MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

NORWAY: NATIONAL DENIAL OF REPRESENTATIONAL MEASUREMENT

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

LOGIT WORKING PAPER No 232 FEBRUARY 2026

www.maimonresearch.com

Tucson AZ

FOREWORD

HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

Norway is often portrayed as distinct within the European HTA landscape. Its “managed introduction” system was designed explicitly to control the adoption of new technologies through a centralized, transparent, and nationally coordinated process. The Norwegian framework integrates prioritization criteria including benefit, resource use, and severity within an explicit political and ethical discourse about fairness and sustainability. Unlike jurisdictions where HTA evolved incrementally or through fragmented institutional pathways, Norway’s model was constructed deliberately as a coherent national architecture. This reputation for structured governance and principled priority setting raises a natural question: might Norway be better positioned than other countries to embed the standards of normal science in its quantitative claims? If any national system were to align cost-effectiveness reasoning with the axioms of representational measurement insisting on unidimensionality, invariance, and lawful arithmetic it could plausibly be one that already emphasizes explicit criteria and methodological discipline. Norway therefore provides an important test case: not whether it follows international HTA conventions, but whether it transcends them by anchoring its evaluative framework in defensible measurement.

The objective of this study is to apply the 24-item canonical representational measurement diagnostic to the Norwegian national HTA knowledge base in order to determine whether the axioms of normal science function as binding constraints within its evaluative architecture. The analysis does not assess procedural transparency, institutional efficiency, or political legitimacy. Instead, it interrogates the scale properties of the quantitative constructs embedded in Norwegian HTA discourse, guidance, and decision logic. Specifically, the study asks whether unidimensionality, invariance, admissible transformations, and the requirement that arithmetic follow lawful measurement are demonstrably present in the national Norwegian knowledge base. By translating categorical endorsement of each canonical statement into normalized logits, the study generates a structured epistemic profile that reveals whether Norway’s cost-effectiveness reasoning rests on ratio-valid measurement or on composite arithmetic whose scale properties are assumed rather than established.

The findings indicate systematic non-possession of representational measurement axioms within the Norwegian national HTA framework. Statements asserting that measurement must precede arithmetic, that multiplication requires ratio-scale properties, and that Rasch transformation is necessary to convert ordinal observations into interval measures register strongly negative logit values, including repeated collapses to the −2.50 floor. In contrast, statements endorsing the QALY as a ratio measure, permitting aggregation of composite utility scores, and treating reference case simulations as generating falsifiable claims cluster strongly positive. The resulting logit profile demonstrates structural alignment with the international cost-per-QALY paradigm and exclusion of scale-type discipline as a binding constraint. Norway’s HTA knowledge base is coherent within global conventions but does not satisfy the axioms required for arithmetic legitimacy under representational measurement theory.

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 NATIONAL NORWEGIAN KNOWLEDGE BASE

Norway’s HTA architecture is widely regarded as systematic and centralized. The “New Methods” (Nye metoder) system was introduced to ensure that new health technologies entering the specialist health service are evaluated through a uniform national process before adoption. This framework integrates assessment, appraisal, and decision-making across the regional health authorities, with defined roles for evidence synthesis and economic evaluation. Norway’s prioritization criteria, benefit, resource use, and severity are explicitly articulated and embedded in national policy discourse. The system is often described as transparent, consistent, and ethically grounded.

Economic evaluation plays a central role within this architecture. Cost-utility analysis using QALYs functions as the dominant quantitative instrument in many assessments. Health outcomes are typically expressed through multiattribute instruments such as EQ-5D, with preference weights derived from population surveys. These weights are multiplied by time to produce QALYs, which serve as the denominator in incremental cost-effectiveness ratios. Costs are aggregated in monetary terms, combining heterogeneous resource inputs into composite totals. The resulting cost-per-QALY ratios are compared against implicit or explicit thresholds within the decision-making framework.

