MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

THE MEASUREMENT FUTURE IN HEALTH TECHNOLGY ASSESSMENT

DENMARK: MEASUREMENT FAILURE IN LEGACY HEALTH TECHNOLOGY ASSESSMENT

Paul C Langley PhD Adjunct Professor, College of Pharmacy, University of Minnesota, Minneapolis, MN

LOGIT WORKING PAPER No 1077 AUGUST 2026

www.maimonresearch.com

Tucson AZ

ABSTRACT

Background: *Denmark has developed a mature and highly institutionalized health technology assessment (HTA) environment incorporating economic modelling, preference-based health outcomes, QALYs and cost-effectiveness analysis. The scientific legitimacy of these methods, however, depends upon a logically prior requirement: the quantities subjected to arithmetic must first possess the measurement properties required for the operations performed upon them.*

Objective: *To determine whether the Danish national HTA knowledge base recognizes and reinforces the principles of representational measurement required to establish measurement before arithmetic, or instead exhibits measurement inversion in which established quantitative operations precede demonstration of their measurement foundations.*

Methods: *A large language model interrogation assessed the publicly identifiable Danish HTA knowledge environment against 24 canonical TRUE/FALSE propositions addressing scale properties, representational measurement, admissible arithmetic, utilities, QALYs, Rasch measurement, latent attributes and falsifiability. Categorical endorsement probabilities were transformed to normalized logits ranging from −2.50 to +2.50. These probabilities characterize conceptual reinforcement within the knowledge base and are not statistical frequencies or estimates of individual practitioner beliefs.*

Results: *The Danish profile demonstrates pronounced measurement inversion. The TRUE proposition that measurement precedes arithmetic received an endorsement probability of 0.10 (−2.20), while the requirement that representational measurement axioms govern arithmetic received 0.05 (−2.50). Multiplication requiring ratio measurement received 0.10 (−2.20). In contrast, the FALSE propositions that the QALY is a ratio measure, is dimensionally homogeneous and can be aggregated each received 0.90 (+2.20). Time was strongly recognized as a ratio measure at 0.95 (+2.50), but the corresponding requirements governing multiplication were effectively absent. Core Rasch propositions concerning measurement of latent attributes were at or near the floor of the interrogation, with key items receiving 0.05 (−2.50).*

Conclusions: *The Danish national HTA knowledge base strongly reinforces the arithmetic conventions of legacy HTA while weakly representing the measurement principles required to justify them. This pattern is structural rather than a deficiency attributable to individual practitioners and reflects Denmark's inheritance of wider European and international HTA conventions. Greater modelling sophistication cannot repair measurement properties absent from the inputs. The required transition is therefore fundamental: restore the scientific sequence of attribute → measurement → scale → admissible arithmetic → therapy-impact claim → empirical evaluation, replication and falsification. Measurement must again become the gatekeeper of arithmetic.*

INTRODUCTION

Denmark has developed a mature and highly institutionalized framework for health technology assessment and the economic evaluation of health technologies. Clinical evidence, economic modelling, preference-based measures of health-related quality of life and cost-effectiveness analysis contribute to an evaluative environment intended to support transparent and systematic decisions over reimbursement and healthcare resource allocation. Within this framework, manufacturers may be required to submit economic models estimating incremental costs and health outcomes, with the QALY providing a familiar basis for expressing health benefits and cost-per-QALY comparisons. Danish practice consequently reflects many of the methods that have become established across European and international HTA.

The scientific question addressed in this paper, however, precedes the application of these methods. Are the quantities entering Danish HTA models measures, and do they possess the scale properties required for the arithmetic performed upon them? The ability to assign numerical values to health states, enter those values into economic models and generate apparently precise quantitative outputs does not itself answer that question. Measurement must first be established. Only then can it be determined which mathematical operations are admissible.

This requirement follows directly from representational measurement theory [1 2 3]. Nominal, ordinal, interval and ratio scales possess different mathematical properties and therefore permit different arithmetic operations. Multiplication and division require ratio measurement. Addition and subtraction require the appropriate quantitative structure. Dimensional homogeneity must be maintained. For latent attributes, numerical scores cannot acquire measurement status merely through assignment, summation or statistical transformation. The measurement structure required for the intended arithmetic must be demonstrated independently of the arithmetic itself.

