MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

MEASUREMENT INVERSION AND THE EQ-5D: DEFINING HTA CLOSURE

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

LOGIT WORKING PAPER No 730 MAY 2026

www.maimonresearch.com

Tucson AZ

FOREWORD

HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

The EQ-5D-3L and EQ-5D-5L instruments were developed by the EuroQol Group as generic measures of health status intended for use across diseases, treatments and populations. Their primary purpose is not simply to describe health states but to generate preference-based utility scores that can be used in economic evaluation. The instruments achieve this by classifying respondents according to a small number of health dimensions and then converting those classifications into numerical utility values through the application of valuation algorithms.

The EQ-5D descriptive system comprises five dimensions: mobility, self-care, usual activities, pain/discomfort and anxiety/depression. Respondents select the level that best describes their current health status within each dimension. The resulting combination of responses defines a health-state profile. In the EQ-5D-3L, each dimension has three response levels, while the EQ-5D-5L expands this to five levels in an effort to improve sensitivity and reduce ceiling effects.

The principal objective of the EQ-5D framework is to convert these health-state descriptions into utility scores that are interpreted as representing the value or desirability of a health state relative to full health. These utility scores are then combined with survival time to create quality-adjusted life years (QALYs), the preferred outcome measure in many health technology assessment systems. The EQ-5D therefore occupies a central position within contemporary HTA. It functions not merely as a health-status classification system but as the foundation for utility estimation, QALY generation, cost-effectiveness analysis and reference-case simulation modelling. Consequently, any assessment of the measurement properties of utilities and QALYs must begin with an examination of the conceptual and methodological framework underpinning the EQ-5D instruments.

The objective of this study is to evaluate the EQ-5D knowledge base against the requirements of representational measurement. For the purposes of this assessment, the EQ-5D knowledge base is defined as the conceptual, methodological, valuation and application framework supporting the EQ-5D-3L and EQ-5D-5L instruments, including health-state classification, time trade-off valuation, econometric tariff construction, utility scoring algorithms and QALY generation. The study does not assess the popularity, practical utility or widespread adoption of the EQ-5D instruments. Rather, it seeks to determine whether the framework recognizes and applies the conditions necessary for quantitative claims, including unidimensionality, ratio measurement, dimensional homogeneity, admissible arithmetic operations, latent and manifest attributes, and empirical evaluability. To allow direct comparison with previous interrogations of HTA agencies, academic centres, professional organizations and pharmacy schools, the assessment applies the same 24 canonical statements derived from the axioms of representational measurement.

The interrogation results indicate a consistent pattern of measurement inversion within the EQ-5D knowledge base. Statements that express the requirements of representational measurement, ratio scales, dimensional homogeneity, Rasch measurement and falsifiable claims receive uniformly low endorsement probabilities. In contrast, statements supporting utility construction, QALY generation, preference-based valuation and simulation modelling receive high endorsement probabilities despite conflicting with established measurement principles. The results suggest that the EQ-5D framework accepts arithmetic operations before establishing lawful measurement. Time trade-off valuations are treated as quantitative inputs, econometric coefficients are transformed into decrement weights, utility scores are interpreted as measures and QALYs are treated as ratio-scale outcomes without demonstrating the conditions required for such claims. The overall endorsement profile is closely aligned with previous interrogations of HTA agencies and research centers indicating that the measurement inversion observed throughout contemporary health technology assessment is embedded within one of its most influential methodological frameworks.

The objective of this study is to evaluate the EQ-5D knowledge base against the requirements of representational measurement. For the purposes of this assessment, the EQ-5D knowledge base is defined as the conceptual, methodological, valuation and application framework supporting the EQ-5D-3L and EQ-5D-5L instruments, including health-state classification, time trade-off valuation, econometric tariff construction, utility scoring algorithms and QALY generation. The study does not assess the popularity, practical utility or widespread adoption of the EQ-5D instruments. Rather, it seeks to determine whether the framework recognizes and applies the conditions necessary for quantitative claims, including unidimensionality, ratio measurement, dimensional homogeneity, admissible arithmetic operations, latent and manifest attributes, and empirical evaluability. To allow direct comparison with previous interrogations of HTA agencies, academic centers, professional organizations and pharmacy schools, the assessment applies the same 24 canonical statements derived from the axioms of representational measurement.

