MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

DENMARK: NATIONAL REJECTION OF REPRESENTATIONAL MEASUREMENT IN HEALTH TECHNOLOGY ASSESSMENT

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

LOGIT WORKING PAPER No 234 FEBRUARY 2026

www.maimonresearch.com

Tucson AZ

FOREWORD

HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

Denmark occupies a prominent position within the European health technology assessment landscape, widely regarded as a country committed to rational priority setting, universal access, and evidence-based healthcare decision making. Its publicly funded health system operates within a structured national framework in which therapeutic evaluation, reimbursement, and clinical adoption are coordinated through formal institutional mechanisms. Danish HTA practice reflects strong integration with broader European methodological standards, including reliance on comparative clinical evidence, economic evaluation, and preference-based outcome measures. Academic centers, national advisory bodies, and clinical guideline committees contribute to a coherent evaluative environment that supports decisions affecting therapy availability and resource allocation. This institutional maturity, combined with Denmark's reputation for methodological rigor and transparency, creates the expectation that its evaluative framework would align with the scientific requirements of measurement. A national-level canonical logit assessment therefore provides a critical test of whether this expectation is realized at the level of measurement foundations governing therapeutic value claims.

The objective of this study was to determine whether the Danish national health technology assessment knowledge base embodies the axioms of representational measurement as operational constraints governing quantitative claims of therapeutic value. Denmark is widely regarded as a methodologically advanced HTA jurisdiction, with formalized economic evaluation requirements, structured reimbursement processes, and strong integration with European HTA standards. However, procedural sophistication alone does not establish measurement validity. The critical question is whether the constructs used to quantify therapy impact satisfy the necessary conditions for lawful arithmetic, including unidimensionality, dimensional homogeneity, invariant unit structure, and admissible ratio scale properties. To address this question, the study applied the 24-item canonical representational measurement diagnostic to the Danish national HTA knowledge environment, including methodological guidance, academic publications, reimbursement criteria, and policy frameworks. The objective was not to evaluate administrative competence, but to determine whether the quantitative constructs used to inform pricing, reimbursement, and clinical access decisions constitute valid empirical measures or represent composite scoring systems treated as if they were measures.

The logit profile demonstrates systematic non-possession of core representational measurement axioms within the Danish national HTA knowledge base. Statements asserting that multiplication requires ratio measurement, that measurement must precede arithmetic, and that latent constructs require Rasch transformation to achieve invariant measurement collapse to floor or near-floor logit values, indicating structural exclusion of these principles from the evaluative framework. Conversely, false statements asserting the legitimacy of QALYs as ratio measures, the admissibility of aggregating preference-weighted composite indices, and the acceptability of simulation-based cost-effectiveness ratios as empirical evaluative objects register positive logit values, indicating active endorsement. This pattern demonstrates that the Danish HTA knowledge environment substitutes composite utility scoring systems for measurement-valid quantities. Arithmetic operations are performed on constructs whose scale properties are not established, and cost-effectiveness ratios are interpreted as if they represent homogeneous quantitative relationships. The resulting evaluative system functions as a coherent administrative framework but does not operate within the constraints required for measurement-based scientific inference.

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 DANISH NATIONAL KNOWLEDGE BASE

Denmark's national HTA knowledge base reflects a mature and highly institutionalized evaluative environment shaped by national policy bodies, academic research centers, and reimbursement decision frameworks. Economic evaluation plays a central role in therapeutic assessment, with cost-effectiveness analysis serving as the primary quantitative mechanism for comparing therapies and informing reimbursement decisions. Manufacturers seeking reimbursement must provide economic models estimating incremental costs and incremental health outcomes, typically expressed in cost-per-QALY terms. These models integrate clinical evidence from trials and observational studies with preference-weighted outcome measures derived from multiattribute utility instruments such as the EQ-5D.

The Danish evaluative framework presents itself as a structured and transparent system intended to support rational allocation of healthcare resources. Decision criteria incorporate evidence of clinical effectiveness, disease severity, and economic impact. Economic modeling serves as the integrative quantitative component, translating heterogeneous clinical and resource inputs into a single summary ratio intended to support cross-therapy comparison. This structure aligns closely with international HTA practice and reflects methodological diffusion from European and global HTA agencies, particularly those that have adopted QALY-based reference case frameworks.

At the core of this knowledge base lies the QALY construct, which combines survival time with preference-based utility weights intended to represent health-related quality of life. Time is a manifest attribute measured on a ratio scale with a true zero and invariant unit structure. Utility weights, however, are derived from preference elicitation over multidimensional health state descriptions. These weights are generated through valuation protocols and scoring algorithms that assign numerical values to composite health states. The resulting values are treated operationally as if they possess interval or ratio scale properties, permitting multiplication by time and aggregation across individuals and populations.