The distinction is particularly important for the QALY. Its familiar construction combines survival time with a preference-based utility value. Time can possess the properties of a manifest ratio measure. The critical question is whether the utility value possesses the measurement properties required to function as a proportional adjustment to time. Danish HTA practice employs preference-weighted outcomes derived from instruments such as the EQ-5D, with resulting values treated operationally as though they possess the interval or ratio properties required for multiplication and aggregation. The existence of a scoring algorithm, however, cannot confer those properties.

The same problem cannot be corrected downstream. If the utility value has not been established as an appropriate measure, multiplication by time cannot manufacture a QALY measure. Subtraction to produce incremental QALYs cannot repair the initial failure, and subsequent division of incremental costs by an inadmissible difference in QALYs cannot establish a scientifically meaningful cost-per-QALY ratio. Nor can probabilistic sensitivity analysis, increasingly sophisticated simulation or greater computational precision resolve a measurement failure that occurred before modelling began. The Danish knowledge base itself illustrates the distinction: numerical precision is generated by the internal coherence of the modelling framework without a prior demonstration that the utility weights constitute invariant measures of a unidimensional attribute.

This creates the possibility of measurement inversion. In normal quantitative science, the sequence is measurement followed by admissible arithmetic. Measurement inversion occurs when numerical operations become established components of an analytical framework while the measurement properties required to justify those operations remain unexamined or assumed. A framework can consequently become methodologically sophisticated, institutionally embedded and procedurally rigorous while failing the more elementary requirement that its numerical inputs first qualify as measures.

Denmark provides an important setting in which to examine this possibility. Its HTA environment demonstrates procedural rigor, methodological consistency and institutional stability. Yet the national knowledge base examined here does not require demonstration that the constructs employed in economic evaluation satisfy the axioms of representational measurement before they are subjected to arithmetic. Composite utility scores function as decision variables despite the absence of demonstrated invariant unit structure or ratio properties. This is not an assessment of the competence or beliefs of individual Danish researchers, analysts or decision-makers. The unit of analysis is the publicly identifiable national HTA knowledge environment: the methodological assumptions, guidance, analytical conventions and practices through which quantitative claims are generated and justified.

The paper therefore interrogates the Danish national HTA knowledge base against 24 canonical statements derived from representational measurement theory and Rasch measurement principles. Each statement has an independently specified TRUE or FALSE status. Categorical endorsement probabilities are used to assess the extent to which the knowledge base recognizes, neglects or behaves as though each proposition were true. A high endorsement of a false proposition does not imply that Danish practitioners consciously believe something false; it indicates that the knowledge environment operates as though the proposition were valid. Conversely, low endorsement of a true proposition indicates that the required measurement principle has little identifiable presence within that knowledge environment.

The central question is consequently straightforward. Does Danish HTA establish measurement before arithmetic, or has it inherited an international analytical framework in which arithmetic has been allowed to substitute for measurement? If the latter is demonstrated, the problem is not a shortage of modelling sophistication or procedural rigor. It is more fundamental: the quantitative framework has been constructed in the wrong scientific order.

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 HTA KNOWLEDGE BASE

The Danish national HTA knowledge base represents a mature and highly institutionalized environment for the evaluation of health technologies. It is shaped by national policy and reimbursement processes, economic evaluation requirements, academic health-economics research and the wider European methodological environment within which Danish HTA operates. The knowledge base is therefore not identified with a single institution or document. It comprises the publicly available methodological guidance, analytical conventions, technical reports, economic evaluation practices, academic contributions and decision frameworks that collectively define the quantitative language of Danish HTA.

Economic evaluation occupies a prominent position within this environment. Cost-effectiveness analysis provides a framework for comparing the additional costs and expected health consequences of alternative therapies, while manufacturer submissions may incorporate economic models projecting incremental costs and health outcomes. Preference-based measures, including instruments such as the EQ-5D, provide utility values that can be combined with survival time to generate QALYs. These quantities can then enter incremental analyses and cost-per-QALY calculations intended to inform reimbursement and resource-allocation decisions.

The Danish knowledge base is consequently quantitatively sophisticated. Clinical trial evidence, observational data, epidemiological assumptions, costs, resource utilization, survival estimates and preference-based health-state values can be incorporated within decision models projecting outcomes beyond the period directly observed. Modelling provides the mechanism through which heterogeneous inputs are combined into a common analytical structure, with uncertainty explored through sensitivity and probabilistic analyses. This approach is consistent with the broader international development of contemporary HTA and with the diffusion across European systems of QALY-based economic evaluation.

