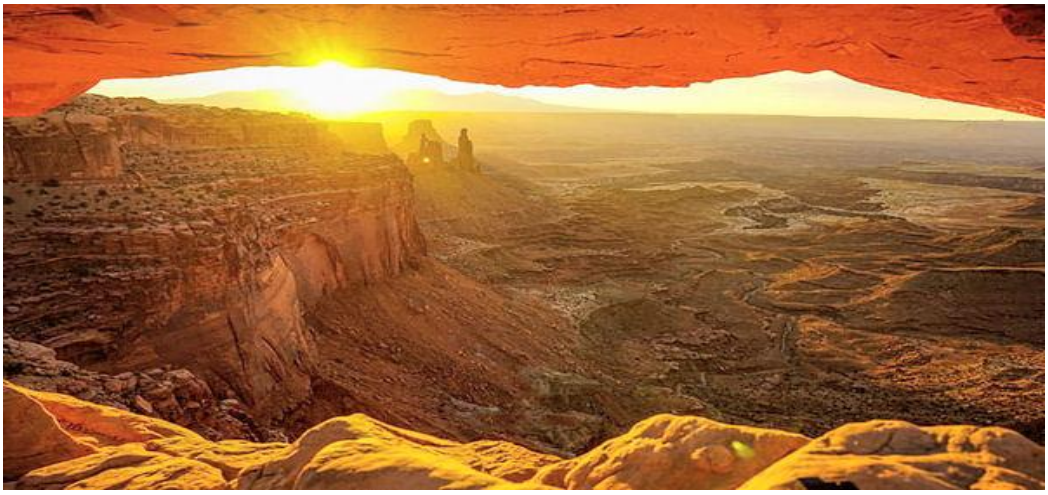
MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION



REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

UNITED STATES: COSMIN AND MEASUREMENT INVERSION - THE CERTIFICATION OF NON-MEASURES

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

LOGIT WORKING PAPER No 237 JANUARY 2026

www.maimonresearch.com

Tucson AZ

FOREWORD

HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

COSMIN (Consensus-based Standards for the Selection of Health Measurement Instruments) is an international initiative developed to provide methodological standards for evaluating, selecting and reviewing health measurement instruments, particularly patient-reported outcome measures [1] [2]. Its principal aim is to improve consistency in the assessment of instruments used to evaluate symptoms, health status, quality of life and related subjective outcomes in healthcare research and practice.

The COSMIN framework provides structured guidance for assessing measurement properties including reliability, validity, responsiveness, internal consistency and structural validity. It also offers checklists and risk-of-bias tools intended to support systematic reviews of measurement instruments. Researchers can use COSMIN standards to judge whether an instrument is appropriate for a target population and whether its reported measurement properties satisfy accepted methodological requirements.

COSMIN has become influential across outcomes research, clinical trials and health technology assessment because it provides a common language for evaluating PROMs and other health-related instruments. The framework also recognizes psychometric methods including item response theory and Rasch analyses as approaches for examining instrument structure and performance. Through its role in defining acceptable measurement instruments and guiding systematic reviews, COSMIN occupies an important position within the broader evidence generation process. Instruments evaluated under COSMIN standards frequently become inputs into clinical studies, evidence synthesis and HTA evaluations.

The objective of this assessment was to interrogate the HTA-related aspects of the COSMIN knowledge base to determine the extent to which its framework for evaluating health measurement instruments aligns with the principles of representational measurement. COSMIN occupies an important position within outcomes research because it establishes standards for selecting and evaluating patient-reported outcome measures and other subjective instruments used in clinical research and health technology assessment. Given its role in defining acceptable evidence inputs, an important question is whether COSMIN distinguishes between psychometric assessment and lawful measurement. The interrogation employed the standardized twenty-four item canonical framework previously applied across agencies, journals, educational programs and professional organizations to determine endorsement probabilities and normalized logits for propositions concerning representational measurement, Rasch measurement, arithmetic admissibility and falsifiability.

The findings suggest that COSMIN occupies an unusual position within the broader measurement inversion landscape. Unlike many HTA interrogations, the COSMIN knowledge base

demonstrates partial recognition of concepts such as unidimensionality and instrument structure. However, endorsement of foundational propositions concerning measurement preceding arithmetic, representational measurement requirements and Rasch measurement as the basis for latent attribute assessment remains weak. The resulting profile suggests that COSMIN recognizes psychometric concepts while remaining only weakly connected to the axioms governing lawful quantitative claims. Rather than complete absence of measurement language, the profile points toward a more subtle form of measurement inversion where reliability, validity and responsiveness become substitutes for representational measurement itself. The implication is that measurement failure may enter the evidence stream at the point where instruments are evaluated and certified for use.