Simulation modeling functions as the operational mechanism through which these constructs are applied. Economic models project long-term cost and outcome trajectories by integrating clinical inputs, epidemiological assumptions, and utility weights. These projections generate cost-per-QALY ratios used to inform reimbursement decisions and pricing negotiations. The numerical precision of these outputs derives from the internal coherence of the modeling framework rather than from demonstration that the underlying utility weights constitute invariant measures of a unidimensional latent attribute.

The Danish HTA knowledge base demonstrates strong procedural rigor, methodological consistency, and institutional stability. It reflects a coherent administrative framework designed to support transparent and systematic decision making. However, the framework does not require demonstration that the constructs used in economic evaluation satisfy representational measurement axioms. Composite utility scores function as operational decision variables despite lacking invariant unit structure or demonstrated ratio scale properties. Arithmetic operations performed on these constructs therefore proceed without established measurement foundations.

This condition reflects structural inheritance from international HTA methodological conventions rather than explicit demonstration of measurement validity. The Danish system reproduces established evaluative practices that emphasize modeling sophistication, procedural transparency, and quantitative integration. However, measurement validity remains assumed rather than demonstrated. The national knowledge base therefore embodies a quantitatively sophisticated but measurement-constrained evaluative architecture, in which composite scoring systems function as substitutes for empirical measurement in determining therapy value, reimbursement eligibility, and resource allocation decisions.

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: DENMARK

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.

DENMARK: NATIONAL DIFFUSION OF MEASUREMENT FAILURE

The logit profile of Denmark's national health technology assessment knowledge base demonstrates not an isolated misunderstanding of measurement principles but their systematic exclusion from the epistemic architecture that governs therapy evaluation (Table 1). The repeated collapse of foundational measurement axioms to floor values establishes that representational measurement does not operate as a binding constraint on quantitative claims within Danish HTA discourse. This exclusion is not accidental. It reflects the structural adoption of evaluative constructs that were never designed to satisfy measurement requirements and whose continued use depends on the absence of those requirements as operational constraints.

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

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	0.90	+2.20
CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.10	-2.20
QALYS CAN BE AGGREGATED	0	0.90	+2.20

NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.60	+0.40
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.85	+1.75
THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.50	+0.00
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	0.35	-1.25
THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT	1	0.20	-1.40
THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.05	-2.50

The most decisive evidence for measurement failure appears in the repeated floor values of −2.50 for statements asserting the necessity of representational measurement axioms, the requirement that latent constructs must be transformed using Rasch rules to achieve invariant interval scaling, and the recognition that lawful measurement exists only in two forms: linear ratio measurement for manifest attributes and Rasch logit ratio measurement for latent traits. A logit of −2.50 denotes effective non-possession. These propositions do not operate as constraints on what constitutes an admissible quantitative claim in Denmark's HTA knowledge base. Their absence is structural rather than incidental. They are not debated, rejected, or refuted. They simply do not function as organizing principles within the evaluative framework.

This absence has immediate consequences for arithmetic operations performed within HTA. Measurement precedes arithmetic. Arithmetic operations derive their meaning from the scale properties of the quantities to which they are applied. Without ratio scale properties, multiplication and division lack empirical meaning. Yet the Danish logit profile demonstrates that arithmetic operations are routinely performed on constructs whose measurement properties are not established. The strongly positive logit values associated with statements endorsing QALYs as ratio measures, permitting aggregation of utilities, and treating preference-weighted composite scores as objects of arithmetic manipulation demonstrate that the knowledge base substitutes scoring systems for measurement systems. Numbers are generated and manipulated without demonstration that they correspond to invariant empirical quantities.

The treatment of time provides a revealing contrast. Time is correctly recognized as a ratio measure, reflected in a logit of +2.50. This demonstrates that the Danish knowledge base is capable of recognizing lawful measurement structures when evaluating manifest attributes. Time possesses a true zero, invariant units, and supports meaningful multiplication and division. However, this recognition is not extended to latent constructs such as quality of life. Instead, latent constructs are treated as if measurement were unnecessary. Preference-weighted composite indices derived from multiattribute instruments are multiplied by time to generate QALYs without demonstration that the underlying utility scores possess ratio scale properties. This asymmetry reflects a structural division within the evaluative framework: lawful arithmetic for manifest attributes, unlawful arithmetic for latent constructs.

This asymmetry cannot be justified scientifically. Arithmetic admissibility is determined by scale properties, not by the nature of the attribute being evaluated. Latent constructs require transformation through Rasch measurement to achieve invariant unit structure. Without such transformation, numerical scores represent ordinal rankings rather than measurable quantities. The collapse of Rasch-related statements to floor values demonstrates that this transformation is not part of the Danish HTA knowledge base. Latent constructs are scored rather than measured.