The issue for the present assessment is not whether these procedures are technically sophisticated. It is whether the knowledge base establishes the measurement properties of the quantities entering them before permitting arithmetic operations upon those quantities. This distinction is fundamental. Numerical assignment is not measurement, and the existence of a scoring algorithm does not establish a measurement scale. Similarly, the ability of a computer model to add, subtract, multiply or divide numerical values does not demonstrate that those operations have empirical meaning.

The QALY provides the clearest example. Within the Danish evaluative environment, survival time is combined with preference-based utility values intended to represent health-related quality of life. Time can satisfy the requirements of manifest ratio measurement. Utility values, however, are generated from preference elicitation over multidimensional health-state descriptions and scoring algorithms. They are subsequently treated operationally as though they possess the quantitative properties required for multiplication by time and for subsequent aggregation. The Danish knowledge base does not first require independent demonstration that these utility values possess an invariant unit structure and the ratio properties required for these operations.

This distinction extends throughout the economic model. Simulation can integrate clinical evidence, epidemiological assumptions, costs and utility values to generate internally consistent projections and apparently precise cost-per-QALY estimates. But internal mathematical consistency is not evidence of external measurement validity. Greater computational sophistication cannot establish measurement properties that were absent from the inputs. The numerical precision of a model therefore needs to be distinguished from the empirical legitimacy of the quantities being manipulated within it.

The same issue applies to latent attributes. Health-related quality of life, pain, fatigue, physical function and similar patient-reported constructs cannot acquire measurement status simply because responses have been assigned numbers, summed or transformed by an algorithm. A latent attribute must first be defined as a unidimensional construct and an appropriate measurement structure established. Within the framework applied in this assessment, Rasch measurement provides the required person-item conjoint structure, invariance and specific objectivity necessary to establish a latent continuum. The relevant question is therefore not whether patient-reported information is incorporated into Danish HTA, but whether the resulting numerical values satisfy the measurement requirements necessary for the quantitative claims made from them.

The Danish knowledge base examined here shows considerable procedural rigor, methodological consistency and institutional stability. These are important characteristics of an established decision system. They do not, however, demonstrate representational measurement. The publicly identifiable framework does not make satisfaction of measurement axioms a prior condition for the use of composite utility scores in economic evaluation. Instead, these scores function as operational decision variables while their invariant unit structure and ratio properties remain undemonstrated. Arithmetic therefore proceeds without measurement operating as its scientific gatekeeper.

This pattern should be understood in its international context. Danish HTA did not independently invent the principal conventions being examined. QALYs, preference-based utility measurement, cost-effectiveness modelling and reference-case approaches form part of a wider international HTA inheritance. Denmark has adopted and institutionalized methods that have achieved widespread professional acceptance across European and global HTA. The resulting measurement problem is therefore better understood as inherited and reproduced rather than uniquely Danish. The knowledge base reflects methodological diffusion in which established analytical conventions have been transmitted without requiring prior demonstration of their measurement validity.

This defines the knowledge base subjected to interrogation. The object of assessment is not the knowledge, competence or intentions of individual Danish researchers, faculty, analysts or decision-makers. It is the collective epistemic structure represented by the publicly available Danish HTA discourse. The question is whether that structure possesses and applies the measurement principles required to support its quantitative operations.

The distinction is critical. A highly developed HTA system may possess extensive expertise in economic evaluation, statistical analysis, simulation modelling and decision analysis while failing to possess the prior measurement principles required to determine whether the arithmetic underlying those methods is admissible. Institutional maturity demonstrates that a methodology has become established. It does not demonstrate that the quantities on which that methodology depends are measures.

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

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.

ASSESSMENT: NATIONAL DIFFUSION OF MEASUREMENT FAILURE

The Danish probability and normalized-logit profile present a remarkably coherent picture of measurement inversion. It does not reveal a knowledge base that possesses the principles of representational measurement but applies them inconsistently (Table 1). Nor does it suggest a marginal disagreement over one or two technical aspects of economic evaluation. The pattern is more fundamental. The principles required to establish measurement before arithmetic receive some of the weakest endorsements in the entire interrogation, while propositions incompatible with those principles are strongly reinforced by established HTA practice

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 central result is difficult to misinterpret. **"**Measurement precedes arithmetic" is a TRUE statement, yet its endorsement probability is only 0.10, with a normalized logit of −2.20. Even more strikingly, the proposition that meeting the axioms of representational measurement is required for arithmetic receives an endorsement probability of only 0.05 and the floor normalized logit of −2.50. The requirement that multiplication demands a ratio measure performs little better at 0.10 and −2.20. These are not peripheral propositions. They define the conditions under which quantitative operations acquire empirical meaning. Their location at or close to the negative extreme indicates that measurement does not function as an operative constraint within the Danish HTA knowledge base.

