MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

UNITED KINGDOM: DECONSTRUCTING THE EPISTEMIC KNOWLEDGE BASE OF THE EQ-5D-3L AND EQ-5D-5L INSTRUMENTS

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

LOGIT WORKING PAPER No 43 FEBRUARY 2026

www.maimonresearch.com

Tucson AZ

ABSTRACT

This paper examines the EQ-5D-3L and EQ-5D-5L instruments not as descriptive questionnaires or psychometric tools, but as epistemic objects embedded within the United Kingdom health technology assessment (HTA) system. In HTA practice, instruments are commonly treated as neutral measurement devices whose numerical properties are fixed at development and merely applied thereafter. This assumption is rejected. Instruments acquire numerical authority through use. Numbers become treated as quantities when professional communities routinely subject them to arithmetic operations, regardless of whether the axioms required for measurement are satisfied.

The EQ-5D occupies a distinctive position within HTA because it functions as the primary descriptive and valuation foundation of the reference-case framework. Its outputs are routinely analyzed as outcomes, transformed into utilities, multiplied by time to generate quality-adjusted life years, and aggregated for policy decision-making. Over time, this repeated application has conferred an appearance of quantitative legitimacy that is rarely examined at the level of measurement theory. The transition from EQ-5D-3L to EQ-5D-5L is commonly interpreted as methodological refinement, yet whether such expansion alters the epistemic status of the instrument remains unexamined.

To address this gap, the paper defines the EQ-5D knowledge base as an epistemic corpus encompassing developers, users, national value sets, methodological guidance, educational materials, and analytic infrastructure. This corpus is interrogated using a reduced canonical diagnostic grounded in representational measurement theory and Rasch principles. The diagnostic evaluates whether the knowledge base recognizes the axioms required for quantitative measurement, including unidimensionality, invariant transformation, scale-type coherence, and arithmetic admissibility. Endorsement probabilities are classified and transformed into normalized logits to reveal structural patterns of epistemic reinforcement or absence.

The results display a coherent and invariant profile. Core measurement axioms receive no positive reinforcement, while assumptions enabling arithmetic treatment of utilities remain normalized. Rasch requirements for latent-trait measurement are entirely absent, consistent with the multiattribute, preference-based architecture of the EQ-5D system. The expansion from three to five response levels does not alter these foundational commitments. The knowledge base exhibits non-possession of measurement principles rather than misunderstanding or dispute.

The analysis demonstrates that the EQ-5D functions as a valuation framework embedded within an epistemic system that treats valuation as measurement. Because the reference-case framework depends upon EQ-5D outputs, non-measurement is inherited rather than corrected at the policy level. The paper concludes that refinement of descriptive detail cannot resolve this condition and that restoring measurement as a precondition for arithmetic is necessary if HTA is o generate quantitative knowledge rather than numerical appearance. EQ-5D and similar multiattribute instruments should be abandoned for other than purely descriptive purposes.

FOREWORD

HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

This Logit Working Paper series documents a finding as extraordinary as it is uncomfortable: health technology assessment (HTA), across nations, agencies, journals, and decades, has developed as a global system of non-measurement. It speaks the language of numbers, models, utilities, QALYs, "value for money," thresholds, discounting, incremental ratios, extrapolations, and simulations. It demands arithmetic at every turn, multiplication, division, summation, aggregation, discounting, yet it never once established that the quantities to which these operations are applied are measurable. HTA has built a vast evaluative machinery on foundations that do not exist. The probabilities and normalized logits in the country reports that follow provide the empirical confirmation of this claim. They show, with unsettling consistency, that HTA presents a world of measurement failure.

The objective of this study is to evaluate the EQ-5D-3L and EQ-5D-5L instruments against the axioms of representational measurement using the 24-item canonical statement diagnostic. The assessment is intended to determine whether these instruments, as deployed within health technology assessment, possess the measurement properties required to support arithmetic operations central to cost-utility analysis, including the construction of QALYs and cost-effectiveness ratios. Particular attention is given to unidimensionality, admissible scale transformations, ratio requirements for multiplication, dimensional homogeneity, and the role of Rasch measurement in transforming ordinal responses for latent traits.