The starting point is simple and inescapable: *measurement precedes arithmetic*. This principle is not a methodological preference but a logical necessity. One cannot multiply what one has not measured, cannot sum what has no dimensional homogeneity, cannot compare ratios when no ratio scale exists. When HTA multiplies time by utilities to generate QALYs, it is performing arithmetic with numbers that cannot support the operation. When HTA divides cost by QALYs, it is constructing a ratio from quantities that have no ratio properties. When HTA aggregates QALYs across individuals or conditions, it is combining values that do not share a common scale. These practices are not merely suboptimal; they are mathematically impossible.

The modern articulation of this principle can be traced to Stevens' seminal 1946 paper, which introduced the typology of nominal, ordinal, interval, and ratio scales [3]. 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) [4] . 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 [5]. 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 [6].

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 COSMIN KNOWLEDGE BASE

The COSMIN knowledge base assessed in this interrogation is defined by the materials, standards and methodological guidance developed by the Consensus-based Standards for the Selection of Health Measurement Instruments initiative as they relate to health outcomes measurement and its interface with health technology assessment. The focus is not on the wider psychometric literature or broader measurement theory but on the specific concepts reinforced through COSMIN guidance, checklists, methodological publications and standards used to evaluate health measurement instruments, particularly patient-reported outcome measures.

COSMIN was established to provide a standardized framework for evaluating the quality of health measurement instruments. Its principal objective is to improve consistency in determining whether instruments are suitable for use in clinical studies, outcomes research and systematic reviews. The framework provides criteria for evaluating reliability, validity, responsiveness, structural validity, content validity and internal consistency. COSMIN also offers methodological guidance and risk-of-bias tools designed to support systematic reviews of measurement instruments.

Particular emphasis within the COSMIN framework is placed on psychometric properties. Instrument performance is evaluated according to predefined criteria intended to establish methodological quality and appropriateness for specific populations and applications. Structural validity and dimensionality receive attention, and psychometric approaches including item response theory and Rasch methods are acknowledged as analytical techniques relevant to instrument assessment.

Importantly, the knowledge base examined here concerns the assumptions embedded within these standards rather than external measurement theory. The issue is not whether representational measurement exists elsewhere in the literature but whether it forms a meaningful component of the concepts reinforced by COSMIN itself. This distinction is critical because the interrogation seeks to determine the degree to which the target knowledge structure behaves as though propositions concerning measurement are accepted or rejected.

Because COSMIN occupies an upstream position within healthcare evidence generation, its influence extends beyond instrument selection. Instruments assessed under COSMIN standards frequently enter clinical studies, systematic reviews and health technology assessments. Consequently, assumptions embedded within COSMIN concerning the nature of measurement have broader implications for the evidence pipeline itself. The interrogation therefore provides insight not only into instrument evaluation practices but also into whether foundational concepts concerning attributes, measurement and admissible arithmetic are embedded within the standards governing evidence generation.

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: COSMIN

Table 1 presents, the endorsement 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.

TABLE 1: ITEM STATEMENT, RESPONSE, ENDORSEMENT AND NORMALIZED COSMIN

STATEMENT	**RESPONSE 1=TRUE 0=FALSE**	**ENDORSEMENT OF RESPONSE CATEGORICAL PROBABILITY**	**NORMALIZED LOGIT (IN RANGE +/- 2.50)**
INTERVAL MEASURES LACK A TRUE ZERO	1	0.35	-0.60
MEASURES MUST BE UNIDIMENSIONAL	1	0.65	+0.60
MULTIPLICATION REQUIRES A RATIO MEASURE	1	0.15	-1.75

TIME TRADE-OFF PREFERENCES ARE UNIDIMENSIONAL	0	0.65	+0.60
RATIO MEASURES CAN HAVE NEGATIVE VALUES	0	0.70	+0.85
EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES	0	0/65	+0.60
THE QALY IS A RATIO MEASURE	0	0.70	+0.85
TIME IS A RATIO MEASURE	1	0.75	+1.10
MEASUREMENT PRECEDES ARITHMETIC	1	0.20	-1.40
SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES	0	0.80	+1.40
MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC	1	0.15	-1.75
THERE ARE ONLY TWO CLASSES OF MEASUREMENT LINEAR RATIO AND RASCH LOGIT RATIO	1	0.10	-2.20
TRANSFORMING SUBJECTIVE RESPONSES TO INTERVAL MEASUREMENT IS ONLY POSSIBLE WITH RASH RULES	1	0.25	-1.10
SUMMATION OF LIKERT QUESTION SCORES CREATES A RATIO MEASURE	0	0.75	+1.10
THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE	0	0.70	+0.85
CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.15	-1.75
QALYS CAN BE AGGREGATED	0	0.75	+1.10
NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.30	-0.85
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.70	+0.85
THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.60	+0.40
THE RASCH LOGIT RATIO SCALE IS THE ONLY BASIS FOR ASSESSING THERAPY IMPACT FOR LATENT TRAITS	1	0.20	-1.40
A LINEAR RATIO SCALE FOR MANIFEST CLAIMS CAN ALWAYS BE COMBINED WITH A LOGIT SCALE	0	0.85	+1.75

THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT	1	0.35	-0.60
THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.25	-1.10

COSMIN: THE MEASUREMENT ILLUSION

The interrogation of the HTA aspects of the COSMIN knowledge base presents a particularly important extension of the measurement inversion program. Unlike agencies concerned directly with reimbursement decisions or organizations centered on cost-effectiveness modeling, COSMIN occupies a position upstream in the evidence generation process. Its stated purpose is to guide the selection and evaluation of health measurement instruments, especially patient-reported outcome measures (PROMs). Through standards for reliability, validity, responsiveness, structural assessment and systematic reviews of instruments, COSMIN influences decisions regarding which outcomes are considered acceptable and which instruments are regarded as scientifically credible. For this reason, one might reasonably expect COSMIN to exhibit a particularly strong commitment to representational measurement principles. If an organization devoted to measurement instruments does not understand measurement itself, the implications extend far beyond psychometrics and reach directly into HTA.

The interrogation suggests a more complicated and ultimately troubling profile. COSMIN does not display the same almost complete absence of measurement language observed in many HTA environments. Indeed, concepts such as dimensionality, validity, reliability and Rasch analysis appear within COSMIN materials. Yet the endorsement profile indicates that representational measurement itself remains weakly embedded. The result is not complete measurement ignorance but a form of institutionalized measurement inversion where measurement terminology exists while the axioms governing lawful quantitative claims remain weakly supported.

The strongest positive feature of the COSMIN profile concerns unidimensionality. The proposition “Measures must be unidimensional” receives a probability of endorsement of 0.65 with a normalized logit of +0.60. Relative to many HTA interrogations this is comparatively strong. This is unsurprising. COSMIN recognizes structural validity and dimensional assessment as important features of instrument evaluation. Instruments are expected to demonstrate coherent internal structure and dimensional consistency.

Yet from the perspective of representational measurement this support remains incomplete. Unidimensionality is necessary but not sufficient. A scale can demonstrate internal structure and still fail to achieve measurement. Internal consistency and structural coherence indicate relationships among responses; they do not establish quantity. Representational measurement asks a prior question: do the observations support a lawful scale structure permitting arithmetic? COSMIN appears more comfortable discussing psychometric organization than addressing the stronger requirements of measurement itself.

COSMIN therefore creates a subtle but important substitution. Representational measurement asks whether an attribute exists, whether observations support quantitative structure and whether arithmetic operations are admissible. COSMIN asks different questions: reliability, responsiveness, content validity and internal consistency. These are not equivalent. Psychometric performance gradually substitutes for measurement itself. Instruments become accepted because they perform consistently rather than because they generate lawful measures. The result is a framework where methodological adequacy increasingly replaces representational legitimacy.

This limitation becomes evident when considering the proposition "Measurement precedes arithmetic." Here endorsement falls to $p = 0.20$ with a normalized logit of -1.40. This finding is critical because it captures perhaps the most important principle in representational measurement. Arithmetic operations are permissible only after scale properties have been demonstrated. Numbers alone do not create measures.

The implication is that COSMIN appears willing to evaluate instruments using psychometric criteria while remaining largely silent regarding whether arithmetic operations applied to resulting scores are legitimate. Instruments may be reliable, responsive and internally consistent, yet these characteristics do not establish ratio properties or admissible transformations. The weak endorsement of measurement preceding arithmetic therefore suggests a framework concerned with evaluating instruments without fully confronting whether the outputs of those instruments constitute measures.

This issue becomes especially important for subjective responses. Patient-reported outcomes begin with observations expressed through response categories: no difficulty, mild difficulty, moderate difficulty and severe difficulty. Such responses are ordinal. They establish order but not quantity. The distinction is fundamental because ordinal responses support ranking but not arithmetic involving means, ratios or multiplication.

