MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

UNITED KINGDOM: SCOTTISH MEDICINES CONSORTIUM - DEVOLUTION WITH NUMERICAL STORYTELLING

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

LOGIT WORKING PAPER No 30 JANUARY 2026

www.maimonresearch.com

Tucson AZ

FOREWORD

HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

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

The objective of this assessment is to interrogate the epistemic foundations of the Scottish Medicines Consortium (SMC) through application of the 24-item diagnostic grounded in representational measurement theory. Rather than evaluating individual reimbursement decisions or appraising the administrative efficiency of the SMC process, this analysis examines the belief system embedded in the analytical framework that the Consortium employs to define admissible evidence, legitimate arithmetic, and acceptable claims of therapy impact. The purpose is to determine whether the quantitative objects relied upon by the SMC such as utilities, QALYs, incremental cost-effectiveness ratios, and reference-case modeling outputs satisfy the axioms required for scientific measurement, falsification, and cumulative knowledge development. By translating endorsement patterns into canonical logit values, the assessment seeks to reveal not surface methodological preferences, but the deeper ordering principles that govern how the SMC understands quantity, evidence, and decision validity.

The findings demonstrate a familiar but nonetheless stark pattern. The SMC knowledge base exhibits strong endorsement of arithmetic operations central to cost-utility analysis while simultaneously rejecting or marginalizing the axioms that make those operations meaningful. Principles requiring that measurement precede arithmetic, that multiplication requires ratio scales, and that unidimensionality be demonstrated rather than assumed are weakly endorsed or rejected outright. In contrast, propositions sustaining the legitimacy of QALYs, utility aggregation, preference algorithms, and simulation-based decision variables are strongly reinforced. Rasch measurement, the only framework capable of constructing invariant measures for latent attributes, is effectively absent. The resulting logit profile reveals not methodological uncertainty but structural inversion: arithmetic is treated as authoritative, while measurement is treated as optional. In consequence, the SMC operates within an evaluative architecture that produces administrative closure without generating empirically evaluable or falsifiable claims.

The modern articulation of the principal 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 SMC KNOWLEDGE BASE

The knowledge base of the Scottish Medicines Consortium is best understood as an institutional extension of the UK reference-case tradition that emerged in the 1990s and was subsequently formalized through NICE. Although the SMC operates under a distinct remit and emphasizes timely access decisions within NHS Scotland, its analytical foundations mirror the same conceptual architecture: cost-utility analysis anchored in preference-based utilities, QALYs, and incremental cost-effectiveness ratios generated through simulation modeling. These constructs are treated as legitimate quantitative evidence rather than as conditional artifacts whose admissibility depends on prior measurement validation.

Within this knowledge base, numerical credibility is established procedurally rather than empirically. Evidence is deemed acceptable when it conforms to prescribed modeling conventions, methodological templates, and reporting standards. The framework does not require demonstration that outcome variables possess the properties necessary to support the arithmetic operations applied to them. Instead, legitimacy flows from alignment with accepted practice. Measurement is therefore not a gatekeeping condition but an implicit assumption embedded within the analytic ritual.

Central to this system is the treatment of utilities as quantitative measures. Health state preferences derived from ordinal questionnaire responses are mapped through algorithms and treated as interval or ratio quantities despite lacking a true zero, permitting negative values, and failing to demonstrate invariant unit structure. These utilities are then multiplied by time to produce QALYs, which are treated as dimensionally homogeneous measures capable of aggregation across individuals and disease states. The knowledge base does not interrogate whether such operations are permissible under representational measurement theory; it presumes that widespread use confers legitimacy.

Latent attributes such as quality of life, symptom burden, and functioning are invoked continuously but never measured in the strict sense. They are represented through summated scores or preference mappings rather than constructed through measurement models capable of producing invariant quantities. The distinction between ordinal scoring and measurement is not operationalized. Statistical performance characteristics—responsiveness, reliability, and model fit—are treated as substitutes for measurement, despite their inability to establish scale type or permissible arithmetic.

