MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

SWITZERLAND: NATIONAL ACCEPTANCE OF FALSE MEASUREMENT IN HEALTH TECHNOLOGY ASSESSMENT

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

LOGIT WORKING PAPER No 237 FEBRUARY 2026

www.maimonresearch.com

Tucson AZ

FOREWORD

HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

Switzerland occupies a distinctive position within the global health technology assessment landscape. Unlike jurisdictions with a single dominant authority such as NICE in England or TLV in Sweden, Switzerland operates through a distributed federal structure anchored in the Federal Office of Public Health (Bundesamt für Gesundheit, BAG/FOPH), statutory reimbursement criteria, and a network of academic, policy, and consensus institutions. This decentralized architecture reflects Switzerland's broader political and administrative tradition, emphasizing procedural rigor, transparency, and formalized evaluation. The national HTA program, established within the BAG, provides systematic assessment of clinical effectiveness, appropriateness, and economic efficiency to inform reimbursement and coverage decisions under the compulsory health insurance system. These evaluations explicitly incorporate comparative effectiveness evidence and economic evaluation, often relying on composite utility measures and modeled projections of therapeutic impact.

Switzerland's reputation for precision, scientific integrity, and institutional discipline makes it a particularly important jurisdiction for evaluating the status of measurement within HTA. A system widely regarded as methodologically rigorous would be expected to enforce the foundational requirements of quantification, ensuring that arithmetic operations are restricted to constructs possessing admissible scale properties. The Swiss national knowledge base therefore provides an ideal test case for determining whether contemporary HTA practice operates within the axioms of representational measurement or whether it reproduces the same structural reliance on composite indices and simulated outputs observed internationally. The logit-based canonical statement assessment presented here examines whether Switzerland's evaluative framework embodies the measurement discipline required for empirically valid and evaluable claims regarding therapeutic value.

The objective of this study was to determine whether the Swiss national health technology assessment (HTA) knowledge base satisfies the axioms of representational measurement required to support lawful arithmetic operations in therapeutic evaluation. Switzerland is widely regarded as a jurisdiction characterized by procedural rigor, methodological discipline, and scientific precision. Its HTA framework, coordinated primarily through the Federal Office of Public Health (Bundesamt für Gesundheit, BAG) and supported by academic, policy, and consensus institutions, plays a central role in informing reimbursement decisions within the compulsory health insurance system. This study applies the 24-item canonical representational measurement diagnostic to the national Swiss HTA corpus, including methodological guidance, economic evaluation requirements, policy documentation, and academic literature. The objective is not to assess administrative competence or procedural transparency, but to determine whether the Swiss evaluative framework recognizes and enforces the necessary measurement preconditions for arithmetic, including unidimensionality, invariance, dimensional homogeneity, and the restriction

of multiplication and division to ratio-scale quantities. The study therefore addresses a foundational question: whether Switzerland's national HTA system operates within the measurement constraints required for empirically evaluable therapeutic value claims.

The logit profile demonstrates systematic exclusion of representational measurement axioms from the Swiss national HTA knowledge base. Core propositions asserting that measurement must precede arithmetic, that multiplication requires ratio measurement, that latent constructs require Rasch transformation to establish invariant interval structure, and that composite utility indices fail dimensional homogeneity requirements collapse to floor or near-floor logit values, indicating non-possession of these principles within the national evaluative framework. Conversely, false propositions asserting the legitimacy of QALYs as ratio measures, the admissibility of aggregating preference-weighted composite scores, and the use of simulation-derived cost-effectiveness ratios as evaluative evidence register positive logit values, demonstrating endorsement. These findings establish that the Swiss HTA knowledge base operationalizes composite utility constructs and arithmetic operations without enforcing the measurement constraints necessary for empirical validity. The resulting framework demonstrates administrative coherence and methodological sophistication but lacks the measurement foundations required for falsifiable, replicable, and empirically grounded therapeutic value claims.