The Norwegian knowledge base reflects strong procedural discipline. Assessment reports document assumptions, modeling structures, data sources, and uncertainty analyses. Sensitivity analyses explore parameter variation. Scenario analyses are conducted to assess robustness. This procedural rigor contributes to Norway’s reputation for methodological seriousness. However, the canonical assessment reveals that explicit adherence to representational measurement constraints is not embedded within this architecture. Unidimensionality is not enforced as a prerequisite for aggregating multiple health domains. The distinction between ordinal and interval scales does not function as a limiting condition on arithmetic operations. The requirement of a true zero for ratio multiplication is not operationalized in practice. Rasch modeling, which provides the only lawful mechanism for transforming ordinal latent trait observations into invariant interval logit measures, does not appear as a structural requirement for latent construct claims.

As a result, Norway’s HTA knowledge base treats preference-weighted composite indices as if they possessed ratio properties. Utility scores derived from multiattribute instruments are multiplied and aggregated, and the resulting QALYs are treated as homogeneous quantities suitable for threshold comparison. Monetary costs, themselves aggregates of heterogeneous resource units valued at context-dependent prices, are combined with these composite denominators to form cost-effectiveness ratios. Simulation models project outcomes over extended horizons, incorporating extrapolated survival curves and assumed utility trajectories. These projections are recalibrated through sensitivity testing but are not framed in terms of falsifiable measurement claims.

Norway’s HTA framework is therefore procedurally coherent and institutionally disciplined, yet the canonical logit profile demonstrates that representational measurement axioms do not operate as binding principles within its knowledge base. The system exemplifies structured implementation of the international cost-per-QALY paradigm rather than adherence to the measurement constraints required for arithmetic legitimacy in normal science.

CATEGORICAL PROBABILITIES

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

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

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

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

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

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

The LLM computes these categorical probabilities from three sources:

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

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

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

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

INTERROGATION STATEMENTS

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

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

Measurement Theory & Scale Properties

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

Measurement Preconditions for Arithmetic

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

Rasch Measurement & Latent Traits

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

Properties of QALYs & Utilities

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

Falsifiability & Scientific Standards

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

Logit Fundamentals

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

Latent Trait Theory

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

AI LARGE LANGUAGE MODEL STATEMENTS: TRUE OR FALSE

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

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

INTERPRETING TRUE STATEMENTS

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

- measurement preceding arithmetic
- unidimensionality
- scale-type distinctions
- dimensional homogeneity

- impossibility of ratio multiplication on non-ratio scales
- the Rasch requirement for latent-trait measurement

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

INTERPRETING FALSE STATEMENTS

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

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

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

2. SUMMARY OF FINDINGS FOR TRUE AND FALSE ENDORSEMENTS: NORWAY

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.

NORWAY: NEW PROCESS FOR MEASUREMENT FAILURE

Norway is often described as a "managed introduction" exemplar: a country that attempted to systematize adoption decisions for new technologies through a centralized national framework, with defined processes, role separation, and consistent decision pathways. The "New Methods" system (*Nye metoder*) was introduced to provide a uniform national process for assessment and decision-making on public funding for new health technologies in the specialist health service. This reputation rests on process design: commissioned assessments, structured evidence synthesis, formal appraisal, and a decision forum representing the regional health authorities. It is also reinforced by the visibility of priority-setting criteria—benefit, resources, and severity—embedded in reimbursement decision-making. The logit profile describes the commitment to measurement failure (Table 1).

TABLE 1: ITEM STATEMENT, RESPONSE, ENDORSEMENT AND NORMALIZED LOGITS   NORWAY

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	0.90	+2.20
THE QALY IS A RATIO MEASURE	0	0.95	+2.50
TIME IS A RATIO MEASURE	1	0.95	+2.50
MEASUREMENT PRECEDES ARITHMETIC	1	0.10	-2.20
SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES	0	0.90	+2.20
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	0.95	+2.50
THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE	0	0.90	+2.20
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	0.60	+0.40
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	0.40	-0.45
THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT	1	0.25	-1.10
THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.05	-2.50

The question posed by the canonical 24-item diagnostic is not whether Norway has a coherent process. It is whether Norway’s HTA knowledge base is anchored in the axioms that make quantification possible. In normal science, numbers represent attributes only under strict conditions. Those conditions are specified by representational measurement theory: unidimensionality, invariance, admissible transformations, and scale-type constraints that determine when arithmetic is meaningful. If those axioms do not operate as binding constraints, then HTA may be procedurally impressive while remaining epistemically fragile. The logit profile for Norway demonstrates exactly this configuration: high procedural seriousness combined with systematic non-possession of measurement discipline.