The canonical assessment demonstrates a uniform and decisive pattern of measurement failure for both EQ-5D-3L and EQ-5D-5L. Statements asserting foundational measurement requirements—unidimensionality, the priority of measurement over arithmetic, the necessity of ratio scales for multiplication, and the requirement of Rasch rules for transforming ordinal responses—collapse to floor or near-floor logit values, indicating effective non-possession within the instruments' justificatory framework. Conversely, false propositions central to the HTA memeplex, including the treatment of EQ-5D indices as interval or ratio measures, the dimensional coherence of QALYs, and the legitimacy of reference-case simulation outputs, are strongly endorsed. The results show that neither instrument operates within a measurement-valid framework; both function as preference-weighted scoring systems whose numerical outputs are treated as quantities without satisfying the axioms required for lawful measurement or empirical falsifiability.

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

The foundational theory that establishes *when* and *whether* a set of numbers can be interpreted as measurements came with the publication of Krantz, Luce, Suppes, and Tversky's *Foundations of Measurement* (1971) [2] . Representational Measurement Theory (RMT) formalized the axioms under which empirical attributes can be mapped to numbers in a way that preserves structure. Measurement, in this framework, is not an act of assigning numbers for convenience, it is the discovery of a lawful relationship between empirical relations and numerical relations. The axioms of additive conjoint measurement, homogeneity, order, and invariance specify exactly when interval scales exist. RMT demonstrated once and for all that measurement is not optional and not a matter of taste: either the axioms hold and measurement is possible, or the axioms fail and measurement is impossible. Every major construct in HTA, utilities, QALYs, DALYs, ICERs, incremental ratios, preference weights, health-state indices, fails these axioms. They lack unidimensionality; they violate independence; they depend on aggregation of heterogeneous attributes; they collapse under the requirements of additive conjoint measurement. Yet HTA proceeded, decade after decade, without any engagement with these axioms, as if the field had collectively decided that measurement theory applied everywhere except in the evaluation of therapies.

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

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

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

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

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

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

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

Paul C Langley, Ph.D

Email: langleylapaloma@gmail.com

DISCLAIMER

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

1. INTERROGATING THE LARGE LANGUAGE MODEL

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

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

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

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

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

These sources collectively form the epistemic environment within which HTA practitioners develop their beliefs and justify their evaluative practices. The boundary of interrogation is thus not the whole of medicine, economics, or public policy, but the specific textual ecosystem that sustains HTA reasoning. . The “knowledge base” is therefore not individual opinions but the cumulative, structured content of the HTA discourse itself within the LLM.

THE EQ-5D KNOWLEDGE BASE

In health technology assessment, instruments are commonly treated as neutral devices that exist independently of the systems in which they are used. Once developed, they are assumed to generate numerical outputs whose meaning is fixed, transportable, and stable across applications. Under this view, the analytical task lies not in questioning whether numbers measure anything, but in deciding how best to deploy them. This assumption is deeply misleading. Instruments do not acquire numerical legitimacy through construction alone. They acquire it through use.

Numbers become treated as quantities only when a community repeatedly subjects them to arithmetic operations and accepts the results as meaningful. That acceptance is rarely explicit. It emerges through routine practice, institutional endorsement, educational transmission, and methodological repetition. Over time, what begins as assumption becomes belief, and belief becomes norm. Once embedded in this way, the question of whether an instrument truly measures anything at all no longer appears as a scientific problem. It becomes epistemically invisible.

The EQ-5D-3L and EQ-5D-5L instruments provide a particularly revealing illustration of this process [5] [6] Originally developed as descriptive systems intended to classify health states, the EQ-5D instruments now occupy a central position within the United Kingdom’s health technology assessment environment. Their numerical outputs are treated as utilities, incorporated into economic models, multiplied by time to generate quality-adjusted life years, and aggregated to support population-level decision-making. These operations are not occasional or experimental. They are routine. Through repetition, numerical treatment has come to substitute for measurement justification; failing to meet the axioms of representational measurement ..

The object of analysis is therefore not the EQ-5D instrument in isolation, nor the intentions of its developers, nor the technical details of its valuation protocols. The object of analysis is the epistemic system within which EQ-5D outputs function as if they were quantitative measures. This paper refers to that system as the user epistemic system; it does not confer measurement legitimacy.

The user epistemic system consists of the collective practices through which EQ-5D values are interpreted as meaningful numbers. It includes researchers who analyze EQ-5D data, reviewers who evaluate manuscripts, editors who publish results, health technology assessment agencies that accept EQ-5D utilities as inputs, educators who train analysts in their application, and software environments that embed scoring algorithms. Together, these actors form a distributed but coherent knowledge base. No single participant determines its structure, yet each reinforces it.