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

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

The foundational theory that establishes *when* and *whether* a set of numbers can be interpreted as measurements came with the publication of Krantz, Luce, Suppes, and Tversky's *Foundations of Measurement* (1971) [2] . Representational Measurement Theory (RMT) formalized the axioms under which empirical attributes can be mapped to numbers in a way that preserves structure. Measurement, in this framework, is not an act of assigning numbers for convenience, it is the

discovery of a lawful relationship between empirical relations and numerical relations. The axioms of additive conjoint measurement, homogeneity, order, and invariance specify exactly when interval scales exist. RMT demonstrated once and for all that measurement is not optional and not a matter of taste: either the axioms hold and measurement is possible, or the axioms fail and measurement is impossible. Every major construct in HTA, utilities, QALYs, DALYs, ICERs, incremental ratios, preference weights, health-state indices, fails these axioms. They lack unidimensionality; they violate independence; they depend on aggregation of heterogeneous attributes; they collapse under the requirements of additive conjoint measurement. Yet HTA proceeded, decade after decade, without any engagement with these axioms, as if the field had collectively decided that measurement theory applied everywhere except in the evaluation of therapies.

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

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

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

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

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

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

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

Paul C Langley, Ph.D

Email: langleylapaloma@gmail.com

DISCLAIMER

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

1. INTERROGATING THE LARGE LANGUAGE MODEL

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

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

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

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

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

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

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

THE SWISS NATIONAL KNOWLEDGE BASE

The Swiss national HTA knowledge base is anchored in the statutory framework governing compulsory health insurance and coordinated primarily through the Federal Office of Public Health (BAG). Under Swiss law, reimbursed therapies must satisfy three formal criteria: effectiveness, appropriateness, and cost-effectiveness (Wirksamkeit, Zweckmässigkeit, Wirtschaftlichkeit, commonly abbreviated as WZW). These criteria establish the conceptual foundation for therapeutic evaluation and reimbursement decisions. The BAG oversees assessment processes, often supported by external academic groups, expert committees, and structured methodological frameworks designed to evaluate clinical evidence and economic impact.

Economic evaluation occupies a central position within this framework. Manufacturers seeking reimbursement or continued listing must submit comparative clinical and economic evidence demonstrating therapeutic benefit relative to existing alternatives. These submissions typically include cost-effectiveness analyses expressed in terms of incremental cost per quality-adjusted life year (QALY) or equivalent composite outcome measures. These analyses integrate clinical trial data, observational evidence, epidemiological projections, and modeled estimates of long-term health outcomes. Preference-based utility instruments, particularly multiattribute health status classification systems such as the EQ-5D, serve as the primary source of utility weights used to construct QALYs. These instruments assign numerical scores to multidimensional health state descriptions, allowing analysts to calculate composite indices representing overall health status.

Simulation modeling is widely employed to extrapolate clinical outcomes beyond observed trial periods. These models integrate transition probabilities, mortality projections, and utility weights to estimate lifetime health outcomes and associated costs. The resulting outputs generate cost-effectiveness ratios intended to inform reimbursement decisions and pricing negotiations. This modeling framework provides a structured approach to comparing therapies across disease areas and patient populations, creating a standardized quantitative architecture for evaluating therapeutic value.

The Swiss HTA knowledge base reflects extensive methodological diffusion from international HTA institutions, particularly NICE in the United Kingdom and similar European agencies. Methodological guidance emphasizes transparency, systematic evidence synthesis, and structured economic analysis. Academic health economics groups within Switzerland contribute to the development, application, and refinement of these evaluative methods. Peer-reviewed publications, policy documents, and technical guidelines reinforce the use of composite utility measures and modeled cost-effectiveness analysis as central evaluative tools.

However, the quantitative constructs embedded within this framework originate from preference-weighted composite indices that do not satisfy representational measurement axioms. Utility scores derived from multiattribute instruments represent aggregated ordinal preferences rather than invariant measures of a unidimensional attribute. Their numerical values depend on scoring

algorithms and valuation conventions rather than demonstration of measurement invariance. Despite this, these scores are treated as arithmetic quantities, multiplied by time and aggregated across individuals to construct QALYs. These composite measures are then combined with cost estimates to produce cost-effectiveness ratios that serve as decision variables.