The interrogation results indicate a consistent pattern of measurement inversion within the EQ-5D knowledge base. Statements that express the requirements of representational measurement, ratio scales, dimensional homogeneity, Rasch measurement and falsifiable claims receive uniformly low endorsement probabilities. In contrast, statements supporting utility construction, QALY generation, preference-based valuation and simulation modelling receive high endorsement probabilities despite conflicting with established measurement principles. The results suggest that the EQ-5D framework accepts arithmetic operations before establishing lawful measurement. Time trade-off valuations are treated as quantitative inputs, econometric coefficients are transformed into decrement weights, utility scores are interpreted as measures and QALYs are treated as ratio-scale outcomes without demonstrating the conditions required for such claims. The overall endorsement profile is closely aligned with previous interrogations of HTA agencies and research centers, indicating that the measurement inversion observed throughout contemporary health technology assessment is embedded within one of its most influential methodological frameworks.

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 endorsement 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 EQ-5D HTA KNOWLEDGE BASE FOR HEALTH TECHNOLOGY ASSESSMENT

The EQ-5D knowledge base is defined pragmatically and operationally. Rather than focusing solely on the EQ-5D-3L and EQ-5D-5L descriptive instruments, the knowledge base encompasses the conceptual, methodological, valuation and application framework that supports the generation and use of utility scores and QALYs. The corpus therefore consists of:

• the EQ-5D-3L and EQ-5D-5L descriptive systems and associated classification frameworks

• EuroQol methodological manuals, user guides and technical documentation

• time trade-off (TTO), standard gamble and related preference-elicitation methodologies used in health-state valuation

• national and international EQ-5D valuation studies and tariff-development programs

• econometric models used to estimate preference algorithms and utility scoring systems

• utility-scoring algorithms, crosswalk methodologies and value-set construction procedures

• conceptual and methodological literature supporting the interpretation of utilities as quantitative measures of health status

• QALY construction methodologies based upon the multiplication of utility scores by time

• cost-effectiveness reference cases and HTA guidelines that specify or recommend the use of EQ-5D-derived utilities

• reference-case simulation models, modelling templates and technical reports employing EQ-5D utility values

• academic publications, textbooks, methodological reviews and task-force recommendations supporting the use of EQ-5D utilities in economic evaluation

• teaching materials, training programs, workshops and institutional guidance documents concerned with EQ-5D valuation, utility estimation and QALY-based decision making

Taken together, these components define the intellectual framework that links health-state classification to preference valuation, econometric estimation, utility construction, QALY generation and health technology assessment. The interrogation therefore focuses not merely on the descriptive instrument itself, but on the entire methodological system that supports the use of EQ-5D utilities as inputs to claims regarding health outcomes, cost-effectiveness and resource allocation.

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: EQ-5D HTA KNOWLEDGE BASE

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.

EQ-5D HTA: THE EXTENT OF MEASUREMENT INVERSION

The interrogation of the EQ-5D knowledge base provides one of the clearest examples of measurement inversion in contemporary health technology assessment. The target knowledge base was defined as the conceptual, methodological, valuation and application framework supporting the EQ-5D-3L and EQ-5D-5L instruments, including time trade-off valuation, econometric tariff construction, utility scoring and QALY generation. This definition is important because the EQ-5D is not merely a questionnaire. It is an entire methodological system that begins with health-state classification and ends with the production of utilities, QALYs and inputs to cost-effectiveness models.

TABLE 1: ITEM STATEMENT, RESPONSE, ENDORSEMENT AND NORMALIZED LOGITS EQ-5D HTA KNOWLEDGE BASE

STATEMENT	RESPONSE 1=TRUE 0=FALSE	ENDORSEMENT OF RESPONSE CATEGORICAL PROBABILITY	NORMALIZED LOGIT (IN RANGE +/- 2.50)
INTERVAL MEASURES LACK A TRUE ZERO	1	0.15	-1.75
MEASURES MUST BE UNIDIMENSIONAL	1	0.10	-2.20
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	0.80	+1.40
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	0.80	+1.40
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	0.20	-1.40
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.86	+1.75
THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.70	+0.85
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	0.50	0.00
THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT	1	0.05	-2.50
THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.05	-2.50

The results show a consistent pattern (Table 1). Statements expressing the foundational requirements of measurement receive low endorsement. Statements supporting the utility/QALY framework receive high endorsement, even when they are false from the perspective of representational measurement. This is the signature of measurement inversion: arithmetic is accepted before measurement is established.