Importantly, this system does not operate through explicit agreement about measurement theory. There is no formal declaration that EQ-5D utilities satisfy the axioms required for quantification. Instead, authority arises through practice. Means are reported. Differences are compared. Utilities are multiplied by time. Each step appears innocuous in isolation. Collectively, they construct a powerful presumption: if the numbers are used as quantities, they must be quantities. Use becomes evidence; repetition becomes validation.

Within such a system, epistemic responsibility is diffuse. Developers may point to widespread adoption. Users may point to methodological guidance. Agencies may point to precedent. Educators may point to accepted curricula. Each component defers foundational justification to another. The result is epistemic closure: numerical practice persists without ever encountering the conditions that would authorize or prohibit it.

This distinction between ignorance and non-possession is critical. The issue is not that users of the EQ-5D fail to understand measurement theory. It is that measurement theory does not function as a governing authority within the system. Where axioms are not recognized, they cannot constrain practice. Arithmetic proceeds not because rules are violated, but because the rules are absent.

The transition from EQ-5D-3L to EQ-5D-5L exemplifies this dynamic. The expansion of response levels is commonly interpreted as methodological progress, improving sensitivity and discrimination. Yet increased descriptive granularity does not address the prior question of whether the resulting numbers represent a measurable attribute. Refinement operates entirely within an already accepted numerical framework. It presupposes measurement rather than establishing it.

This paper therefore does not ask whether EQ-5D instruments are useful, convenient, or widely adopted. It asks a more fundamental question: does the knowledge base that authorizes their numerical use contain the axioms required for measurement? By treating EQ-5D as an epistemic object embedded within a user system rather than as a technical artifact, the analysis shifts attention from instrument performance to the conditions that make numerical authority possible.

Having established that numerical authority arises through use rather than construction alone, the next task is to define what constitutes the knowledge base of the EQ-5D system. This cannot be limited to the original development papers of the EuroQol Group, nor can it be confined to formal descriptions of the instrument. Once released into applied domains, an instrument becomes embedded within a far broader epistemic environment. Its authority is sustained not by its design history, but by the network of texts, practices, institutions, and routines that treat its outputs as quantities. The EQ-5D knowledge base must therefore be understood as an epistemic corpus.

This corpus includes the foundational publications introducing the EQ-5D-3L and later the EQ-5D-5L. These texts define the descriptive architecture of the instrument: a multiattribute classification system covering mobility, self-care, usual activities, pain/discomfort, and anxiety/depression, each expressed through ordered response levels. From the outset, the instrument was framed not as a direct measure of health, but as a system for describing health states to be valued through preference elicitation. This framing is crucial. It establishes valuation, rather than measurement, as the conceptual foundation of the instrument.

However, the authority of the EQ-5D does not persist because these development papers are repeatedly interrogated. It persists because subsequent users treat the resulting numerical outputs as if they were quantitative measures. The dominant component of the knowledge base therefore lies in applied research. Thousands of clinical trials, observational studies, and economic evaluations report EQ-5D values as outcomes. Means are calculated, changes interpreted, and group differences compared. These operations are typically presented without discussion of scale type, unidimensionality, or invariance. Yet the absence of such discussion is itself epistemically powerful. It signals that justification is unnecessary.

Health technology assessment agencies constitute a second and particularly influential layer of the EQ-5D knowledge base. Within the United Kingdom, NICE guidance explicitly endorses EQ-5D as the preferred instrument for estimating health-related quality of life. This endorsement confers institutional authority. It does not arise from demonstration that EQ-5D values satisfy the axioms of measurement, but from their consistency with the reference-case framework. Once incorporated into official guidance, the instrument's numerical status becomes administratively secured rather than theoretically established.

Methodological documents further reinforce this authority. Reference-case manuals, submission templates, and technical support documents routinely specify EQ-5D utilities as required or preferred inputs. These texts treat utilities as interchangeable numerical entities, abstracted from their descriptive origins. The instrument becomes a standardized component of analytic workflow rather than an object of epistemic scrutiny. At this stage, the question of what kind of numbers EQ-5D produces is displaced by the assumption that numbers are required.

