the journal’s conception of utility, welfare, and admissible arithmetic.

False measurement in this context does not mean occasional error or misuse. It means the systematic acceptance of numerical constructs that fail the axioms required for quantity. Cardinal utility derived from ordinal preference data, aggregation of heterogeneous attributes into single indices, multiplication of non-ratio scales, and ratio formation without dimensional homogeneity are not marginal practices. They are the enabling assumptions of the journal's core evaluative logic. The canonical logit assessment demonstrates that propositions asserting the priority of measurement axioms, scale-type constraints, and Rasch transformation collapse to floor values. This indicates that these principles do not operate as binding constraints within the journal's knowledge base. Without those constraints, arithmetic proceeds unmoored from measurement.

For the journal to escape this legacy, it would have to reverse the ordering on which its quantitative authority rests. Measurement would have to precede modeling. Scale type would have to constrain admissible operations. Unidimensionality would have to be demonstrated rather than assumed. Latent traits would require invariant transformation rather than preference aggregation. This is not a marginal methodological adjustment. It would invalidate large portions of the journal's published corpus and render central constructs such as utility functions, cost-effectiveness ratios, welfare aggregates representationally indefensible.

That is the first obstacle: path dependence. A flagship journal cannot easily declare that decades of accepted quantitative practice rest on inadmissible arithmetic. Such an acknowledgment would destabilize not only past publications but the training, peer review norms, and professional identities built around them. The cost of correction is therefore not technical but institutional.

The second obstacle is conceptual insulation. Health economics has long treated utility as a modeling primitive rather than a measurement problem. Cardinality is assumed for analytical convenience. Comparability is imposed by convention. The axioms of representational measurement are rarely engaged because they threaten to dissolve the object of analysis itself. The canonical logit profile captures this insulation empirically: statements asserting that measurement precedes arithmetic or that multiplication requires ratio scales do not register as admissible propositions within the journal's evaluative environment. They are excluded not by rebuttal but by absence.

The third obstacle is reinforcement through application. The *JHE* does not operate in isolation. Its norms are reinforced by NICE reference-case methodology, by applied HTA journals, by policy manuals, and by international diffusion through organizations such as ISPOR. Each layer stabilizes the others. Even if individual editors or authors recognized the measurement problem, institutional alignment would exert pressure toward continuity rather than rupture.

Could the journal change course? Only if it explicitly recognized representational measurement axioms as non-negotiable constraints rather than optional philosophical considerations. That would require rejecting or radically reframing constructs that cannot satisfy those axioms. It would require acknowledging that some forms of quantitative reasoning currently treated as authoritative are not merely imperfect but invalid. It would require privileging lawful measurement over modeling elegance.

There is no evidence that such a shift is underway. Editorial policies do not require demonstration of scale type. Peer review does not enforce unidimensionality or admissible arithmetic. Rasch measurement remains marginal. Instead, the journal continues to publish increasingly sophisticated models applied to constructs whose measurement status is unexamined. The appearance of rigor deepens as the foundations remain unaddressed.

The legacy of false measurement is therefore not a historical artifact that can be quietly outgrown. It is an active structural condition. Escape would require deliberate abandonment of arithmetic practices that the discipline depends upon for its evaluative claims. That is not reform. It is reconstruction.

Until such reconstruction occurs, the *JHE* cannot plausibly claim to have escaped its legacy. It continues to confer legitimacy on quantitative claims that violate the axioms defining measurement. In doing so, it stabilizes a form of numerical authority that looks scientific while remaining epistemically hollow. The issue is not whether the journal can adapt incrementally. It is whether it is willing to accept that true measurement is prior to, and constraining of, all quantitative analysis. On the present evidence, that willingness is absent.

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.

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