The Swiss HTA knowledge base therefore represents a structured and administratively coherent evaluative system grounded in established international conventions. It demonstrates methodological consistency, procedural rigor, and institutional stability. However, the framework does not require demonstration that the quantitative constructs used in evaluation satisfy the axioms of representational measurement. Composite utility scores function as operational decision variables despite lacking invariant unit structure. As a result, the Swiss national HTA system embodies a quantitatively sophisticated but measurement-unverified framework in which arithmetic operations are performed on constructs whose admissible scale properties are not established.

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: SWITZERLAND NATIONAL

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.

SWITZERLAND: THE NATIONAL ABSENCE OF REPRESENTATIONAL MEASUREMENT

The Swiss national logit profile presents a familiar but still unsettling pattern: a system that is procedurally sophisticated, institutionally careful, and rhetorically committed to "evidence-based" decision-making, while simultaneously excluding, at the level of binding constraints, the axioms that make quantitative claims scientifically meaningful (Table 1). Switzerland's HTA environment is not built around a single NICE-style authority, but it does have a clearly articulated national HTA program housed within the Federal Office of Public Health (FOPH/BAG), alongside a long-running ecosystem of policy reports, academic outputs, and methodological consensus efforts. The BAG's HTA program explicitly frames its work as systematic evaluation that feeds into reimbursement decisions. Yet the canonical diagnostic shows that the "measurement discipline" required for evaluable claims is not what governs the arithmetic at the core of Swiss HTA. What governs it is the inherited administrative convenience of composite utility scoring and cost-per-QALY modelling; an architecture that produces outputs that look like quantities, behave like decision thresholds, and are treated as though they are empirically testable, even when their scale properties are never demonstrated.

TABLE 1: ITEM STATEMENT, RESPONSE, ENDORSEMENT AND NORMALIZED LOGITS SWITZERLAND NATIONAL

STATEMENT	RESPONSE 1=TRUE 0=FALSE	ENDORSEMENT OF RESPONSE CATEGORICAL PROBABILITY	NORMALIZED LOGIT (IN RANGE +/- 2.50)
INTERVAL MEASURES LACK A TRUE ZERO	1	0.20	-1.40
MEASURES MUST BE UNIDIMENSIONAL	1	0.15	-1.75
MULTIPLICATION REQUIRES A RATIO MEASURE	1	0.10	-2.20
TIME TRADE-OFF PREFERENCES ARE UNIDIMENSIONAL	0	0.85	+1.75
RATIO MEASURES CAN HAVE NEGATIVE VALUES	0	0.90	+2.20
EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES	0	0.85	+1.75
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	0.05	-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	0.65	+0.50
CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.15	-1.75
QALYS CAN BE AGGREGATED	0	0.90	+2.20

NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.55	+0.52
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.85	+1.75
THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.65	+0.50
THE RASCH LOGIT RATIO SCALE IS THE ONLY BASIS FOR ASSESSING THERAPY IMPACT FOR LATENT TRAITS	1	0.05	-2.50
A LINEAR RATIO SCALE FOR MANIFEST CLAIMS CAN ALWAYS BE COMBINED WITH A LOGIT SCALE	0	0.60	+0.40
THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT	1	0.25	-1.87
THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.05	-2.50

Start with the basic asymmetry that runs through the entire table. Time is correctly recognized as a ratio measure (+2.50). This is important, because it shows that the Swiss knowledge base is not numerically illiterate in general; it can recognize a manifest extensive quantity with a true zero and invariant units. The system is capable of lawful arithmetic when the object is genuinely measurable. But that competence does not carry over into the objects that dominate HTA decision-making: "quality," "utility," "health-related quality of life," and their aggregation into QALYs. Here, the floor collapses appear precisely where they must if the system is built to keep cost-per-QALY modelling intact. "Measurement precedes arithmetic" sits at −2.20. "Multiplication requires a ratio measure" sits at −2.20. "Meeting the axioms of representational measurement is required for arithmetic" sits at the absolute floor (−2.50). These are not optional philosophical claims; they are the operating rules for when numbers can represent empirical attributes and when they cannot. Their low endorsement indicates that they do not function as gatekeeping conditions in Swiss HTA. They may be known somewhere in the wider scientific world, but within the HTA knowledge base they do not operate as constraints on what is admissible.