The contrast with the treatment of the QALY is stark. The proposition "the QALY is a ratio measure" is FALSE, yet the Danish knowledge base has an endorsement probability of 0.90 and a normalized logit of +2.20. The equally false proposition that the QALY is a dimensionally homogeneous measure receives exactly the same values: 0.90 and +2.20. The proposition that QALYs can be aggregated, again FALSE under the measurement framework applied here, is also endorsed at 0.90 and +2.20. The knowledge base therefore strongly reinforces the quantitative status of the QALY precisely while failing to reinforce the measurement principles that would be required to establish that status.

This is measurement inversion in unusually clear form. The arithmetic is strongly possessed; the conditions that would make the arithmetic admissible are not. The QALY is accepted operationally as though its measurement properties had already been established, while the principles necessary to establish those properties are effectively absent. The direction of scientific reasoning has been reversed. Instead of establishing measurement and then determining permissible arithmetic, established arithmetic appears to confer legitimacy retrospectively upon the numerical constructs required to sustain it.

The contradiction becomes sharper when the components of the QALY are considered separately. Denmark strongly recognizes that time is a ratio measure: the TRUE statement receives an endorsement probability of 0.95 and the ceiling normalized logit of +2.50. Yet recognition of the ratio status of time is accompanied by almost complete failure to recognize that multiplication itself requires ratio measurement. Knowing that one component is a ratio measure does not establish the measurement status of the other component. Nor can the legitimate measurement properties of time be transferred to a preference-based utility value merely by multiplying the two together. The Danish profile therefore recognizes the measurement status of time while failing to apply the very scale requirements necessary to determine whether its multiplication by utility is admissible.

The treatment of preference-based and subjective scores reinforces the same conclusion. The false proposition that time trade-off preferences are unidimensional is endorsed at 0.85 and +1.75. The false proposition that EQ-5D-3L preference algorithms create interval measures is likewise endorsed at 0.85 and +1.75. Summations of subjective instrument responses being ratio measures and summated Likert scores creating ratio measures, both FALSE, each receive endorsement probabilities of 0.85 and normalized logits of +1.75. Most strikingly, the false proposition that ratio measures can have negative values receives an endorsement probability of 0.90 and +2.20. Taken together, these results expose a knowledge environment in which numerical assignment has effectively displaced measurement demonstration. Scores can be generated, algorithms applied, responses summed and preference values manipulated, after which the resulting numbers are permitted to enter economic models. What is missing is the prior scientific question: what attribute is being measured, what measurement structure has been demonstrated, what scale has been created, and what arithmetic does that scale permit? The profile indicates that these questions are not merely underemphasized. They have little effective authority over the quantitative framework.

The results for latent attributes are particularly severe. The TRUE proposition identifying the Rasch logit ratio scale as the basis for assessing therapy impact for latent traits receives an endorsement probability of 0.05 and the floor normalized logit of −2.50. The proposition linking Rasch measurement rules to the axioms of representational measurement is likewise at 0.05 and −2.50. The proposition concerning transformation of subjective responses under Rasch rules is also at 0.05 and −2.50. Even recognition that the outcome of interest for a latent attribute concerns possession of that attribute reaches only 0.20 and −1.40.

This is not a minor omission from an otherwise measurement-based system. The machinery required to establish latent measurement is effectively missing from the epistemic architecture being interrogated**.** At the same time, numerical representations of subjective and preference-based constructs are routinely accommodated within the economic framework. The result is a profound substitution: scoring is permitted to perform the role that measurement should perform, while modelling subsequently gives those scores an appearance of quantitative legitimacy.

