MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

UNITED STATES: COCHRANE AND MEASUREMENT INVERSION - THE AGGREGATION OF NON-MEASURES

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

LOGIT WORKING PAPER No 238 JANUARY 2026

www.maimonresearch.com

Tucson AZ

FOREWORD

HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

The Cochrane Collaboration is an international organization dedicated to improving healthcare decision making through the production and dissemination of systematic reviews and evidence synthesis [1]. Established in 1993, its central objective is to evaluate and summarize evidence concerning the effects of healthcare interventions using standardized and transparent methodological procedures. Cochrane reviews are widely regarded as a major component of evidence-based medicine and frequently occupy the highest level within hierarchies of evidence.

Cochrane develops detailed methodological guidance through its *Cochrane Handbook for Systematic Reviews of Interventions*, which provides standards for protocol development, literature searching, study selection, data extraction, risk-of-bias assessment, evidence grading and statistical analysis. The organization places particular emphasis on reducing bias and improving consistency in the interpretation of clinical evidence.

A major feature of Cochrane reviews is the use of meta-analysis, where results from multiple studies may be statistically combined to generate pooled estimates of treatment effects. Continuous outcomes, patient-reported outcomes and quality-of-life measures frequently enter these analyses through mean differences or standardized mean differences. Cochrane also supports approaches such as GRADE to assess the certainty of evidence and inform interpretation.

Through its influence on clinical guidelines, policy development and health technology assessment, Cochrane occupies an important position in the broader architecture of healthcare evidence production.

The objective of this assessment was to interrogate the HTA-related aspects of the Cochrane knowledge base to determine the extent to which evidence synthesis within systematic reviews and meta-analysis aligns with the principles of representational measurement. Cochrane occupies a central position within evidence-based medicine because it establishes methodological standards for evaluating, synthesizing and interpreting healthcare evidence. Given its influence on clinical guidelines, policy development and health technology assessment, an important question is whether Cochrane recognizes the requirement that measurement must precede arithmetic or whether evidence synthesis assumes that quantitative properties already exist. The assessment employed the standardized twenty-four item canonical framework previously applied across agencies, journals, professional organizations and educational institutions to estimate endorsement probabilities and normalized logits for propositions concerning representational measurement, Rasch measurement, arithmetic admissibility and falsifiability.

The findings suggest that Cochrane exhibits a distinctive form of measurement inversion centered on evidence synthesis itself. Unlike organizations focused on utilities or reference-case models, Cochrane is primarily concerned with the aggregation and interpretation of evidence. Yet the endorsement profile indicates weak support for representational measurement principles together

with strong support for assumptions that permit arithmetic operations on subjective and composite outcomes. Rasch measurement and latent trait possession approaches approach floor values while arithmetic assumptions surrounding synthesis receive relatively stronger support. The resulting profile suggests that Cochrane assumes measurement has already occurred before evidence enters systematic review. The implication is that evidence synthesis may not correct measurement failure but rather normalize and amplify it. Measurement inversion therefore becomes embedded not through utility construction itself but through the institutional processes responsible for legitimizing and aggregating evidence.

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 [2]. 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) [3] . 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 [4]. 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 [5].

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, PhD

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 COCHRANE KNOWLEDGE BASE

The Cochrane knowledge base assessed in this interrogation is defined by the concepts, standards and methodological assumptions reinforced through the Cochrane Collaboration and its guidance for systematic reviews and evidence synthesis. The focus is not on the broader literature of evidence-based medicine or the extensive statistical literature surrounding systematic review methods. Rather, the interrogation is directed toward the specific knowledge structures embodied within Cochrane guidance, methodological standards, handbook chapters and review practices as they relate to healthcare evidence generation and health technology assessment.

The Cochrane Collaboration was established to improve healthcare decision making through transparent and standardized approaches to evaluating and synthesizing evidence. Its principal methodological resource, the *Cochrane Handbook for Systematic Reviews of Interventions*, provides extensive guidance covering protocol development, literature searching, study selection, data extraction, risk-of-bias assessment, evidence grading and statistical synthesis. Through these procedures Cochrane seeks to reduce bias and improve consistency in evidence interpretation.

A central component of the Cochrane framework is evidence synthesis through systematic review and meta-analysis. Studies addressing similar interventions and outcomes may be combined statistically to produce pooled estimates of treatment effects. Continuous outcomes frequently enter these analyses through mean differences and standardized mean differences. Patient-reported outcomes, symptom scales, quality-of-life instruments and utility measures may also become candidates for synthesis where reviewers judge them to represent similar constructs.