Education plays a central role in reproducing this knowledge base. In graduate training programs and professional short courses, students are taught how to apply EQ-5D utilities in modeling exercises. They learn to compute quality-adjusted life years, to compare incremental cost-effectiveness ratios, and to interpret numerical thresholds. Rarely are they taught to interrogate whether the utilities themselves possess the properties required for arithmetic. By the time analysts enter practice, numerical legitimacy has already been internalized. The instrument is encountered not as a theoretical proposition, but as a given.

The epistemic reach of the EQ-5D knowledge base extends further through analytic infrastructure. Software packages, economic models, and spreadsheet templates embed EQ-5D scoring algorithms and national value sets. Once encoded, assumptions become invisible. Users interact with outputs without encountering the conceptual premises that authorize their numerical treatment. In this way, epistemic commitment is no longer expressed through argument or citation, but through automation.

The introduction of the EQ-5D-5L illustrates how the knowledge base absorbs modification without altering its foundations. The expansion from three to five response levels is presented as improved sensitivity and reduced ceiling effects. Yet this refinement does not alter the underlying epistemic architecture. The instrument remains multiattribute. Valuation remains preference-based. Utilities continue to allow negative values. Arithmetic compatibility with the QALY remains assumed rather than demonstrated. The proliferation of versions therefore signals elaboration within a fixed belief system rather than epistemic change.

Crucially, the EQ-5D knowledge base is not unified by explicit theoretical agreement. There is no authoritative text asserting that EQ-5D utilities satisfy representational measurement axioms. Instead, unity emerges through coordinated silence. Measurement theory is not debated because it is not invoked. Scale properties are not defended because they are not questioned. The absence of foundational discourse functions as a stabilizing mechanism.

This distributed structure explains the resilience of the EQ-5D system. Developers can point to widespread use. Users can point to agency guidance. Agencies can point to precedent. Educators can point to standard curricula. Each component defers epistemic responsibility to another. The result is a closed loop in which numerical authority circulates without ever encountering measurement constraints.

Defining the EQ-5D knowledge base in this way is essential for the analysis that follows. The purpose of interrogation is not to assess individual publications or intentions, but to determine whether the epistemic environment as a whole recognizes measurement axioms as governing rules. Only by treating the instrument as embedded within this broader corpus can its numerical status be meaningfully evaluated.

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

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.

RESULTS AND DISCUSSION

The EQ-5D-3L and EQ-5D-5L occupy a privileged position in global health technology assessment. They are not simply instruments among many; they function as the canonical gateway through which "health-related quality of life" is converted into the numbers required for cost-utility analysis. The EuroQol Group's descriptive systems of five dimensions, with three levels in the 3L and five levels in the 5L are widely treated as if they instantiate a measurement architecture capable of supporting arithmetic operations that culminate in QALYs and cost-per-QALY ratios. The question, however, is not whether these instruments are widely used, convenient, or institutionally embedded. The question is whether they can be defended as measurement systems in the sense required by representational measurement theory and by normal science: unidimensional, invariant, and governed by admissible transformations that license arithmetic. On that standard, EQ-5D-3L and EQ-5D-5L do not merely fall short. They exemplify a multi-layered measurement failure that is then stabilized and protected by the broader HTA evaluative framework (Tables 1 and 2). .

TABLE 1: ITEM STATEMENT, RESPONSE, ENDORSEMENT AND NORMALIZED LOGITS EQ-5D-3L

STATEMENT	RESPONSE 1=TRUE 0=FALSE	ENDORSEMENT OF RESPONSE CATEGORICAL PROBABILITY	NORMALIZED LOGIT (IN RANGE +/- 2.50)
INTERVAL MEASURES LACK A TRUE ZERO	1	0.10	-2.20
MEASURES MUST BE UNIDIMENSIONAL	1	0.05	-2.50
MULTIPLICATION REQUIRES A RATIO MEASURE	1	0.05	-2.50
TIME TRADE-OFF PREFERENCES ARE UNIDIMENSIONAL	0	0.90	+2.20
RATIO MEASURES CAN HAVE NEGATIVE VALUES	0	0.90	+2.20
EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES	0	0.95	+2.50
THE QALY IS A RATIO MEASURE	0	0.95	+2.50
TIME IS A RATIO MEASURE	1	0.90	+2.20
MEASUREMENT PRECEDES ARITHMETIC	1	0.05	-2.50
SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES	0	0.90	+2.20
MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC	1	0.05	-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	0.90	+2.20
THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE	0	0.95	+2.50
CLAIMS FOR COST-EFFECTIVENESS FAIL THE	1	0.10	-2.20