The first pattern is the collapse of foundational measurement propositions into strongly negative logits. Measurement precedes arithmetic sits at −2.20. Multiplication requires a ratio measure also sits at −2.20. Meeting the axioms of representational measurement is required for arithmetic sits at −2.20. These are not methodological preferences. They are gatekeeping rules. They specify the conditions under which multiplication and division are admissible. If the knowledge base does not possess them, then arithmetic proceeds without prior demonstration that the constructs involved can bear that arithmetic.

The Rasch-related propositions collapse to the floor. That there are only two classes of measurement linear ratio for manifest claims and Rasch logit ratio for latent traits, registers at −2.50. Transforming subjective responses to interval measurement is only possible with Rasch rules registers at −2.50. Rasch rules are identical to the axioms of representational measurement registers at −2.50. These floor values signify effective non-possession: these propositions do not function as binding principles in the Norwegian HTA knowledge base. Even if Rasch is occasionally mentioned in scattered contexts, it does not operate as a required transformation architecture for latent trait claims. In consequence, latent constructs remain scored rather than measured, and the system proceeds as if scoring and measurement were interchangeable.

Against this, the positive cluster is unmistakable. The QALY is treated as a ratio measure at +2.50. QALYs can be aggregated at +2.50. Summation of Likert scores creates a ratio measure at +2.50. EQ-5D preference algorithms are treated as creating interval measures at +2.20. Reference case simulations are treated as generating falsifiable claims at +2.20. These endorsements define the operational architecture of the Norwegian HTA memeplex: multiattribute preference scoring is elevated to the status of measurement; composite arithmetic becomes legitimate by repetition; simulations become evidence by convention.

Norway’s published guidance reinforces the institutional context in which these commitments operate. The Norwegian Medical Products Agency (DMP; formerly NOMA in many English-language references) describes itself as responsible for HTAs of pharmaceuticals, vaccines, and medical devices, and situates HTA explicitly within the problem of prioritization under resource constraints. Its submission guidance for medicinal products emphasizes quantitative or qualitative assessment of prioritization criteria, including cost-utility analysis as an explicit tool in relevant cases. The broader “New Methods” system description and technical guidance make clear that the framework is designed to channel technologies into assessment pathways that may include economic evaluation before a decision forum determines adoption. Norway is therefore not marginally exposed to the QALY paradigm; it is structurally organized around it.

The logit profile helps explain what this means in practice. Consider the core cost-per-QALY ratio. Time is correctly identified as a ratio measure (+2.50). But the utility weight multiplied by time is not demonstrably ratio. Multiattribute instruments represent bundles of attributes: mobility, pain, anxiety, self-care, usual activities. They elicit preferences over health states and map those preferences to a bounded index. This is preference aggregation, not measurement transformation. Anchoring “dead = 0” is a convention; it does not establish a representational true zero corresponding to the absence of the attribute being measured. The allowance of negative utilities signals that the index is not ratio. Yet the knowledge base endorses the proposition “the QALY is a ratio measure” at the ceiling.

This is not a minor technical quibble. Multiplication requires ratio properties in both components. The knowledge base rejects that requirement (−2.20). This single logit alignment is enough to expose the architecture: the system performs multiplication because multiplication is needed to create the QALY, not because the construct satisfies the axioms that permit multiplication. Arithmetic is driving measurement claims, rather than measurement justifying arithmetic.

Now consider the numerator of cost-effectiveness. Monetary units are ratio in the narrow sense (currency units have a true zero). But HTA “cost” is typically not a single manifest attribute; it is an aggregate of heterogeneous resource inputs valued at context-dependent prices. Prices are not invariant; they vary by contract, setting, time, and accounting conventions. If the system’s aim were to maintain ratio measurement discipline, it would prioritize resource units as the outcome: hospital days, ICU days, specialist visits, procedure counts, drug doses—each a potentially unidimensional ratio measure if defined cleanly. Aggregated monetary “cost” is an accounting composite. That composite is then divided by a composite preference index to yield a ratio that lacks dimensional homogeneity. The knowledge base, however, treats these ratios as legitimate objects of threshold comparison and national decision-making.