The knowledge base examined here concerns assumptions embedded within this evidence synthesis framework rather than external theories of measurement. The issue is not whether representational measurement exists elsewhere in methodological literature but whether such principles form a meaningful component of concepts reinforced by Cochrane itself. This distinction is critical because the interrogation seeks to determine how the target knowledge structure behaves toward propositions concerning attributes, arithmetic, latent measurement and falsifiability.

Because Cochrane occupies a central position in evidence hierarchies, its influence extends beyond evidence synthesis alone. Systematic reviews frequently become inputs into clinical guidelines, policy recommendations and health technology assessment. Consequently, assumptions embedded within Cochrane concerning the legitimacy of outcome measures may have broader implications for the evidence production process itself. The interrogation therefore provides insight not merely into review methodology but into whether foundational principles concerning attributes and lawful measurement occupy a meaningful role within the broader architecture of 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: COCHRANE

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 COCHRANE

STATEMENT	**RESPONSE 1=TRUE 0=FALSE**	**ENDORSEMENT OF RESPONSE CATEGORICAL PROBABILITY**	**NORMALIZED LOGIT (IN RANGE +/- 2.50)**
INTERVAL MEASURES LACK A TRUE ZERO	1	0.25	-1.10
MEASURES MUST BE UNIDIMENSIONAL	1	0.30	-0.85
MULTIPLICATION REQUIRES A RATIO MEASURE	1	0.15	-1.75

TIME TRADE-OFF PREFERENCES ARE UNIDIMENSIONAL	0	0.80	+1.40
RATIO MEASURES CAN HAVE NEGATIVE VALUES	0	0.80	+1.40
EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES	0	0.75	+1.10
THE QALY IS A RATIO MEASURE	0	0.80	+1.40
TIME IS A RATIO MEASURE	1	0.65	+0.60
MEASUREMENT PRECEDES ARITHMETIC	1	0.15	-1.75
SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES	0	0.85	+1.75
MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC	1	0.10	-2.20
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	0.10	-2.20
SUMMATION OF LIKERT QUESTION SCORES CREATES A RATIO MEASURE	0	0.85	+1.75
THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE	0	0.80	+1.40
CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.10	-2.20
QALYS CAN BE AGGREGATED	0	0.85	+1.75
NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.25	-1.10
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.75	+1.10
THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.70	+0.85
THE RASCH LOGIT RATIO SCALE IS THE ONLY BASIS FOR ASSESSING THERAPY IMPACT FOR LATENT TRAITS	1	0.05	-2.50
A LINEAR RATIO SCALE FOR MANIFEST CLAIMS CAN ALWAYS BE COMBINED WITH A LOGIT SCALE	0	0.85	+1.75

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

COCHRANE: THE AGGREGATION ILLUSION

The interrogation of Cochrane reviews opens a different and potentially larger avenue for assessing measurement inversion. Cochrane is not an HTA agency in the narrow reimbursement sense. It is not NICE, PBAC, ICER or ISPOR. Its principal role is evidence synthesis. Yet precisely for this reason it occupies a strategic position in the scientific evidence chain. If Cochrane methods assume that outcomes entering systematic reviews and meta-analyses are already valid measures, then measurement inversion becomes institutionalized not only in HTA modeling but in the machinery used to summarize evidence itself.

The Cochrane knowledge base is highly sophisticated regarding review methods. Its Handbook provides detailed guidance for protocols, searching, study selection, bias assessment, statistical analysis, GRADE and interpretation of results. It also provides specific guidance on patient-reported outcomes and continuous outcome measures, including mean differences and standardized mean differences. This sophistication is precisely what makes the interrogation important. The issue is not methodological sloppiness. The issue is whether systematic review methodology asks the prior question that representational measurement requires: are the variables entering synthesis lawful measures of defined attributes?

The profile suggests that Cochrane reviews strongly reinforce arithmetic and synthesis while giving weak recognition to representational measurement. The proposition “measurement precedes arithmetic” receives only $p = 0.15$, with a normalized logit of -1.75. This is central. Meta-analysis is an arithmetic operation applied across studies. It estimates pooled effects, combines means, calculates standardized effects and interprets differences. Yet representational measurement insists that before arithmetic is legitimate, the underlying values must possess appropriate scale properties. Cochrane methods are largely organized around whether studies can be pooled statistically, not whether the underlying outcomes satisfy the axioms required for measurement.