Yet COSMIN appears vulnerable to precisely this problem. The proposition "Summations of subjective instrument responses are ratio measures" receives strong rejection with $p = 0.80$ and logit $+1.40$. At first sight this seems encouraging. However, the broader profile suggests that while explicit ratio interpretation of summed scores may be rejected, the representational implications remain weakly understood. Instrument assessment appears to proceed as though psychometric adequacy largely resolves the issue.

The same issue appears in the proposition "Summation of Likert question scores creates a ratio measure," rejected with $p = 0.75$ and logit $+1.10$. Again, the response appears superficially reassuring. Yet COSMIN continues to treat summed scores as acceptable outputs for reliability, responsiveness and comparative analysis. The distinction between ordinal score construction and lawful measurement remain only partially recognized.

The role of Rasch analysis provides perhaps the clearest example of this ambiguity. COSMIN acknowledges Rasch and item response approaches as acceptable methods for instrument assessment. Yet the proposition "Transforming subjective responses to interval measurement is only possible with Rasch rules" receives weak endorsement with $p = 0.25$ and logit -1.10. This result is important because Rasch is not merely another statistical option. Under representational

measurement Rasch provides the mechanism through which observations concerning latent attributes can be transformed into invariant measures if model requirements are satisfied. Treating Rasch as one psychometric approach among many substantially alters its role.

This interpretation becomes stronger when considering the proposition “The Rasch logit ratio scale is the only basis for assessing therapy impact for latent traits,” which receives p = 0.20 and logit −1.40. This suggests that COSMIN recognizes Rasch procedurally but not foundationally. Rasch appears to occupy the position of a useful analytical technique rather than the central mechanism for lawful latent measurement. The consequence is substantial. A framework may require reporting of Rasch analyses without recognizing why Rasch matters. Reporting on an analysis is not equivalent to understanding its representational significance.

The proposition “There are only two classes of measurement: linear ratio and Rasch logit ratio” receives one of the weakest endorsements in the profile: p = 0.10 with normalized logit −2.20. This finding may represent the most revealing feature of the interrogation. The distinction between manifest and latent attributes appears largely absent.

The endorsement profile also suggests weak recognition that latent attributes differ fundamentally from manifest attributes. Manifest claims such as prescription counts, treatment duration and hospital days possess observable ratio structures. Latent attributes such as symptom burden, treatment satisfaction and need fulfillment require indirect inference through invariant structures. Weak endorsement of propositions concerning Rasch measurement and possession indicates that COSMIN largely treats subjective responses as scores rather than manifestations of attribute possession. Yet therapy evaluation for latent constructs concerns changes in possession, not movement in arbitrary numerical totals.

This absence is important because manifest and latent attributes require fundamentally different measurement approaches. Manifest attributes such as prescription counts or hospital days possess observable structure and support linear ratio scales. Latent constructs such as symptom burden or treatment satisfaction require indirect inference through invariant structures. Without this distinction, subjective responses become vulnerable to arbitrary numerical treatment.

This limitation becomes especially visible in the outcome concept itself. The proposition “The outcome of interest for latent traits is possession of that trait” receives only p = 0.35 and logit −0.60. The implication is that COSMIN remains focused upon scores rather than possession. Yet Rasch measurement is fundamentally concerned with locating persons on a continuum of attribute possession.

The distinction matters because therapy evaluation concerns change in possession of latent attributes rather than movement in arbitrary numerical scores. Score changes do not necessarily indicate quantity changes.

The profile also demonstrates substantial weakness regarding representational constraints on arithmetic. “Meeting the axioms of representational measurement is required for arithmetic” receives p = 0.15 and logit −1.75. Likewise, “Multiplication requires a ratio measure” receives p = 0.15 and logit −1.75.

These results suggest weak understanding of admissible transformations and arithmetic restrictions. This becomes especially relevant because COSMIN-guided instruments frequently enter broader HTA environments where arithmetic operations become extensive.

PROMs evaluated under COSMIN frameworks frequently become incorporated into utility systems, preference constructions and comparative analyses. If representational requirements are not enforced at the instrument stage, measurement failure propagates downstream.

This observation creates an important connection between COSMIN and broader HTA measurement inversion. COSMIN does not directly construct QALYs or reference-case simulations. Rather, it influences the instruments entering those systems.

If instrument standards fail representational requirements, subsequent HTA structures inherit measurement failure before modeling even begins.

The implications become clearer when considering broader evidence structures. COSMIN frequently supports systematic reviews of measurement instruments and interacts closely with evidence synthesis traditions. Here the interrogation raises an uncomfortable possibility. Instrument assessment may create an appearance of rigor while bypassing the prior question of whether measurement has occurred.