The cost-effectiveness result demonstrates the downstream consequence. The TRUE proposition that claims for cost-effectiveness fail the axioms of representational measurement receives an endorsement probability of only 0.10 and a normalized logit of −2.20. This is entirely consistent with the rest of the profile. Once QALYs have been accepted as legitimate quantitative outcomes, the incremental QALY and the cost-per-QALY ratio follow almost automatically within the conventional framework. But additional arithmetic cannot correct the measurement failure embedded in the denominator. If the QALY has not first been established as a measure possessing the properties required for its construction and manipulation, subtraction does not rescue it and division by it does not create a scientifically meaningful ratio.

The profile also identifies a problem at the level of scientific claims. Denmark shows some recognition of the principle that non-falsifiable claims should be rejected, with an endorsement probability of 0.60 and a modest positive normalized logit of +0.40. Yet the false proposition that reference-case simulations generate falsifiable claims receives substantially stronger endorsement at 0.85 and +1.75. The contrast is revealing. A general commitment to falsifiability has limited scientific value if the analytical framework continues to generate modeled claims that cannot be subjected to the empirical testing required to falsify them. The language of normal science can therefore coexist with an evidentiary structure that does not satisfy its requirements.

The overall Danish pattern is consequently not epistemic noise. It is remarkably ordered. TRUE propositions that would constrain established HTA arithmetic repeatedly receive strongly negative logits. FALSE propositions that permit established HTA arithmetic repeatedly receive strongly positive logits. The asymmetry is precisely what would be expected from a mature methodological system whose practices became institutionalized without representational measurement operating as their prior scientific foundation.

This matters because Denmark's HTA environment is not primitive, technically unsophisticated or administratively disorganized. Quite the opposite. Its sophistication makes the result more consequential. Methodological refinement, procedural transparency, increasingly complex modelling and institutional stability have developed around a quantitative architecture in which the legitimacy of the underlying measures was not first established. Forty years of methodological refinement cannot compensate for a measurement requirement that was omitted at the beginning.

Nor can widespread international acceptance provide the missing demonstration. Denmark has inherited much of this architecture from the wider European and international HTA environment. QALYs, utility instruments, cost-effectiveness modelling and reference-case conventions arrived with substantial professional legitimacy already attached to them. But methodological diffusion transmits assumptions as readily as it transmits valid methods. International consensus can explain why a practice became established; it cannot establish the measurement properties of the quantities upon which that practice depends.

The Danish results therefore support a tougher conclusion than simply identifying deficiencies in knowledge of measurement theory. The national HTA knowledge base has institutionalized arithmetic without first institutionalizing the science required to determine whether that arithmetic is admissible. The problem is structural. The established framework does not merely overlook a few concepts from measurement theory; it depends upon numerical operations that would be placed in question if those concepts were applied as binding requirements.

This is why adding greater sophistication to the existing model cannot provide the solution. More elaborate simulations, additional uncertainty analyses, larger evidence networks and increasingly detailed economic models all begin downstream of the same unresolved question. If the inputs have not been established as measures, modelling cannot make them measures. If multiplication is inadmissible, simulation cannot make it admissible. If a QALY has no demonstrated measurement structure, aggregating thousands of projected QALYs does not create one.

The Danish profile therefore documents national diffusion of measurement failure. It describes an HTA knowledge environment in which the rules governing arithmetic are determined predominantly by methodological convention rather than demonstrated measurement properties. Denmark has inherited, refined and institutionalized an international framework in which measurement was never established as the gatekeeper of quantitative claims.

The implication is unavoidable. The issue is not how Danish HTA can further refine the legacy framework. It is whether that framework can continue to be defended once its arithmetic is required to meet the standards of representational measurement. On the evidence of this interrogation, **it cannot.** Measurement must be restored to its proper position: before arithmetic.

CONCLUSION

The interrogation of Denmark's national HTA knowledge base identifies a systematic inversion of the relationship between measurement and arithmetic. The problem is not an absence of quantitative sophistication. Danish HTA operates within a mature evaluative environment characterized by economic modelling, preference-based outcomes, QALYs, cost-effectiveness analysis and structured approaches to reimbursement and resource allocation. The problem is more fundamental: the measurement properties required to justify these quantitative operations have not been established as prior conditions for their use [4] [5].

The probability and normalized-logit results make this inversion explicit. The TRUE proposition that measurement precedes arithmetic receives an endorsement probability of only 0.10 and a normalized logit of −2.20. The requirement that the axioms of representational measurement must be satisfied for arithmetic receives 0.05 and −2.50, while recognition that multiplication requires a ratio measure is similarly weak at 0.10 and −2.20. These are not optional refinements to quantitative analysis. They determine whether numerical operations have empirical meaning.