This is especially important for subjective outcomes. Cochrane guidance recognizes patient-reported outcomes as important because they capture patients’ perspectives on benefit and harm. This is entirely appropriate at the level of relevance. Patients’ experiences matter. The difficulty begins when subjective responses are treated as if they produce quantitative measures suitable for means, standardized mean differences and meta-analysis. A patient-reported outcome instrument may generate ordered responses, summed scores or composite scales. These are not automatically measures. They may rank observations, but ranking does not establish equal intervals, ratio properties, invariance or dimensional homogeneity.

The interrogation reflects this weakness. The proposition “summations of subjective instrument responses are ratio measures” receives p = 0.85, logit +1.75, indicating strong endorsement of the false measurement assumption in the Cochrane review environment. Similarly, “summation of Likert question scores creates a ratio measure” receives p = 0.85, logit +1.75. These values capture the central problem. Systematic reviews frequently inherit trial outcomes expressed as total scores. Once those scores are reported as continuous outcomes, they become candidates for mean differences, standardized mean differences and pooled effects. The score enters the review as if it were a measure. The prior representational question disappears.

The standardized mean difference is particularly revealing. Cochrane guidance notes that continuous outcomes may be compared using mean difference or standardized mean difference, and that when studies measure the same construct using different scales, reviewers must interpret the standardized mean difference or use alternative effect measures. This guidance is statistically coherent within conventional evidence synthesis, but from a representational measurement perspective it is deeply problematic. The phrase “same construct” does heavy work. It assumes that different instruments measure the same attribute. But has this been demonstrated? Do the instruments share a unidimensional attribute? Do they preserve empirical relations? Are they invariant? Are they merely different ordinal scoring systems? Standardization by the pooled standard deviation cannot answer these questions.

This is why the Cochrane interrogation assigns only p = 0.30 to “measures must be unidimensional” and p = 0.10 to “meeting the axioms of representational measurement is required for arithmetic.” Cochrane review methods may acknowledge constructs, domains and outcomes, but they do not appear to place representational measurement at the foundation of synthesis. Studies may be grouped because they are judged clinically or conceptually similar. Instruments may be pooled because they are assumed to address the same construct. But similarity of labels is not measurement equivalence. “Quality of life,” “symptom burden” or “functional status” may conceal multiple attributes, different item structures and different response models.

The problem is not solved by risk-of-bias assessment. Cochrane excels in assessing trial design, allocation concealment, blinding, missing data and selective reporting. These are important. But they are not measurement. A low-risk-of-bias trial can still use an invalid outcome score. A well-conducted meta-analysis can still combine non-measures. Statistical precision cannot compensate for failure of measurement.

Cochrane therefore creates a second form of substitution. Risk-of-bias assessment, evidence grading and methodological rigor become increasingly detached from the prior issue of whether outcomes themselves possess lawful measurement properties. The result is that procedural rigor may substitute for representational legitimacy. A low-risk-of-bias study using invalid outcomes remains methodologically impressive but measurement deficient. Evidence architecture increasingly replaces measurement architecture.

This gives Cochrane a distinctive inversion profile. Unlike COSMIN, which at least centers on instrument properties, Cochrane centers on evidence synthesis. Its framework is designed to combine evidence, not reconstruct measurement. The result is that measurement assumptions are imported silently from the included studies. If the primary literature reports a score, Cochrane

methods can treat it as an analyzable continuous outcome. The machinery of synthesis then proceeds. In this sense, Cochrane may institutionalize a second-order measurement inversion: it does not merely accept non-measures; it aggregates them.

The Rasch-related items are particularly important. “The Rasch logit ratio scale is the only basis for assessing therapy impact for latent traits” receives p = 0.05, logit −2.50. “The outcome of interest for latent traits is the possession of that trait” also receives p = 0.05, logit −2.50. “The Rasch rules for measurement are identical to the axioms of representational measurement” receives p = 0.05, logit −2.50. These floor-level endorsements indicate that Cochrane’s review logic does not understand latent outcomes as possession claims located on a logit continuum. Instead, latent outcomes appear as instrument scores, scales or continuous variables available for synthesis.

The endorsement profile also suggests weak recognition of the distinction between manifest and latent attributes. Manifest outcomes such as mortality, treatment duration and hospital admissions possess observable ratio properties. Latent attributes such as symptom burden, treatment satisfaction and quality of life require indirect inference through invariant measurement structures. Yet Cochrane review logic largely treats both classes of outcomes as continuous variables available for synthesis. The consequence is that latent traits become scores rather than manifestations of possession.