Reliability becomes emphasized. Responsiveness becomes emphasized. Structural validity becomes emphasized. Yet representational legitimacy remains largely absent. The resulting framework risks certifying instruments that remain fundamentally ordinal.

Importantly, this should not be interpreted as criticism of individual analysts or researchers. The endorsement structure suggests something broader: institutional assumptions inherited across psychometric and HTA traditions. Analysts work within inherited conventions. Instrument developers satisfy accepted standards. Reviewers apply recognized frameworks. The issue therefore is systemic.

The significance of COSMIN lies precisely here. Because it occupies an upstream position within evidence generation, its assumptions influence a large downstream ecosystem of PROM development, evidence synthesis and HTA evaluation. The interrogation suggests that COSMIN may therefore represent a critical transmission mechanism within the broader HTA memeplex.

Unlike reference-case modeling organizations where measurement inversion appears explicit, COSMIN demonstrates a subtler form. Measurement language is present. Technical sophistication is present. Yet the axioms governing lawful quantitative claims remain weakly embedded. This distinction makes COSMIN particularly important. It demonstrates that measurement inversion does not require overt rejection of representational measurement. It can arise through procedural substitution where psychometric adequacy replaces measurement itself.

The implications extend beyond COSMIN. If frameworks governing instrument evaluation do not understand representational measurement, then evidence pipelines become contaminated at origin.

The issue therefore is not whether HTA models later misuse outcomes. The issue may be that measurement failure enters before outcomes ever reach HTA.

COSMIN may therefore occupy a more important position than initially suspected. It does not directly construct QALYs or reference-case models. Rather, it determines what counts as acceptable evidence before evidence generation begins. If representational measurement is absent at this stage, subsequent evidence synthesis and HTA frameworks merely inherit and amplify earlier assumptions. Measurement inversion may therefore begin not with cost-effectiveness models but with the certification of non-measures themselves.

This interrogation therefore opens an important new avenue for understanding measurement inversion. Previous assessments focused on QALYs, utilities and reference-case models. COSMIN points toward a broader architecture where instrument standards themselves institutionalize assumptions inconsistent with representational measurement.

For over forty years this structure remained largely invisible because criticism focused on outputs. The present interrogation suggests attention should move upstream toward the standards governing evidence generation itself. The result is potentially one of the most important findings in the measurement inversion program. COSMIN was expected to represent measurement expertise. Instead, the interrogation suggests a framework that discusses measurement extensively while remaining only weakly connected to the axioms required for lawful measurement. That distinction may prove decisive.

CONCLUSION

The interrogation of the COSMIN knowledge base points toward a form of measurement inversion that is both subtle and potentially far-reaching. Unlike many HTA organizations where false measurement assumptions appear explicit, COSMIN presents a more complicated profile. Measurement language is present. Concepts such as dimensionality, reliability, validity and Rasch analysis appear throughout its framework. Yet the interrogation suggests that representational measurement itself remains only weakly embedded. The result is not rejection of measurement but a procedural substitution where psychometric adequacy increasingly replaces lawful measurement.

This distinction is important because COSMIN occupies a critical upstream position within evidence generation. It determines which instruments are regarded as scientifically credible and therefore which outcomes become candidates for clinical studies, systematic reviews and ultimately health technology assessment. If representational measurement requirements are weakly recognized at this stage, downstream systems simply inherit these assumptions. Measurement failure enters the evidence stream before cost-effectiveness models, utilities or QALYs are ever constructed.

The implications extend beyond COSMIN itself. The interrogation suggests that the problem may not be isolated methodological oversight but a broader institutional structure in which ordinal and composite constructions acquire legitimacy through accepted psychometric standards. Instruments become accepted because they perform consistently rather than because they produce lawful

measures. Reliability, responsiveness and structural validity become proxies for measurement itself.

COSMIN may therefore occupy a more important role than initially anticipated. It does not directly create QALYs or reference-case models. Instead, it determines what counts as evidence before evidence generation begins. Measurement inversion may therefore begin not with the construction of economic models but with the certification of non-measures themselves. If so, COSMIN represents one of the earliest transmission points within the broader evidence production memeplex.

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] Mokkink L, Terwee C, Patrick D et al. The COSMIN Checklist for Assessing the Methodological Quality of Studies on Measurement Properties of Health Status Measurement Instruments: An International Delphi Study. *Quality of Life Research* 2010; 19 (4): 539–49

[2] Mokkink L, Terwee C, Knol D et al. 2010. The COSMIN Checklist for Evaluating the Methodological Quality of Studies on Measurement Properties: A Clarification of Its Content. *BMC Medical Research Methodology* 2010; 10:22

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

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

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

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