The same structural logic explains the Rasch-related collapses. "Transforming subjective responses to interval measurement is only possible with Rasch rules" is at −2.50. "There are only two classes of measurement linear ratio and Rasch logit ratio" is at −2.50. "The Rasch logit ratio scale is the only basis for assessing therapy impact for latent traits" is at −2.50. "Rasch rules are identical to the axioms of representational measurement" is at −2.50. These are not marginal technicalities. If HTA insists on using latent constructs to support comparative claims, then the question is not whether a preference algorithm exists, but whether the resulting numbers have invariant unit structure. Invariance is not a rhetorical flourish; it is the condition for replication and the possibility of learning. Without it, comparisons across persons, across time, across studies, and across therapies are structurally unstable. The Swiss national environment—like others you have

assessed—excludes that entire measurement pathway, which effectively means it cannot truthfully claim to "measure" latent traits. It can only score them.

Once that is recognized, the positive logit cluster becomes diagnostic rather than merely descriptive. Switzerland's knowledge base strongly endorses the propositions that keep the QALY machinery operational: that EQ-5D preference algorithms create interval measures (+1.75), that QALYs can be aggregated (+2.20), that summated ordinal scores can be treated as ratio measures (+1.75), that Likert summations create ratio measures (+1.75), that the QALY is dimensionally homogeneous (+1.75), and most strikingly that the QALY is a ratio measure (+2.20). These are not innocent misunderstandings; they are the specific false beliefs required for the reference-case style arithmetic to proceed without embarrassment. If the system conceded that utility scores are ordinal (or at best undeclared), the multiplication by time would immediately be in question. If it conceded that composite multiattribute indices are not unidimensional measures, aggregation would be in question. If it conceded that negative values contradict ratio interpretation, threshold comparisons and ratios would become indefensible. Yet the Swiss knowledge base endorses the entire package of false measurement claims. Table 1 is therefore not reporting random confusion. It is reporting a stabilized belief system: the set of propositions that must be treated as "true enough" for the system to keep functioning.

This is where Switzerland's institutional context matters. The BAG's HTA program is not a theoretical seminar; it is explicitly tied to reimbursement decision processes and structured HTA reporting. The Swiss system also embeds statutory evaluation criteria, notably the EAE criteria (effectiveness, appropriateness, and economic efficiency) in health insurance law, which are operationalized through official documentation. "Economic efficiency" is precisely where the impossible QALY construct becomes attractive as an administrative shortcut: it promises a single index, comparable across diseases, that can be multiplied, divided, thresholds applied, and used to claim commensurability. It is the dream of closure: generate a number, compare it to a boundary, and end the discussion. Switzerland's decentralization does not remove that impulse; it redistributes it across institutions, reports, and consensus documents. SwissHTA, for example, has explicitly discussed evaluation processes, cost-effectiveness concepts, and even threshold-style reasoning in its materials, reflecting the same international methodological inheritance.

The national logit profile also clarifies why "non-falsifiability" occupies an ambiguous middle space (+0.52 rather than a floor collapse). Switzerland, like many jurisdictions, can easily endorse the rhetoric that non-falsifiable claims should be rejected; it is a safe slogan, consistent with a general "evidence-based" posture. The problem is that the system does not operationalize falsifiability where it matters most: in the central quantitative objects that decide access and pricing. When the table shows high endorsement for the proposition that reference case simulations generate falsifiable claims (+1.75), it is recording the institutional substitution of model coherence for empirical risk. This is not a Swiss peculiarity; it is the reference-case logic everywhere. But Switzerland's HTA program and the broader Swiss evaluation environment provide ample evidence that modelling, cost-effectiveness, and QALY-style outcomes are in active use in Swiss health economic analyses and HTA reports. A model output can be recalculated indefinitely, but recalculation is not falsification. If the objects being multiplied and aggregated are not measures, there is no empirical quantity to be wrong about; only a set of conventions to be varied.