Rasch measurement occupies no structural role within the SMC framework. Although patient-reported outcomes are frequently cited as evidence of benefit, the transformation of subjective responses into invariant measures of latent trait possession is not required. As a result, latent

attributes are treated as if they were already quantified, enabling their incorporation into models without confronting whether they exist as measurable quantities at all.

The knowledge base further relies on long-horizon simulation modeling to generate decision variables that cannot be falsified through observation. Model outputs are conditional on assumptions about disease progression, treatment persistence, utilities, and extrapolation beyond observed data. Sensitivity analysis is used to explore assumption variability, but this is not equivalent to empirical refutation. The framework thus produces numerical certainty without empirical risk, allowing decisions to be closed administratively rather than provisionally tested through real-world replication.

What ultimately defines the SMC knowledge base is not methodological diversity but epistemic constraint. Only those claims compatible with the reference-case architecture are admissible. Single-attribute claims expressed on demonstrable measurement scales such as event reduction, time-to-event, or resource utilization are subordinated to composite value constructs. The result is a system optimized for consistency, speed, and administrative decisiveness rather than for the generation of evaluable, falsifiable knowledge.

In this sense, the SMC functions not as an assessor of measurement-valid therapy claims, but as a regulator of numerical storytelling. Its framework produces decisions, but it does not produce measurements. The logit diagnostic makes clear that this is not accidental. It is the predictable outcome of an evaluative architecture that privileges arithmetic closure over measurement discipline.

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: SCOTTISH MEDICINES CONSORTIUM

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 SCOTTISH MEDICINES CONSORTIUM

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	0.20	-1.40
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	+2.50
MEASUREMENT PRECEDES ARITHMETIC	1	0.10	-2.20
SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES	0	0.85	+1.75
MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC	1	0.10	-2.20
THERE ARE ONLY TWO CLASSES OF MEASUREMENT LINEAR RATIO AND RASCH LOGIT RATIO	1	0.05	-2.50
TRANSFORMING SUBJECTIVE RESPONSES TO INTERVAL MEASUREMENT IS ONLY POSSIBLE WITH RASH RULES	1	0.05	-2.50
SUMMATION OF LIKERT QUESTION SCORES CREATES A RATIO MEASURE	0	0.90	+2.20
THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE	0	0.85	+1.75
CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.15	-1.75

QALYS CAN BE AGGREGATED	0	0.95	+2.50
NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.65	+0.60
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.90	+2.20
THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.60	+0.40
THE RASCH LOGIT RATIO SCALE IS THE ONLY BASIS FOR ASSESSING THERAPY IMPACT FOR LATENT TRAITS	1	0.05	-2.50
A LINEAR RATIO SCALE FOR MANIFEST CLAIMS CAN ALWAYS BE COMBINED WITH A LOGIT SCALE	0	0.70	+0.85
THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT	1	0.20	+0.85
THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.05	-2.50

SCOTTISH MEDICINES CONSORTIUM: DEVOLUTION WITH THE ABSENCE OF MEASUREMENT

The SMC profile is not merely "similar" to the wider UK HTA posture; it is a concentrated re-enactment of the same governing inversion that made the UK template so exportable: arithmetic is treated as primary, while measurement is treated as optional commentary. The table does not describe a set of minor technical disagreements. It describes a settled epistemic constitution. The SMC corpus signals, with near-ceiling endorsement, that the principal objects required to keep the QALY system operational are to be treated as if they were already admissible measures. Simultaneously, it signals, at the floor. that the axioms that would determine whether those objects are measures are not merely absent, but functionally excluded.

Start with the gatekeeping rule that should define any quantitative evaluative enterprise: measurement precedes arithmetic. In the SMC profile, that foundational claim sits at $p = 0.10$ (-2.20). This is the signature of institutional permission: the system authorizes calculation before it authorizes meaning. Once that permission exists, everything that follows becomes administratively easy. You can multiply, aggregate, threshold, and rank. You can do all of this without ever confronting the prior question: "What are these numbers, exactly, and what operations are they allowed to support?" Table 1 shows that SMC's practical answer is: "Whatever operations the reference case requires."