The positive logit values associated with false statements concerning EQ-5D preference algorithms and QALY aggregation demonstrate the institutional normalization of composite constructs. Multiattribute utility instruments decompose health into multiple dimensions, assign preference weights, and aggregate those weights into composite scores. These scores reflect subjective valuations rather than invariant measurement. Their numerical properties depend on scoring algorithms rather than empirical invariance. Yet the Danish knowledge base treats these scores as if they possessed measurement properties. The resulting arithmetic operations create the appearance of quantification without satisfying the conditions required for measurement.

Simulation modeling reinforces this substitution of computation for measurement. Simulation models generate cost-per-QALY ratios by combining cost estimates with projected QALY gains. These projections depend on assumptions regarding disease progression, treatment effect duration, and utility weight stability. Simulation outputs cannot be falsified because they do not represent measured quantities. They represent computational outputs generated by model structure. Yet the logit profile demonstrates that simulation outputs are treated as legitimate evaluative objects. The positive logit associated with the false proposition that simulation models generate falsifiable claims demonstrates the normalization of this substitution.

The consequences extend beyond methodology to the evolution of objective knowledge. Scientific progress depends on measurement. Measurement enables falsification, replication, and cumulative knowledge development. Without measurement, numerical claims cannot be empirically evaluated. They become immune to refutation because they lack empirical referents. The Danish HTA knowledge base operates within such an environment. Numerical claims can be recalculated but not empirically tested. They exist within a closed computational framework.

This closure reflects institutional stabilization rather than empirical validation. Once evaluative frameworks become embedded in national guidelines, academic training, and methodological practice, they acquire institutional legitimacy independent of measurement validity. Their continued use reinforces the appearance of scientific rigor. Numerical outputs acquire authority through procedural acceptance rather than empirical correspondence.

Denmark's integration into the broader European HTA ecosystem reinforces this stabilization. Danish evaluative practice reflects methodological diffusion from international sources, particularly NICE and ISPOR. These frameworks institutionalized cost-per-QALY analysis as the central evaluative paradigm. Denmark adopted these conventions as part of its integration into global HTA practice. The logit profile demonstrates that Denmark inherited not measurement-valid constructs but the absence of measurement constraints.

This inheritance has implications for duty of care. HTA agencies influence therapy access, reimbursement, and clinical availability. Their decisions affect patient outcomes. When evaluative frameworks rely on constructs lacking measurement validity, decision-making becomes detached from empirical reality. Numerical outputs acquire administrative authority without empirical grounding.

The Danish logit profile demonstrates that this condition is structural. The repeated floor values associated with representational measurement axioms indicate that these principles are not recognized as necessary preconditions for arithmetic operations. Instead, arithmetic operations are performed on constructs lacking measurement foundations. This reversal of logical order transforms measurement from prerequisite to assumption. The absence of Rasch measurement is particularly consequential. Without invariant unit structure, latent construct comparisons lack stability. Differences between therapies cannot be interpreted as quantitative differences in patient outcomes. They represent differences in scoring outputs rather than differences in measured quantities. This distinction defines the boundary between measurement and scoring. Measurement produces quantities that exist independently of the scoring process. Scoring produces numbers that exist only within the scoring system. Denmark's HTA knowledge base operates on the latter. The persistence of this framework reflects epistemic stabilization. Measurement-valid alternatives, linear ratio measures for manifest attributes and Rasch logit ratio measures for latent constructs, are excluded. Their exclusion is reflected in repeated floor values across Rasch-related statements.

The implications extend beyond Denmark. They demonstrate that measurement failure is not jurisdiction-specific but systemic across global HTA practice. Denmark's logit profile aligns with those of Norway, Sweden, Finland, and other European countries. The pattern is consistent: lawful measurement principles are absent; composite scoring systems dominate.

Recovery requires structural reconstruction. Evaluative frameworks must restrict arithmetic operations to constructs satisfying measurement axioms. Manifest attributes must be measured using linear ratio scales. Latent constructs must be measured using Rasch transformation to establish invariant logit ratio scales. Only within such a framework can arithmetic regain empirical meaning. Only within such a framework can HTA re-enter the domain of measurement-based science.

The Danish logit profile demonstrates that this reconstruction has not yet occurred. Denmark's HTA knowledge base continues to operate within an evaluative framework that institutionalizes false measurement. Numerical sophistication masks measurement absence. Arithmetic proceeds without measurement foundation. Evaluation becomes computation without quantification.

The logit evidence demonstrates that this condition is structural. It reflects national diffusion of a memeplex rather than local methodological error. Until representational measurement axioms become operational constraints rather than ignored abstractions, Denmark's HTA knowledge base will remain epistemically detached from measurement-based science.

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