A Swiss-specific nuance strengthens, rather than weakens, the critique: the well-known absence (historically) of a Swiss EQ-5D value set forced analysts to borrow foreign value sets, making explicit that the "utility" is not a measured attribute of Swiss patients but a convention imported from elsewhere. That should have served as a warning signal: if the scoring algorithm can be swapped across countries and still treated as the same "quantity," then what exactly is being measured? In measurement science, the unit structure and the admissible transformations are not optional; they define the meaning of the numbers. Yet within the HTA memeplex, portability of scoring is treated as practicality rather than as evidence of non-measurement.

The statement "measures must be unidimensional" at −1.75 and the rejection of "TTO preferences are unidimensional" at +1.75 show the deeper epistemic inversion at work. The system is not merely ignoring unidimensionality; it is normalizing multidimensional preference constructions as if they were measurement. Multiattribute instruments are built precisely by decomposing health into dimensions and then recombining them via preference weights. That architecture makes unidimensionality structurally impossible at the level of the final index. If the system were to enforce unidimensionality, it would have to abandon the multiattribute index. The logit profile says it does not enforce it. That is why the critique must be framed as structural and inevitable: it is not that Switzerland has made a few correctable mistakes; it is that the core architecture depends on violating the axioms that would otherwise constrain it.

Now consider the cost side, because Switzerland's national context invites a particularly sharp clarification. It is tempting for defenders of cost-per-QALY arithmetic to say "cost is money, money is ratio." But in HTA practice, "cost" is typically a composite constructed from heterogeneous resource items, each with context-dependent prices, then bundled into a single currency sum. Currency units may have a true zero in accounting terms, but their invariance as a measure of resource consumption is not guaranteed across settings, time, contracts, or payment systems. If the goal is measurement-valid manifest claims, the correct approach is to use single attribute resource units (hospital days, ED visits, physician visits, lab tests, procedures) as ratio measures, then allow the health system to attach local prices. Switzerland's complex, decentralized health financing environment makes this point more, not less, important. If the numeric inputs are context-sensitive composites, then the arithmetic is not discovering a stable empirical quantity; it is assembling an administrative artifact. That matters because the HTA output is then treated as if it can legitimately drive access decisions.

All of this converges on duty of care, which is where the Swiss profile should be read without sentimentality. HTA is not a scholastic exercise. It is an evaluative system that influences therapy availability, reimbursement, and ultimately patient access. Switzerland's national HTA program exists explicitly to support policy advice and reimbursement decision processes. If the quantitative outputs are built on non-measures, then the system is not merely "methodologically contestable." It is operating without scientific accountability. The more refined the procedures become, the more damaging this becomes, because procedural sophistication amplifies the authority of the numbers. The Swiss environment can produce beautifully documented HTA reports, systematic reviews, and modelled cost-effectiveness outputs, but if the central outcome is a composite preference index masquerading as a ratio measure, then the system's numeric authority outstrips its epistemic warrant. That is exactly how a memeplex operates: it stabilizes its core beliefs, rewards conformity, and treats internal coherence as legitimacy.

Switzerland has a reputation for caution, precision, and scientific seriousness. Demonstrating the same floor collapses here as in all other countries evaluated with this LLM diagnostic shows that the failure is not a parochial "training deficit" but a structural feature of the pseudoscientific international HTA template. The national HTA unit established within BAG/FOPH as part of intensified HTA efforts reinforces that this is not an accidental peripheral activity; it is institutionally embedded. The Swiss HTA consensus materials reinforce that the Swiss environment participates in the same methodological language of meaningless cost-effectiveness and thresholds. In short: Switzerland is not outside the memeplex; it is one of its disciplined hosts.