The same inversion appears again, equally bluntly, in the companion statement that makes the gate explicit: meeting the axioms of representational measurement is required for arithmetic. It sits at p = 0.10 (−2.20). This is not an oversight. It is an institutional stance. If SMC treated representational measurement axioms as binding constraints, the central machinery of cost-utility analysis would be stopped at the door. QALYs would be reclassified as inadmissible composite artifacts. Utilities would be treated as ordinal outputs, not ratio-ready quantities. "Cost per QALY" would collapse as a meaningful ratio claim. The system cannot permit those consequences and still function as it currently functions. The near-floor endorsement is therefore not "ignorance"; it is self-protection.

The next item exposes the same structure with particular cruelty: multiplication requires a ratio measure. That statement is true, and it is the mathematical gatekeeper for the entire cost-utility enterprise. Yet in the SMC profile it sits at p = 0.15 (−1.75). That number is devastating because it means the SMC corpus operates as if the most basic condition for its flagship arithmetic operation does not matter. The QALY is literally built as time multiplied by a preference weight. Time is correctly recognized as a ratio measure at p = 0.95 (+2.50). The system knows what ratio measurement looks like when it is dealing with manifest quantities. Then it crosses the border into preferences and simply exempts itself from the same rule. This is selective discipline. It is not that the system "doesn't understand ratio"; it understands ratio perfectly well when ratio is convenient, and it suspends that understanding when ratio is inconvenient.

The enabling falsehoods then appear at the ceiling. The statement "the QALY is a ratio measure" is endorsed at p = 0.90 (+2.20). The statement "QALYs can be aggregated" is endorsed at p = 0.95 (+2.50). The statement that EQ-5D preference algorithms create interval measures is endorsed at p = 0.90 (+2.20). These are not decorative assumptions. These are load-bearing beams. If you remove them, the entire apparatus loses the right to perform its own arithmetic. Yet the SMC corpus does not debate them as measurement claims; it repeats them as administrative necessities. The table is therefore not simply a scorecard. It is a map of what the system must believe in order to maintain closure.

The "negative values" accommodation tells you how far that closure has been institutionalized. The claim "ratio measures can have negative values" is false, yet it is endorsed at p = 0.90 (+2.20). This is the point at which terminology becomes a mask. The system insists on calling utilities and QALYs "ratio" while simultaneously accepting states "worse than dead," meaning negative valuations. In a measurement-literate environment, the presence of negative values where a true zero is claimed would trigger immediate reclassification: you do not possess the scale properties you claim to possess. In the SMC environment, the contradiction is not a signal to stop; it is a signal to normalize. The contradiction becomes a professional habit.

The table also shows that SMC treats unidimensionality as expendable when it is inconvenient, while asserting it when it is required by the narrative. "Measures must be unidimensional" sits at p = 0.20 (−1.40). Yet "time trade-off preferences are unidimensional," a statement marked false here, is endorsed at p = 0.85 (+1.75). This is exactly the structural behavior of a memeplex: enforce constraints where they preserve the replicator; suspend constraints where they threaten the replicator. Unidimensionality is not treated as a property to be demonstrated; it is treated as a label to be assigned to whatever needs to function as a single continuum inside the model.

The most consequential part of the SMC profile is not, however, the QALY block. It is the Rasch block. The Rasch-related statements are not "low." They are at the floor. The claim that there are only two admissible classes of measurement a linear ratio for manifest attributes and Rasch logit ratio for latent traits sits at p = 0.05 (−2.50). The claim that transforming subjective responses to interval measurement is only possible with Rasch rules sits at p = 0.05 (−2.50). The claim that the Rasch logit ratio scale is the only basis for assessing therapy impact for latent traits sits at p = 0.05 (−2.50). The claim that Rasch rules are identical to representational measurement axioms sits at p = 0.05 (−2.50). These values do not mean "SMC never mentions Rasch." They mean something more decisive: the SMC evaluative culture does not treat Rasch as a compulsory measurement law for latent trait claims. It treats Rasch, if it appears at all, as optional ornamentation rather than as a gatekeeper. That is fatal, because without Rasch there is no lawful pathway from ordinal subjective responses to invariant measures.