AXIOMS OF REPRESENTATIONAL MEASUREMENT			
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	0.90	+2.20
THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.40	-0.45
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	0.80	+1.40
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

TABLE 2: ITEM STATEMENT, RESPONSE, ENDORSEMENT AND NORMALIZED LOGITS EQ-5D-5L

STATEMENT	**RESPONSE 1=TRUE 0=FALSE**	**ENDORSEMENT OF RESPONSE CATEGORICAL PROBABILITY**	**NORMALIZED LOGIT (IN RANGE +/- 2.50)**
INTERVAL MEASURES LACK A TRUE ZERO	1	0.10	-2.20
MEASURES MUST BE UNIDIMENSIONAL	1	0.05	-2.50
MULTIPLICATION REQUIRES A RATIO MEASURE	1	0.05	-2.50
TIME TRADE-OFF PREFERENCES ARE UNIDIMENSIONAL	0	0.90	+2.20
RATIO MEASURES CAN HAVE NEGATIVE VALUES	0	0.95	+2.50
EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES	0	0.90	+2.20
THE QALY IS A RATIO MEASURE	0	0.95	+2.50

TIME IS A RATIO MEASURE	1	0.90	+2.20
MEASUREMENT PRECEDES ARITHMETIC	1	0.05	-2.50
SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES	0	0.90	+2.20
MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC	1	0.05	-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	0.90	+2.20
THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE	0	0.95	+2.50
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	0.90	+2.20
THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.40	-0.45
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	0.80	+1.40
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 first point is architectural. Both EQ-5D versions are explicitly multiattribute: they decompose "health" into multiple dimensions (mobility, self-care, usual activities, pain/discomfort, anxiety/depression) and then represent a respondent's status as a profile across those dimensions. Whatever their merits as a descriptive classification, the profile is not a measure of a single attribute. Unidimensionality is not a negotiable refinement; it is the defining requirement for any quantity that purports to represent magnitude along one dimension. Multiattribute instruments begin by denying that requirement. They implicitly assert that an assemblage of qualitatively distinct attributes can be collapsed into a single number without loss of meaning and without violating the conditions under which numerical representation is valid. That is the first and most basic contradiction: a multidimensional classification is being pressed into service as if it were a unidimensional measure.

The second point is that the conversion from a descriptive profile to an index value is not a transformation that creates measurement; it is a preference-weighting exercise. EuroQol values are "attached" to profiles using weights that reflect average preferences, anchored to 1 for full health and 0 for dead, with negative values permitted for states judged worse than dead. This anchoring is not a demonstration of a true zero; it is an imposed convention driven by the requirements of economic evaluation. The EuroQol's own terminology makes this clear: anchoring at 1 and 0 is "required by their use in economic evaluation," and negative values reflect states regarded as "worse than dead." The conceptual problem is immediate. A constructed scale anchored by convention to facilitate downstream arithmetic is not thereby granted the properties needed to support that arithmetic. A label "0=dead" does not create a natural origin. A true zero is not a policy decision; it is an empirical property of the attribute being measured. If the attribute is not itself measurable as an extensive quantity, the anchoring is cosmetic: it enables multiplication in spreadsheets, not multiplication in science.

The third point is that the EQ-5D index is routinely treated as if it were an interval measure, sometimes even as if it were a ratio measure, when it is neither. Even the claim that preference algorithms yield "interval" values is not established by decree or by the sophistication of the valuation protocol. It would require a demonstration of constant unit meaning across the scale and invariance of comparisons under admissible transformations. Instead, what is offered is a scoring algorithm mapping profiles to numbers, with the numbers interpreted as if they were distances on a quantitative continuum. Anchoring at dead and permitting negative values further undermines any ratio interpretation, because it invites the false inference that values can be meaningfully compared as ratios, or that a negative "utility" is a permissible ratio magnitude. The EuroQol documentation explicitly acknowledges negative values. In measurement terms, negative values are not inherently disallowed on all ratio scales (temperature in Celsius is a classic interval example, while Kelvin has a true zero), but that observation misses the point: the EQ-5D index is not a demonstrated ratio measure of a single attribute in the first place. The issue is not merely sign; it is the absence of the measurement structure that would make sign meaningful.