The weak endorsement of the proposition that measurement precedes arithmetic is central. This statement receives a probability of 0.10 and a normalized logit of -2.20. This captures the intellectual structure of the EQ-5D framework. Health-state descriptions are valued through TTO exercises. These values become the dependent variable in econometric models. Coefficients are estimated, reinterpreted as decrement weights, combined in additive algorithms and transformed into utility scores. These scores are then multiplied by time to create QALYs. At every stage arithmetic is performed, yet the prior measurement question is left unresolved.

The weak endorsement of unidimensionality is equally important. Measures must be unidimensional, yet this statement receives an endorsement probability of only 0.10 and a normalized logit of -2.20. The EQ-5D descriptive system is explicitly multidimensional. It combines mobility, self-care, usual activities, pain/discomfort and anxiety/depression. The TTO valuation of a health state such as 21123 is therefore a valuation of a composite description. No evidence is provided that the resulting value represents a single measurable attribute. Yet without a single attribute there can be no lawful measure. The utility score is therefore not the result of measurement but the result of scoring.

The results also show little recognition of the requirements for multiplication. The statement that multiplication requires a ratio measure receives an endorsement probability of 0.10 and a normalized logit of -2.20. This is a critical failure because the QALY depends upon multiplication. Time is a ratio measure, and this is correctly recognized with an endorsement probability of 0.95 and a normalized logit of 2.50. But the problem is not time. The problem is the utility score. If the utility score is to discount time, it must itself possess ratio properties. It must have a meaningful zero, support ratio comparisons and refer to a single attribute. None of these requirements is demonstrated.

The interrogation indicates strong endorsement of the false proposition that the QALY is a ratio measure, with a probability of 0.90 and a normalized logit of 2.20. This is not surprising because the QALY occupies a central position in the EQ-5D application framework. The utility score is treated as a proportion of full health and multiplied by time. A utility of 0.5 over twelve months is interpreted as six months in perfect health. Yet this interpretation is only meaningful if the utility score is a lawful ratio-scale discount factor. The EQ-5D framework assumes this condition but does not demonstrate it.

The endorsement of the claim that ratio measures can have negative values is also high, with a probability of 0.90 and a normalized logit of 2.20. This reflects the treatment of health states worse than death within utility scoring systems. From the perspective of ratio measurement, however, this is a decisive problem. A ratio scale requires a meaningful non-arbitrary zero. Values below zero are incompatible with a scale where zero represents the absence of the attribute. The presence of negative utilities therefore raises fundamental questions about the claimed ratio status of utility scores. If the utility scale permits negative values, it cannot simultaneously function as a ratio measure suitable for proportional interpretation and multiplication by time.

The interrogation also indicates strong endorsement of the false proposition that the QALY is dimensionally homogeneous. This statement receives a probability of 0.90 and a normalized logit of 2.20. Dimensional homogeneity is essential for admissible arithmetic. The QALY combines a utility score with time. Time is a unidimensional ratio measure. The utility score, however, is derived from a multidimensional classification system and an econometric scoring algorithm. No evidence is provided that the utility score has the dimensional properties required to combine lawfully with time. The QALY therefore lacks a defensible dimensional foundation.

The role of the econometric model is particularly revealing. Regression analysis estimates relationships among variables. It does not create measurement. The dependent variable in the EQ-5D tariff model is the TTO valuation assigned by respondent to health state. This value is a preference score. It is not a demonstrated measure of health, quality of life, utility or any other single attribute. The regression model decomposes variation in that preference score using dummy variables representing EQ-5D levels. It does not transform the preference score into a ratio measure.

The coefficients estimated by the regression model inherit the measurement properties of the dependent variable. If the TTO valuation lacks demonstrated ratio properties, then the coefficients cannot acquire ratio properties through estimation. They remain average conditional effects expressed in the units of the dependent variable. This is the critical point. The econometric model

can produce numbers. It can produce statistically significant coefficients. It can improve prediction. It cannot create measurement properties that are absent from the TTO values themselves.

The dummy variables in the EQ-5D tariff model are not measures. They are classifications. They identify whether a health-state description includes Mobility Level 2, Pain Level 3 or Anxiety Level 2. The coefficient attached to such a dummy variable is not a measure of mobility, pain or anxiety. It is an estimated average conditional effect on the TTO valuation. The subsequent transformation of that coefficient into a decrement weight is therefore a conceptual leap. A statistical effect becomes a scoring weight. A scoring weight then becomes part of an alleged utility measure. This transition is assumed rather than justified.