This is why the journal-level ecosystem audits undertaken matters here. Bodies like SMC sit downstream from the instrument factories and the methods journals, yet they also reinforce what those upstream entities will continue to supply. If SMC demanded Rasch-based logit measures for any claim about latent traits then entire instrument families would be disqualified overnight. The supply chain would have to change. Instead, the SMC profile indicates that the supply chain is protected: summation is normalized, Rasch is marginalized, and the resulting pseudo-quantities are treated as legitimate inputs to cost-utility arithmetic.

That protection is visible in the endorsement of the summation propositions. "Summation of Likert question scores creates a ratio measure" (false) sits at p = 0.90 (+2.20). "Summations of subjective instrument responses are ratio measures" (false) sits at p = 0.85 (+1.75). Those are not minor errors. Those are the decisive permissions that allow an evaluative culture to treat scored responses as if they were measured quantities. Once that belief is in place, everything becomes possible: mapping, cross-walking, preference algorithms, "utilities," and then QALYs. SMC does not need to prove measurement because it has already accepted a substitute: the belief that scoring is measurement.

The table also contains a revealing "half-competence" zone, which is exactly what one expects in mature memeplex operation. The statement "non-falsifiable claims should be rejected" is endorsed at p = 0.65 (+0.60). That allows the system to perform virtue. It can speak the language of science: evidence, robustness, uncertainty, rejection of non-testable claims. Yet that virtue collapses immediately when the system turns to its operational core. "Reference case simulations generate falsifiable claims" (false) is endorsed at p = 0.90 (+2.20). That is the laundering step. A simulation is a conditional projection; it is not falsifiable in the Popperian sense unless it is bound to prospective protocols with defined endpoints that risk refutation. Reference case culture substitutes "sensitivity analysis" for falsification. It replaces empirical risk with internal model stability. The statement endorsement at +2.20 makes clear that the SMC corpus has adopted this substitution as normal practice.

This is also why "the QALY is dimensionally homogeneous" (false) sits at p = 0.85 (+1.75). Dimensional homogeneity is not a rhetorical flourish. It is the condition under which aggregation can even pretend to be meaningful. If you are multiplying a ratio time measure by an ordinal preference weight produced by an algorithm that permits negative values, then claiming homogeneity and ratio status is not merely wrong—it is the kind of wrong that exists to keep the

machine running. The SMC profile shows the machine is running, and the corpus is trained to declare the machine legitimate.

What, then, is the real significance of this logit structure? It is that SMC operates as a closure institution. It must deliver a decision environment where therapies can be priced, accepted, restricted, or rejected with the appearance of quantitative justification. The reference case is an administrative technology that offers exactly that: a standardized narrative format that yields a number, a threshold comparison, and an implied conclusion. To achieve closure, the system cannot permit the foundational question "Are the inputs measures?" to become governing. If it did, closure would become provisional and revisable over the product life span, because claims would be treated as falsifiable hypotheses rather than as model outputs. The SMC profile is therefore what closure looks like in numeric form: the rejection of measurement constraints and the endorsement of the arithmetic fictions that enable finality.

This is why the SMC profile should not be interpreted as merely "following NICE." It is following NICE in the only way that matters: by adopting the same measurement amnesty. Scotland does not merely inherit the UK HTA style; it inherits the UK HTA epistemic structure. It inherits the refusal to treat representational measurement axioms as binding constraints. It inherits the permission to treat utilities as if they were interval or ratio measures. It inherits the permission to aggregate QALYs. And it inherits the permission to treat reference case model outputs as if they were decision-relevant evidence rather than conditional projections.