The fourth point is dimensional. Even if one were to grant, hypothetically, that the EQ-5D index were an interval scale, the QALY construction treats it as if it were at least interval with meaningful multiplication by time. Standard HTA descriptions state that QALYs are calculated by assigning a utility to health states and multiplying by time spent in those states. Multiplication, however, is not a universal right. It is an operation that requires ratio properties and dimensional coherence. If time is ratio (it is), then multiplying time by a non-ratio, non-unidimensional value does not yield a quantity with a coherent dimensional interpretation. At best, it yields a composite score. At worst, it yields a policy token falsely treated as a physical-like magnitude. The critical point is that the QALY is not rescued by the fact that one component (time)has ratio properties. A lawful product requires that both components be quantities of the right type and that the resulting product correspond to a meaningful empirical attribute. Otherwise, it is merely a numerical artifact.

The fifth point is that the 5L does not solve these problems. The EQ-5D-5L was introduced to improve sensitivity and reduce ceiling effects compared to the 3L. This is a legitimate psychometric and descriptive objective: more response levels can increase discrimination and reduce clustering at "no problems." Yet improving descriptive sensitivity is not the same as establishing measurement. A finer classification does not become unidimensional by adding categories. A preference-weighted index does not become a ratio measure by becoming more responsive. A scoring algorithm does not become an admissible transformation simply because it yields more gradations. The 5L improves the instrument's ability to represent differences in profiles; it does not create the measurement properties required for arithmetic on a single attribute. The epistemic status remains unchanged: a multiattribute descriptive system is still being collapsed into a single index through preference weights, and that index is still treated as a quantitative measure.

These points are captured in the canonical assessment tables. Statements that define what measurement requires such as unidimensionality, the priority of measurement over arithmetic, the ratio requirement for multiplication, and the Rasch requirement for transforming ordinal responses to interval scales collapse to the floor (−2.50) or adjacent values. The interpretation is not that these propositions are controversial; they are axiomatic in measurement science. The interpretation is that the EQ-5D knowledge base, understood as the set of claims, conventions, and justificatory narratives surrounding the instruments, does not operationalize these axioms as constraints. They are not possessed. Instead, the instruments presuppose that quantification can be achieved by scoring and weighting and that downstream arithmetic (QALYs; cost per QALY) is thereby licensed. That is exactly the "closure" characteristic of a memeplex: the framework does not engage measurement axioms because it does not need them to reproduce itself.

The Rasch items in particular diagnose the core category error. When responses are ordinal, as they are in EQ-5D dimension levels, any claim to interval measurement requires a lawful transformation model. Rasch measurement provides the only model that can, under strict conditions, convert ordinal observations into linear measures with invariance properties. The EQ-5D family does not proceed by Rasch transformation. It proceeds by preference weighting and aggregation across attributes. That is not a technical alternative; it is a different activity. Rasch is measurement; preference weighting is valuation. The canonical logit floor values on Rasch statements are therefore not incidental. They reflect the deep structural exclusion of measurement science from the justificatory logic of multiattribute utilities. In this respect, EQ-5D is not merely an instrument used by the HTA memeplex; it is one of the memeplex's principal carriers.

The falsifiability items add the final closure mechanism. The EQ-5D index is not used to make testable claims about patient outcomes; it is used to populate reference case simulations whose outputs are cost-per-QALY ratios. Those simulations are treated as evidence in deliberative processes even when they cannot be refuted in the Popperian sense, because their outputs depend on assumptions, horizon choices, and model structures that can always be revised without conceding failure. The NICE description of QALYs as time multiplied by utility is explicit, and it reveals the mechanism: once the index is accepted as a quantity, the multiplication and aggregation steps follow automatically. The system is therefore not merely a set of instruments; it is an evaluative assembly line. EQ-5D produces utilities; utilities produce QALYs; QALYs support cost-per-QALY; cost-per-QALY is used for pricing and access decisions. At no stage is the chain forced to confront measurement axioms as binding constraints.

A frequent defense is that EQ-5D values are "anchored" for use in economic evaluation, and that anchoring at dead is a practical requirement. EuroQol states precisely that. Yet "required for economic evaluation" is not a measurement argument; it is a policy argument. It reveals the administrative origin of the instrument's numeric form. The instrument exists to provide a single index compatible with an evaluative calculus, not to measure an empirical attribute according to the requirements of measurement science. In that sense, the anchoring statement is inadvertently candid. It admits that the scale is structured to satisfy the needs of a downstream decision algorithm. It does not claim, because it cannot, that the scale's numeric operations are licensed by the structure of an empirical attribute.