The results also show strong endorsement of the false proposition that summations of subjective instrument responses are ratio measures, with a probability of 0.80 and a normalized logit of 1.40. This is consistent with the broader utility framework, where numerical aggregation is treated as though it can create measurement. Yet summing responses, coefficients or decrements does not create a ratio measure. Measurement properties must exist before arithmetic is undertaken. Arithmetic cannot manufacture a meaningful zero, unidimensionality or invariance.

The low endorsement of the axioms of representational measurement is among the most serious findings. The statement that meeting these axioms is required for arithmetic receives a probability of 0.05 and a normalized logit of -2.50. This indicates that the EQ-5D knowledge base does not recognize the formal standards required for lawful quantitative claims. The axioms of representational measurement exist precisely to distinguish numerical assignment from measurement. Without these axioms, any scoring system can be mistaken for a measure.

The absence of Rasch measurement is equally striking. The statements that there are only two classes of measurement, linear ratio and Rasch logit ratio, and that transforming subjective responses requires Rasch rules both receive endorsement probabilities of 0.05 and normalized logits of -2.50. The statement that the Rasch logit ratio scale is the only basis for assessing therapy impact for latent traits also receives 0.05 and -2.50. These results indicate that the EQ-5D knowledge base does not recognize the measurement framework required for latent attributes.

This omission is important because many of the attributes relevant to therapy assessment are latent. Pain, fatigue, anxiety, depression, physical functioning and need fulfilment cannot be directly observed. They must be inferred from responses. Rasch measurement provides the required framework for transforming those responses into lawful measures. The EQ-5D framework does not take this path. Instead, it uses health-state descriptions, TTO valuations, econometric coefficients and utility algorithms. The result is a numerical score, not a demonstrated latent measure.

The statement that the outcome of interest for latent traits is possession of that trait receives an endorsement probability of only 0.05 and a normalized logit of -2.50. This confirms the absence of the Rasch concept of attribute possession. In Rasch measurement, the objective is to locate persons and items on a common latent continuum and to estimate the degree to which a person possesses the latent attribute. The EQ-5D framework bypasses this entire measurement logic. It

does not ask whether a respondent possesses more or less of a latent attribute. It asks how a health-state description is valued and then converts that valuation into a utility score.

The reference-case implications are direct. The statement that reference-case simulations generate falsifiable claims receives an endorsement probability of 0.85 and a normalized logit of 1.75, despite being false. This reflects the central role of EQ-5D utilities and QALYs in simulation-based HTA. Yet simulation models can only transform their inputs. If the inputs are not lawful measures, the outputs cannot be lawful measures. If utility scores fail measurement standards, then QALYs fail. If QALYs fail, then cost-per-QALY estimates fail. The reference case therefore inherits the measurement deficiencies of the EQ-5D utility framework.

The statement 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 suggests that the EQ-5D knowledge base does not recognize the downstream consequences of measurement failure. Once utility scores are accepted as measures, the rest of the framework follows automatically. QALYs are calculated. Cost-effectiveness ratios are estimated. Simulation models are populated. Yet the initial measurement failure remains embedded throughout the sequence.

The interrogation also indicates weak endorsement of the rejection of non-falsifiable claims, with a probability of 0.20 and a normalized logit of -1.40. This is consistent with the reference-case tradition, where simulated lifetime claims are treated as meaningful decision inputs despite their lack of direct empirical falsifiability. Scientific claims should be capable of being tested, replicated and potentially refuted. A simulated cost-per-QALY estimate based on utilities of uncertain measurement status does not meet this requirement.

The overall pattern is therefore clear. The EQ-5D knowledge base strongly supports the arithmetic structures required for utility-based HTA while weakly recognizing the measurement standards required to justify those structures. It accepts TTO valuations, econometric tariff construction, additive utility scoring and QALY generation. It does not demonstrate that the quantities involved satisfy ratio measurement, dimensional homogeneity or representational measurement.

The importance of this finding extends beyond the EQ-5D instruments themselves. The EQ-5D is one of the most widely used instruments in HTA. It supplies utility values for QALY construction and reference-case modelling across numerous jurisdictions. If the EQ-5D knowledge base is characterized by measurement inversion, then the same inversion is transmitted into HTA agencies, academic research centers, journals and educational programs. This helps explain why similar patterns have been observed across national and institutional interrogations.