If SMC wished to function as a genuine gatekeeper for therapy impact claims, the table would be reversed. Measurement-before-arithmetic would sit above neutrality, not at −2.20. Multiplication-requires-ratio would be strongly endorsed, not rejected. QALY ratio status and QALY aggregation would be rejected rather than celebrated. Summation-as-measurement would be rejected as a category error. Rasch would sit as a compulsory foundation for latent trait claims rather than as a quarantined curiosity. That reversal does not require "more data" or "better modeling." It requires a different constitution: the adoption of measurement as the prerequisite for calculation.

The brutal implication is that under the present SMC epistemic posture, the committee's numerical outputs cannot serve as cumulative scientific knowledge. They can be repeated, refined, scenario-tested, and sensitivity-analyzed, but they cannot be falsified in the strong sense because the dependent variables are not demonstrably measured quantities. Without measures, there are no invariant units; without invariant units, there is no stable object to reproduce; without reproducible objects, there is no evolution of objective knowledge; only the evolution of narrative technique. This is why the logit extremes matter. Probabilities can look modestly different across targets, but logits reveal where the system has placed its absolutes. Here the absolutes are clear: Rasch is at the floor; QALY arithmetic is at the ceiling; measurement-first is rejected.

Finally, SMC's role in the supply chain must be stated plainly. It is not a passive consumer of the memeplex. It is a reinforcer. What SMC accepts becomes what manufacturers continue to submit, what academic centers continue to teach, and what instrument developers continue to publish. When SMC treats pseudo-measurement as admissible, it creates demand for pseudo-measurement products. When it does not require Rasch-based possession measures for latent traits, it ensures

that scored instruments remain the currency of "patient-centered" evidence. When it treats reference case outputs as decision variables, it ensures that modeling replaces falsification as the standard of proof. In that sense, SMC is not merely part of the global numerical storytelling memeplex; it is one of its operational transmission nodes.

If the aim is to end numerical storytelling, the target is not only the most visible institution. The target is the evaluative culture that treats measurement as optional and arithmetic as destiny. The SMC table shows that culture in its canonical form. The conclusion is not that SMC needs "better methods." The conclusion is that SMC needs a different rulebook: measurement as gatekeeper, claims as single-attribute propositions, manifest outcomes restricted to linear ratio measures, latent outcomes restricted to Rasch logit ratio measures of possession, and protocols designed for reassessment rather than models designed for closure. Until that constitution changes, SMC's decisions may be administratively final, but they cannot claim scientific legitimacy in the sense required by representational measurement theory.

CAN THE SCOTTISH MEDICINES CONSORTIUM CONTINUE TO APPLY THE REFERENCE CASE FOR PRICING AND ACCESS

For more than two decades, the Scottish Medicines Consortium (SMC) has operated within an analytical framework closely aligned with the NICE reference case. Although institutionally independent, the SMC inherited the same evaluative architecture: preference-based utilities, QALYs, reference-case simulation models, and threshold-informed judgments about value for money. This framework has been treated as settled methodology, a technical apparatus that enables consistent pricing and access decisions across therapeutic areas. The question now is whether this apparatus can continue to function with any claim to scientific legitimacy.

The challenge confronting the SMC is no longer philosophical or academic. It is epistemic. The reference case depends on arithmetic operations whose validity rests entirely on the measurement properties of the variables involved. Yet those properties are absent. Utilities are not ratio measures. QALYs are not dimensionally homogeneous quantities. Composite health-state descriptions do not define unidimensional attributes. Simulation outputs are not falsifiable claims. These are not matters of interpretation; they are violations of representational measurement theory that have been established for more than half a century.

The central problem is structural. The reference case presumes that numbers become meaningful through modeling. Measurement theory requires the opposite: numbers are meaningful only if the attribute being represented possesses the structure required for arithmetic. In the SMC framework, this ordering is inverted. Arithmetic precedes measurement. Models are constructed first, and questions of scale type are never treated as gatekeeping conditions. Once this inversion is accepted, any numerical object can be treated as decision-relevant, provided it is generated consistently.

This inversion explains why the reference case has proven administratively attractive. It allows decisions to be produced under conditions of limited data. It offers apparent comparability across diseases. It provides closure. But administrative convenience is not scientific justification. Closure obtained by bypassing measurement is not evidence-based decision making; it is numerical governance.