The same applies to the widespread acceptance of negative values ("worse than dead"). Negative values are not merely a curiosity; they are a structural signal that the index is not behaving as a measure of a single extensive attribute with a natural origin. The EuroQol terminology makes the presence of negative values explicit. This is regularly defended as reflecting preferences: some health states are judged worse than death. But preference intensity is not measurement of health. It is measurement of valuation, and even then it is not measurement in the representational sense unless the conditions for such measurement are demonstrated. The system moves between "health" and "preference about health" as if these were interchangeable. That conceptual slippage is one reason the instruments can be treated as measures without meeting measurement requirements: the object being quantified is constantly re-described to fit the needs of the framework.

The 5L's improvements (reduced ceiling effects; improved sensitivity) should be acknowledged as descriptive refinements, but they do not touch the foundational critique. Sensitivity gains are not equivalent to lawful quantification. A better thermometer does not justify measuring length with it; likewise, a more sensitive classification does not justify treating a weighted index as a ratio measure. The improvement narrative is therefore perfectly compatible with continued measurement failure. Indeed, it can deepen it: by making the index appear more responsive, it may increase confidence in downstream arithmetic without addressing admissible transformations, unidimensionality, or invariance. The refinements improve the instrument's rhetorical plausibility as "measurement" while leaving its measurement status unchanged.

What, then, is the scientific standing of EQ-5D-3L and EQ-5D-5L within HTA? They can be defended as standardized descriptive frameworks that allow comparisons of reported problems across a fixed set of dimensions. They can be defended as tools for eliciting structured self-reports and for mapping those reports into preference-based indices useful for certain administrative comparisons. But they cannot be defended as measurement systems that yield quantities appropriate for multiplication, aggregation, or the construction of cost-effectiveness ratios. The leap from classification to measurement is not a matter of improved valuation studies, larger samples, better regression models, or more culturally tailored value sets. The problem is architectural: multidimensional descriptions are collapsed into a single number through valuation, and that number is then used as if it were a unidimensional quantitative measure.

The implication is not that EQ-5D instruments should be "improved" and retained as the basis of QALYs. The implication is that they belong to a class of constructs that must be reclassified: they are scoring and valuation systems, not measurement systems. In a scientific evaluative framework, their outputs would be treated as inputs into deliberation as ordinal or classificatory information, not as quantities on which one performs multiplication and ratio arithmetic. If HTA seeks to recover scientific accountability, the successor framework is not a new generation of multiattribute utilities. It is a portfolio of single-attribute claims with lawful measurement: linear ratio measures for manifest outcomes and Rasch logit ratio measures for latent traits, each tied to protocols that expose claims to the risk of failure.

In that context, EQ-5D-3L and EQ-5D-5L have a clear future only if their role is constrained to what they actually are: standardized descriptive systems. Their role as the engine of the QALY calculus, and as the numerical foundation for simulated cost-per-QALY claims, should be regarded as an historical artifact of an evaluative memeplex that prioritized closure over measurement. The instruments did not "reach the end" of a scientific life in HTA; they never possessed one. They are emblematic of a long-standing substitution of valuation for measurement and of a global evaluative practice that has operated without binding measurement constraints.

AN EMBARRASSMENT OF RICHES

If the EQ-5D instruments cannot legitimately go beyond the description of health states, the question facing analysts becomes unavoidable: how can one continue to present multiattribute utility instruments as a scientific basis for closed cost-effectiveness claims? The dilemma is not merely technical; it is professional and epistemic. Analysts are asked to report to clients, manufacturers, payers, or agencies, using tools that are deeply embedded in HTA practice, while evidence accumulates that those tools do not meet the requirements of measurement needed to support the claims being made.

The standard defense is historical. Multiattribute instruments have been in continuous use for over forty years. Tens of thousands of peer-reviewed publications report QALYs, cost-per-QALY ratios, and modeled therapy impacts built on EQ-5D utilities and similar constructs. This volume of output is presented as proof of validity. The implicit argument is sociological rather than scientific: a method so widely used, reviewed, and institutionalized cannot be fundamentally wrong. Yet this is precisely the argument that normal science rejects. Longevity, replication of practice, and publication count are not criteria of truth. They are indicators of stability.