The conclusion is unavoidable. The EQ-5D framework provides numbers. It does not demonstrate lawful measurement. TTO valuations are preference scores. Regression coefficients are average conditional effects. Decrement weights are transformed statistical parameters. Utility scores are outputs of scoring algorithms. QALYs are products of multiplying time by utility scores whose measurement properties remain unproven. At each stage arithmetic is performed without first establishing the conditions required for measurement.

The EQ-5D knowledge base therefore exemplifies measurement inversion. It begins with preference and ends with arithmetic, while never establishing a lawful ratio measure. The result is not merely a problem for one instrument. It is a problem for the entire utility/QALY/reference-case framework that depends upon it. If HTA is to recover scientific credibility, it must abandon the assumption that utility scores are measures and return to first principles: identify the attribute, determine whether it is manifest or latent, construct the appropriate ratio measure and only then undertake admissible arithmetic.

CONCLUSION

The conclusion is unavoidable. The EQ-5D framework provides numbers. It does not demonstrate lawful measurement. TTO valuations are preference scores. Regression coefficients are average conditional effects. Decrement weights are transformed statistical parameters. Utility scores are outputs of scoring algorithms. QALYs are products of multiplying time by utility scores whose measurement properties remain unproven. At each stage arithmetic is performed without first establishing the conditions required for measurement.

The EQ-5D knowledge base therefore exemplifies measurement inversion. It begins with preference and ends with arithmetic, while never establishing a lawful ratio measure. The result is not merely a problem for one instrument. It is a problem for the entire utility/QALY/reference-case framework that depends upon it. If HTA is to recover scientific credibility, it must abandon the assumption that utility scores are measures and return to first principles: identify the attribute, determine whether it is manifest or latent, construct the appropriate ratio measure and only then undertake admissible arithmetic.

Whether these findings are regarded as definitive evidence for HTA closure is ultimately a matter for individual judgment. Nevertheless, the interrogation results provide substantial support for that conclusion. The endorsement profile demonstrates a consistent failure to recognize the requirements of representational measurement while simultaneously endorsing the arithmetic foundations of utility construction, QALY generation and reference-case simulation modelling. The result is a methodological framework in which numerical manipulation is accepted as a substitute for measurement. Utilities, QALYs and simulation outputs are treated as quantitative measures despite the absence of evidence that the conditions necessary for lawful measurement have been satisfied.

The implications are difficult to avoid. If time trade-off valuations are not demonstrated ratio measures, then econometric coefficients derived from those valuations cannot acquire ratio-scale properties through statistical estimation. If those coefficients are not lawful measures, then their transformation into decrement weights is unjustified. If decrement weights do not represent quantities of a common attribute, then their arithmetic aggregation into utility scores lacks a defensible measurement foundation. If utility scores are not ratio measures, then their use as discount factors for time is inadmissible. The QALY therefore becomes an arithmetic construction rather than a measure, while reference-case simulation models become exercises in numerical storytelling rather than empirical science.

Whether abandoning four decades of measurement inversion is an easy decision is another matter. Entire disciplines can become committed to methodological conventions that persist long after their scientific foundations have been called into question. Utilities, QALYs and reference-case simulations have become deeply embedded in HTA agencies, academic programs, journals and reimbursement systems throughout the world. Institutional acceptance, however, is not evidence of validity. Scientific standards are not determined by repetition, consensus or longevity. They are determined by conformity to the requirements of measurement and the capacity to support evaluable, replicable and falsifiable claims.

Importantly, the rejection of the utility framework does not leave HTA without an alternative. On the contrary, a straightforward and scientifically coherent alternative already exists. The assessment of therapy impact begins with the identification of the attribute of interest. If the attribute is manifest and directly observable, then it requires a linear ratio measure. If the attribute is latent and cannot be directly observed, then it requires a Rasch logit ratio measure. These are the only two measurement frameworks capable of supporting lawful quantitative claims. Once valid measures are established, claims regarding therapy impact can be evaluated in real populations over specified time horizons and subjected to replication and falsification.

The choice facing HTA is therefore not between the existing framework and methodological uncertainty. It is between arithmetic without measurement and measurement before arithmetic. The evidence presented here suggests that the future of HTA lies not in increasingly sophisticated utility algorithms, QALYs and simulation models, but in a return to the principles of measurement that govern every other quantitative science.

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