The QALY lies at the center of this problem. In the SMC framework, quality-adjusted life-years are treated as if they represent a quantity that can be multiplied, aggregated, and compared across populations. Yet the utility component of the QALY is derived from ordinal responses to multiattribute health state descriptions. These responses lack equal intervals, lack invariance, and lack a true zero. Negative utilities are permitted while the construct is still described as ratio-scaled. Under the axioms of representational measurement, this arithmetic is disallowed. Multiplying time by an ordinal score does not create a measurable quantity. It creates a numerical artifact.

The SMC's continued use of such constructs therefore raises a fundamental question: on what basis can pricing and access decisions be defended as scientific judgments rather than administrative conventions? Sensitivity analysis does not resolve this problem. Scenario analysis does not resolve it. Internal model coherence does not resolve it. None of these addresses the prior requirement that the dependent variable be measurable.

Until recently, this contradiction could persist largely unchallenged because it was embedded across institutions. NICE used it. Academic journals normalized it. Consulting firms reproduced it. The SMC inherited it. There was no effective mechanism for exposing the belief system itself. That situation has changed.

AI large language model diagnostics now allow interrogation of institutional knowledge bases at scale. When applied to HTA systems, they reveal consistent patterns: near-total rejection of the axioms that govern measurement, combined with strong endorsement of arithmetic operations that depend on those axioms. This is not methodological diversity. It is a stable belief system, a memeplex, that protects itself by excluding the very principles that would invalidate it.

For the SMC, this development creates a governance dilemma. Continuing to apply the reference case now means doing so with full visibility of its epistemic failure. The defense that "this is how HTA is done" no longer suffices. The defense that "international alignment requires it" no longer holds when the alignment itself is demonstrably incoherent.

The issue is not whether the SMC must make difficult decisions about affordability and access. It must. The issue is whether those decisions can continue to be framed as the outputs of scientific evaluation when the underlying framework cannot produce evaluable claims. A system that cannot be falsified cannot learn. A system that cannot generate invariant quantities cannot accumulate objective knowledge. It can only repeat itself.

This is particularly problematic for a publicly accountable body operating within a national health system. The legitimacy of SMC decisions depends not only on procedural fairness but on epistemic integrity. If access restrictions are justified by reference to numbers that cannot, in principle, be measured, then the moral authority of those decisions becomes fragile. Patients are denied access not because a claim was empirically falsified, but because a model generated an unfavorable scenario.

A defensible future for the SMC therefore requires a transition away from the reference case as a decision tool. This does not imply abandoning economic thinking. It implies reordering it. Measurement must precede arithmetic. Claims must be explicit, unidimensional, and protocol-

driven. Manifest attributes, such as events avoided, time to hospitalization, or resource use, must be evaluated using linear ratio measures. Latent attributes such as symptom burden or functioning, must be measured using Rasch logit ratio scales capable of expressing possession invariantly.

Under such a framework, decisions are no longer anchored to composite indices or simulated thresholds. They are anchored to claims that can be tested, reproduced, and revised. Pricing discussions become negotiations informed by evidence rather than adjudications based on imaginary quantities. Access decisions become provisional rather than terminal.

The SMC is not trapped by its history. It is constrained only by the assumption that the reference case is indispensable. It is not. It was a solution to an administrative problem at a time when measurement literacy was low and computational models appeared to offer rigor by default. That era has passed.

The emergence of AI-based diagnostic tools marks a decisive shift. The invisibility that once protected the reference case is gone. The contradictions are now explicit. The emperor, as the phrase goes, has no clothes, not because critics say so, but because the belief system can now be interrogated directly.

The question facing the Scottish Medicines Consortium is therefore stark. It can continue to apply the reference case as an administrative ritual, accepting that its outputs lack scientific legitimacy. Or it can lead a transition toward a framework grounded in representational measurement, evaluable claims, and empirical accountability. Only one of these paths has a future consistent with science. The reference case delivered closure. Measurement delivers knowledge. The SMC must now decide which it intends to defend.

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