This is where Norway’s reputation becomes revealing. Norway is often praised not because it rejects the global HTA template, but because it implements it consistently. The “New Methods” framework is a mechanism for disciplined adoption control. Norway’s prioritization criteria of benefit, resources, severity are explicitly integrated into decision-making. But explicit criteria do not solve measurement. They presuppose that “benefit” is being quantified in a manner that is commensurable, invariant, and lawful for arithmetic. If benefit is operationalized as QALYs constructed from multiattribute preference indices, the criteria are applied to a construct that does not meet representational requirements. The fairness narrative can be sincere while the arithmetic remains illegitimate.

The duty-of-care implications follow directly. A national HTA system does not merely describe evidence; it constrains clinical options. It shapes what physicians can prescribe within reimbursed pathways and what patients can access without catastrophic out-of-pocket burden. If the decisive quantitative criterion is built on constructs that are not measured in the normal-science sense, the system is exposed to a specific type of harm: decisions that are presented as quantitatively justified when the underlying quantities are not established as admissible objects of multiplication, division, and aggregation. This is not simply “uncertainty.” It is category error. The harm is epistemic before it is economic, and it becomes clinical when access is denied, delayed, or rationed.

Norway’s moderate positive logit for rejecting non-falsifiable claims (+0.40) looks like an allegiance to scientific norms. But the strong positive endorsement of “reference case simulations generate falsifiable claims” (+2.20) undermines that allegiance. Simulation models can be iterated; they can be recalibrated; they can be sensitivity-tested. But where the constructs are composite and the scale properties assumed, simulations do not yield falsifiable claims in the Popperian sense. They yield conditional projections. The knowledge base stabilizes this as evidence, thereby substituting procedural robustness checks for empirical risk of refutation. In a mature HTA system, this substitution becomes invisible: it is treated as the natural way evaluation is done.

This brings us back to the deeper question you’ve pressed across jurisdictions: is any country “better placed” than others to implement representational measurement? Institutionally, Norway has features that suggest it could. The managed introduction framework is centralized, systematic, and capable of enforcing methodological requirements upstream of adoption. If Norway chose to require portfolios of single-attribute claims with manifest linear ratio measures for clinical and resource outcomes, and Rasch logit ratio measures for latent traits it has the governance capacity to impose that architecture. It also has an explicit prioritization discourse, which could naturally incorporate an epistemic duty of care: the obligation to use only measures that satisfy admissible arithmetic.

Epistemically, however, the logit profile suggests Norway is not better placed. Norway’s HTA reputation is for disciplined implementation of the international paradigm, not for measurement innovation. The national knowledge base endorses the QALY arithmetic strongly while excluding

Rasch and scale-type discipline. That combination signals institutional investment. In such a system, reform is not an incremental methodological update; it is a foundational repudiation of the evaluative core. It implies that the principal decision metric of cost per QALYhas never been a lawful quantitative object. Few institutions voluntarily adopt that conclusion, because it threatens legitimacy, training, publication norms, and the entire ecosystem of submissions and decisions.

The Norwegian case is not a hopeful exception. It is a confirmation case. It shows that a country can have a sophisticated adoption control system and still operate without representational measurement constraints. It shows that explicit prioritization criteria can coexist with composite arithmetic. It shows that "methodological rigor" in HTA can mean rigorous adherence to a template whose core constructs remain non-measures.

The conclusion, therefore, is not that Norway is uniquely deficient. It is that Norway exemplifies the global HTA structure: a stabilized, administratively effective decision framework anchored in QALY-based arithmetic and simulation outputs, while systematically excluding the axioms that would make those operations meaningful. If Norway were to move to representational measurement, it would need to abandon the cost-per-QALY centerpiece, replace it with a portfolio of evaluable claims grounded in unidimensional ratio measures (manifest events and resource units) and Rasch logit ratio measures (latent traits), and insist that falsification and replication apply to claims, not to models. The logit profile suggests that, as with other mature HTA jurisdictions, that transformation is possible in principle but institutionally unlikely in practice.