The analyst's predicament arises because EQ-5D instruments do provide something of value: standardized descriptions of reported health states across a limited set of dimensions. That descriptive function is not in dispute. What is in dispute is the subsequent transformation of those descriptions into single numerical indices treated as interval or ratio measures, multiplied by time, aggregated across individuals, and embedded in simulation models to produce decisive cost-effectiveness claims. Once the descriptive role is acknowledged as the instrument's epistemic ceiling, the downstream arithmetic becomes indefensible. Description does not license measurement, and classification does not authorize multiplication.

From the analyst's perspective, however, abandoning multiattribute utilities appears impractical. Clients expect cost-effectiveness results because decision processes are structured around them. Guidelines require them. Reviewers expect them. The analyst is therefore incentivized to treat the EQ-5D index not as what it is, a preference-weighted score, but as what the system demands it to be: a quantitative measure suitable for arithmetic. This is not deception in the ordinary sense; it is accommodation. The framework requires closure, and the analyst supplies it.

The argument that "it cannot all be wrong" rests on a misunderstanding of how error accumulates in institutionalized practice. The history of science contains many examples where entire communities operated within frameworks later recognized as fundamentally flawed. What is unusual in HTA is not that error persisted, but that it persisted while claiming the mantle of quantitative science. The scale of publication does not make the claims correct; it makes the belief system resilient. Each new publication cites and reinforces the same assumptions, creating the appearance of confirmation without exposing the framework to refutation.

In this sense, the abundance of results is not evidence of success but an embarrassment of riches. It signals a system that has been extraordinarily productive in generating numbers while being largely insulated from the question of whether those numbers measure anything. Analysts can continue to apply multiattribute instruments for closed cost-effectiveness claims only by implicitly accepting that closure, comparability, and procedural legitimacy are substitutes for measurement validity. Once that acceptance is made explicit, the scientific pretense collapses.

What would be unique in the annals of science is not that a framework was wrong, but that it remained dominant for decades without ever being exposed to the risk of being wrong. Multiattribute utility instruments and the reference case simulation together created precisely that condition. They enabled the continuous production of authoritative-looking quantitative claims that could not be falsified because they were not grounded in measurable attributes. The result is a vast literature that is internally consistent, methodologically refined, and epistemically hollow.

For analysts, the implication is stark. Continuing to deploy multiattribute instruments as the basis for closed cost-effectiveness claims is no longer a neutral methodological choice. It is a commitment to an evaluative fiction. The alternative, accepting the limits of description and insisting on measurement-valid single-attribute claims, disrupts established practice but restores scientific accountability. The embarrassment, ultimately, is not the richness of the literature, but the realization that abundance has been mistaken for evidence.

3. THE TRANSITION TO MEASUREMENT IN HEALTH TECHNOLOGY ASSESSMENT

THE IMPERATIVE OF CHANGE

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

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

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

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

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

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

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

MEANINGFUL THERAPY IMPACT CLAIMS

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

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

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

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

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

THE PATH TO MEANINGFUL MEASUREMENT

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

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

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

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

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

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

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

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

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

TRANSITION REQUIRES TRAINING

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

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

Training must therefore begin with scale types and their permissible operations. Linear ratio measurement applies to manifest attributes that possess a true zero and invariant units, such as time, counts, and resource use. Latent attributes, by contrast, cannot be observed directly and cannot be measured through summation or weighting. They require formal construction through a measurement model capable of producing invariant units. This distinction is the conceptual fulcrum of reform, because it determines which claims are admissible and which are not.

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

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

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

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

A NEW START IN MEASUREMENT FOR HEALTH TECHNOLOGY ASSESSMENT

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

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

DESIGNED FOR CLOSURE

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

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

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

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

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

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

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

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

ACKNOWLEDGEMENT

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

REFERENCES

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

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

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

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

[5] Buchholz I, Mathieu F, Janssen M et al. A Systematic Review of Studies Comparing the Measurement Properties of the Three-Level and Five-Level Versions of the EQ-5D." *PharmacoEconomics* 36, no. 6 (2018): 645–661

[6] Shaw C, Longworth L, Bennett B,. A Review of the Use of EQ-5D for Clinical Outcome Assessment in Health Technology Assessment, Regulatory Claims, and Published Literature." *Patient*, 2024; 17(30): 251