What does the Swiss national logit profile ultimately say? It says that the Swiss HTA knowledge base does not treat representational measurement as the prior condition for arithmetic. It treats arithmetic as the prior condition for decision-making, and it treats measurement as optional rhetoric. The system recognizes but does not understand ratio measurement when the object is manifest and uncontroversial (time), but it suspends ratio discipline the moment the object becomes a preference-weighted composite index. It endorses precisely the false statements needed to preserve the QALY, the multiattribute instrument, and the reference-case model as administratively decisive yet mathematically false constructs. It endorses simulation outputs as if they were falsifiable claims, while excluding the measurement axioms that would make falsifiability possible. It treats Rasch not merely as absent, but as irrelevant despite Rasch being the only lawful route to invariant interval scaling for latent constructs. The consequence is unavoidable: Switzerland's national HTA environment produces numbers that look scientific, travel easily through reports and deliberations, and support closure, but do not meet the conditions required for empirical evaluability.

The long-term for Switzerland is non-negotiable from the perspective of normal science. Switzerland can keep a distributed HTA system or build a centralized one; that is secondary. The primary issue is whether any HTA system will accept the first rule of quantification: measurement must come before arithmetic. Until the Swiss HTA knowledge base treats unidimensional ratio measurement (for manifest claims) and Rasch logit ratio measurement (for latent traits) as binding constraints, the national program cannot claim to support the evolution of objective knowledge. It can only support the evolution of procedures. And procedural evolution without measurement correction is exactly the condition under which everything changes, but nothing changes.

ABANDONING DUTY OF CARE AND THE EVOLUTION OF OBJECTIVE KNOWLEDGE

Health technology assessment presents itself as a scientific enterprise. It claims to evaluate therapies quantitatively, compare their effectiveness, and guide rational resource allocation in the interests of patients and the health system. These claims carry an implicit obligation: that the numerical constructs used in evaluation must correspond to real, measurable attributes. This obligation is not merely methodological. It is ethical. It defines the duty of care owed by agencies, journals, and evaluators to patients and physicians. When numerical claims influence therapy access, reimbursement, and clinical practice, their measurement validity becomes inseparable from institutional responsibility. Abandoning measurement validity is therefore not a neutral methodological choice. It is an abandonment of duty of care.

The concept of duty of care presupposes that decisions affecting patients are grounded in reliable knowledge. In clinical medicine, this principle is obvious. A physician cannot prescribe treatment based on arbitrary numbers or invalid diagnostic tests. The same principle applies at the institutional level. Agencies and journals that evaluate therapies and recommend reimbursement decisions must ensure that the quantities they manipulate represent real attributes. If they perform arithmetic on constructs that lack measurement validity, the resulting conclusions do not constitute scientific knowledge. They constitute administrative artifacts. The distinction is decisive. Scientific knowledge describes reality. Administrative constructs describe procedures.

The axioms of representational measurement define the boundary between these two domains. They specify the conditions under which numbers legitimately represent empirical attributes. Arithmetic operations such as addition, multiplication, and ratio comparison are admissible only when scale properties support those operations. Ratio scales require a true zero and invariant units. Interval scales support addition and subtraction but not multiplication. Ordinal scales support rank ordering but not arithmetic. These principles are not optional conventions. They are logical conditions that must be satisfied before numbers can be treated as quantities. Ignoring them does not create new forms of measurement. It creates numbers without empirical meaning.

The evolution of objective knowledge depends on adherence to these principles. Scientific progress occurs through measurement, falsification, and replication. Measurement establishes quantities that can be compared across observations. Falsification allows empirical claims to be tested and rejected if they do not correspond to reality. Replication ensures that findings are stable and independent of specific observers or methods. Without measurement, falsification is impossible. Without falsification, knowledge cannot evolve. Numerical claims may proliferate, but they cannot be empirically evaluated. They remain insulated from correction.