PLUS ÇA CHANGE, PLUS C'EST LA MÊME CHOSE

Norway's health technology assessment system presents itself as a model of modern, responsible governance. Over the past two decades, its institutional architecture has been refined, its procedures formalized, and its decision-making processes made more transparent. National coordination has strengthened. The "New Methods" system has integrated clinical and economic evaluation into a coherent framework. Ethical criteria to cover severity, benefit, and resource use have been explicitly articulated. Documentation is publicly available. Decision pathways are clearly defined. From an administrative standpoint, everything appears to change. The system evolves, adapts, and presents itself as an exemplar of rational priority setting in a publicly funded health system.

Yet beneath this visible institutional evolution lies a striking epistemic continuity. The quantitative foundation of the system remains unchanged. The core construct, the quality-adjusted life year (QALY), derived from multiattribute utility instruments, continues to function as the central measure of therapeutic value. Cost-per-QALY ratios continue to serve as the principal evaluative tool. Simulation models continue to project long-term outcomes based on composite utility scores. The arithmetic operations performed today are structurally identical to those employed decades ago. The apparatus has become more elaborate, but the measurement foundation remains untouched.

The logit profile makes this continuity visible. Statements asserting that measurement must precede arithmetic, that multiplication requires ratio measurement, that latent traits require Rasch transformation, and that representational measurement axioms must be satisfied register at floor values or near-floor levels. These results demonstrate not disagreement but absence of possession. The axioms do not operate as binding constraints within the Norwegian HTA knowledge base. Instead, composite utility scores are treated as if they were lawful measures, and arithmetic operations are performed without prior demonstration of admissible scale properties. The quantitative framework persists not because its measurement foundations have been validated, but because its institutional role has become entrenched.

This is the defining paradox of the Norwegian system. Procedural refinement has occurred without epistemic correction. Institutional sophistication has increased, but measurement validity has not. The system has become more efficient at producing quantitative outputs, but it has not addressed the foundational question of whether those outputs constitute measurement at all. The appearance of methodological progress conceals the persistence of structural error.

In normal science, progress occurs through falsification and replacement. Measurement-invalid constructs are identified, rejected, and superseded by constructs that satisfy representational criteria. This process ensures that quantitative claims become progressively more accurate, more reliable, and more useful for decision making. In the Norwegian HTA system, however, falsification has not occurred at the level of the central evaluative construct. The QALY remains in place, not because its measurement properties have been demonstrated, but because it continues to serve an administrative function. The system evolves procedurally while remaining epistemically static.

The implications extend beyond methodological purity. Health technology assessment exists to inform decisions affecting patient access to therapy, physician treatment choices, and health system resource allocation. These decisions rely on quantitative claims about therapeutic impact. If those claims are not grounded in valid measurement, then the decision framework lacks scientific accountability. Transparency of procedure cannot compensate for invalidity of measurement. Ethical commitment cannot substitute for empirical discipline. A system may be open, structured, and principled, yet still rely on constructs that do not meet the requirements of quantification.

Norway’s HTA system therefore illustrates a fundamental distinction between administrative change and scientific progress. Everything changes at the level of institutions, procedures, and governance. Nothing changes at the level that matters most: the measurement foundation on which quantitative claims depend. The apparatus becomes more sophisticated, but the epistemic core remains fixed. Without replacement of composite utility constructs by measurement-valid ratio scales with linear ratio measures for manifest attributes and Rasch logit ratio measures for latent traits the system cannot enter the trajectory of normal science. It can refine its procedures indefinitely, but refinement without measurement validity is not progress.

Jean-Baptiste Alphonse Karr (1808–1890). used the phrase in his journal *Les Guêpes* (“The Wasps”) in January 1849, commenting on political upheaval in France after the 1848 Revolution. His point was that despite apparent political transformation with new leaders, new institutions, new rhetoric the underlying structure of power and behavior remained unchanged.

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.

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