Against this near-absence of the principles that should constrain arithmetic stands strong reinforcement of the assumptions required to sustain legacy HTA. The FALSE propositions that the QALY is a ratio measure, that it is dimensionally homogeneous and that QALYs can be aggregated each receive an endorsement probability of 0.90 and a normalized logit of +2.20. The contrast is decisive. The Danish knowledge base strongly possesses the arithmetic conventions of legacy HTA while failing to possess the measurement principles required to justify them. The latent-attribute results reinforce the same conclusion. Core propositions concerning Rasch measurement reach the floor of the interrogation at 0.05 and −2.50. The Rasch measurement pathway therefore has essentially no identifiable presence as a governing requirement for establishing latent measures, while preference-based and subjective numerical scores are permitted to enter the economic evaluation framework. This is not a contest between alternative statistical techniques. It concerns the prior requirement to establish a measure before numerical scores are subjected to arithmetic.

Nor can the problem be repaired downstream. If utility has not been established as possessing the measurement properties required for multiplication by time, the QALY cannot acquire those properties merely because the multiplication is performed. If the QALY is not a legitimate quantitative measure, subtraction cannot manufacture a legitimate incremental QALY. Division of incremental costs by that difference cannot subsequently manufacture a scientifically meaningful cost-per-QALY ratio. Probabilistic sensitivity analysis, simulation and increasingly sophisticated decision models cannot restore measurement properties that were absent before the arithmetic began.

The importance of this finding extends beyond Denmark. The contemporary Danish framework is substantially inherited from the wider European and international development of HTA. QALYs, preference-based utilities, cost-effectiveness analysis and modelling conventions arrived as established components of an internationally accepted methodology. Denmark therefore illustrates how measurement failure can diffuse through a scientific and professional community precisely because the inherited methods already carry institutional legitimacy.

This inheritance explains how the present position arose. It cannot provide its scientific justification**.** Widespread adoption does not establish a scale of measurement. International guidelines cannot make an inadmissible arithmetic operation admissible. Professional consensus cannot supply an invariant unit, create a non-arbitrary zero or transform a multidimensional preference score into a unidimensional ratio measure. Forty years of institutional use cannot confer measurement properties retrospectively upon a numerical construct that never possessed them.

The implications should therefore not be directed toward individual Danish researchers, analysts or decision-makers. They were trained within, and work within, an international framework in which these methods have been treated as established standards of practice. The object of criticism is the knowledge system that made proficiency in those methods possible without requiring prior proficiency in the measurement principles necessary to judge their scientific legitimacy.

But inheritance cannot remain a sufficient explanation once the measurement failure has been exposed. The appropriate response is not another refinement of the legacy framework. Nor is it the construction of increasingly elaborate models around the same unexamined numerical foundations. What was previously accepted can no longer simply be accepted.

Denmark has the institutional and scientific capacity to make the transition. Nothing requires the invention of a new standard of quantitative science. The necessary standards already exist. For manifest attributes, quantitative claims must rest upon appropriate linear ratio measurement before multiplication, division or ratio comparison is undertaken. For latent attributes, the attribute must be defined and a measurement structure established through conjoint measurement, with invariance and the other requirements necessary to construct a legitimate latent continuum. Arithmetic must then be restricted to the operations permitted by the resulting scale. This restores the correct scientific sequence: attribute → measurement → scale → admissible arithmetic → therapy-impact claim → empirical evaluation, replication and falsification.

The significance of the Danish findings therefore lies not simply in documenting another national example of measurement inversion. They identify a choice. Denmark can continue to refine an inherited HTA framework whose fundamental arithmetic precedes demonstration of measurement, or it can require that quantitative therapy evaluation meet the measurement standards expected elsewhere in normal science.

The first course preserves methodological convention. The second restores scientific discipline. Denmark does not need more sophisticated arithmetic built upon unestablished measures. It needs measurement to become the gatekeeper of the arithmetic it already performs. Measurement first. Arithmetic second.

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] Langley P. California: Legacy HTA – 40 years of numerical storytelling. Logit Working Paper No 6765. https://maimonresearch.com/logit-working-paper-no-6765-august-2026/

[5] Langley P. Legacy HTA: Is this the worst we could have done? Logit Working Paper No 7987 https://maimonresearch.com/logit-working-paper-no-7987-august-2026/