Health technology assessment, as currently practiced, disrupts this process. Composite constructs such as the quality-adjusted life year are treated as if they were ratio measures, despite lacking demonstrated ratio scale properties. Preference-weighted utility scores derived from ordinal responses are multiplied by time, aggregated across individuals, and used to generate cost-effectiveness ratios. These operations assume measurement validity without demonstrating it. The resulting numbers possess administrative authority but lack empirical grounding. They cannot be falsified because they do not correspond to measured quantities. They can be recalculated, but recalculation is not empirical testing. It is internal manipulation within a scoring system.

This substitution of scoring for measurement has profound implications for duty of care. Agencies use cost-effectiveness ratios to determine whether therapies are reimbursed. Journals publish analyses that influence clinical and policy decision making. These decisions affect patient access to treatment, physician prescribing options, and the allocation of finite health system resources. If the quantitative constructs underlying these decisions lack measurement validity, the decision framework becomes detached from empirical reality. Numerical outputs may appear precise, but their precision reflects internal consistency rather than empirical truth.

The ethical consequence is unavoidable. Institutions exercising evaluative authority have a duty to ensure that their quantitative claims satisfy the conditions required for measurement. Failure to do so transforms evaluation into administrative computation rather than scientific assessment. This is

not a minor technical oversight. It represents a structural abandonment of the responsibility to base decisions on valid empirical knowledge.

The abandonment is reinforced by institutional stabilization. Once measurement-invalid constructs become embedded in guidelines, curricula, and publication standards, they acquire legitimacy through repetition. Analysts learn to manipulate composite indices without questioning their measurement properties. Journals publish studies using established frameworks. Agencies require submissions to conform to existing evaluative templates. The system reproduces itself. Its numerical outputs create the appearance of scientific rigor, even though the underlying constructs do not satisfy representational measurement axioms. This appearance delays recognition of the foundational problem.

The effect on the evolution of objective knowledge is corrosive. Instead of replacing invalid constructs with measurement-valid alternatives, the system refines procedures for manipulating invalid constructs. Models become more complex. Simulation techniques become more sophisticated. Statistical methods become more elaborate. Yet the measurement foundation remains unchanged. The apparatus evolves administratively while remaining epistemically static. This is not scientific progress. It is procedural elaboration of an unchanged conceptual error.

Recovery requires reestablishing measurement as the prerequisite for arithmetic. Manifest attributes must be evaluated using linear ratio scales that preserve dimensional homogeneity. Latent constructs must be measured using invariant transformation models such as the Rasch model, which establish interval structure grounded in empirical invariance. Only when measurement validity is established can arithmetic operations produce meaningful quantitative claims.

The obligation to make this transition is not merely methodological. It is ethical. Agencies, journals, and evaluators exercise authority over decisions that affect human health and well-being. That authority carries a duty of care grounded in the integrity of the knowledge on which decisions rely. Abandoning representational measurement abandons that duty. It replaces empirical evaluation with numerical ritual. It halts the evolution of objective knowledge and substitutes administrative coherence for scientific validity.

The choice is therefore stark. Health technology assessment can continue to operate within a closed system of measurement-invalid constructs, preserving institutional continuity at the expense of scientific integrity. Or it can reestablish the measurement foundations necessary for empirical evaluation, falsification, and cumulative knowledge development. Only the latter path fulfills the duty of care owed to patients, physicians, and the health system.

III. THE TRANSITION TO MEASUREMENT IN HEALTH TECHNOLOGY ASSESSMENT

THE IMPERATIVE OF CHANGE

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

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

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

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

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

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

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

MEANINGFUL THERAPY IMPACT CLAIMS

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

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

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

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

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

THE PATH TO MEANINGFUL MEASUREMENT

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

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

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

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

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

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

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

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

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

TRANSITION REQUIRES TRAINING

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

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

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

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

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

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

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

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

A NEW START IN MEASUREMENT FOR HEALTH TECHNOLOGY ASSESSMENT

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

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

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

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

DESIGNED FOR CLOSURE

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

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

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

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

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

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

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

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

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

ACKNOWLEDGEMENT

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

REFERENCES

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

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

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

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