This is not a minor omission. For latent attributes, therapy impact should be expressed as change in possession of the attribute. If a therapy is claimed to improve symptom burden, need fulfillment or treatment satisfaction, the claim requires lawful measurement of that latent attribute. Rasch measurement provides a route to this when model requirements are satisfied. Without Rasch transformation, ordinal responses remain ordinal. Yet Cochrane reviews frequently work with whatever outcome scales trials report. If those are summed ordinal scores, they become part of the evidence base.

The QALY and utility-related statements also show strong inversion. The propositions “the QALY is a ratio measure,” “the QALY is a dimensionally homogeneous measure,” and “QALYs can be aggregated” receive p = 0.80, p = 0.80 and p = 0.85 respectively. Cochrane is not primarily a QALY-producing organization, but it includes economics and health technology-relevant evidence within its broader review environment. The Handbook includes economics among its specialized topics, and economic outcomes may enter reviews. The problem is that Cochrane’s general evidence-synthesis orientation does not appear to challenge whether utilities and QALYs satisfy representational measurement. If these outcomes are reported in eligible studies, they may be summarized, interpreted or used in evidence profiles without confronting their measurement status.

This matters because evidence synthesis confers legitimacy. A single flawed measure in one trial is a local problem. A systematic review pooling many flawed measures gives the appearance of stronger evidence. The hierarchy of evidence then magnifies the error. A pooled estimate from multiple trials may be treated as more reliable, but reliability of aggregation does not establish validity of measurement. If each input fails measurement requirements, the pooled output remains invalid. More data do not rescue measurement failure.

The falsification profile is also weak. “Non-falsifiable claims should be rejected” receives p = 0.25, logit −1.10. “Reference case simulations generate falsifiable claims” receives p = 0.75, logit +1.10. Cochrane’s framework is oriented toward empirical studies, so one might expect stronger support for falsifiability. Yet the problem is that systematic review methodology often treats reported outcomes and model outputs as evidence if they satisfy review inclusion criteria. It does not necessarily require that the claims themselves be prospectively falsifiable. In economic and HTA-related contexts, this allows simulation outputs and utility-based projections to enter evidence discussions without satisfying normal science standards.

The Cochrane profile therefore suggests a broader institutional issue. Measurement inversion is not limited to the creation of utilities, QALYs or reference-case models. It can occur when evidence synthesis assumes that reported outcomes are measures because they are numerical. This is a deeper and potentially more damaging form of inversion. It affects how evidence is summarized, how treatment effects are interpreted and how conclusions are translated into guidelines and policy.

Cochrane’s great strength, standardization of evidence synthesis, may also become its vulnerability. Standardization is valuable only if the objects being standardized are legitimate. If outcome measures are invalid, standardization can spread the error more efficiently. The same logic applies to meta-analysis. Pooling is powerful only when what is pooled represents comparable quantities. If studies use different ordinal scales for different latent attributes, standardized mean differences produce comparability by statistical construction rather than measurement equivalence.

The implication is that Cochrane reviews should be interrogated not only for risk of bias but for risk of measurement failure. Before pooling, reviewers should ask: What attribute is being measured? Is it manifest or latent? If latent, has unidimensionality been demonstrated? Has Rasch transformation established a lawful logit scale? Are scores merely ordinal? Are instruments commensurate? Do differences preserve meaning? Is the outcome capable of supporting the arithmetic used in synthesis?

These questions are not currently central to Cochrane methodology. Their absence explains the interrogation profile. Cochrane is strong on review architecture but weak on representational foundations. It asks whether studies are methodologically eligible, whether bias is controlled, whether effects can be estimated and whether certainty can be graded. It does not consistently ask whether the underlying outcome variables are measures.

This creates a new avenue in the measurement inversion program. COSMIN shows how instrument standards may fail to grasp representational measurement. Cochrane shows how evidence synthesis may amplify that failure. Together they indicate an upstream pathway: instruments generate ordinal scores; trials report those scores; systematic reviews pool them; HTA models or guidelines treat the pooled estimates as evidence. At each stage, the same assumption persists: numerical output is treated as measurement.

Cochrane may therefore occupy a more important position than initially suspected. It does not directly create utilities, QALYs or reference-case simulations. Rather, it determines how outcomes

become legitimized through synthesis. If representational measurement is absent at this stage, downstream HTA systems simply inherit and amplify these assumptions. Measurement inversion may therefore arise not from constructing non-measures but from aggregating them into accepted evidence.

The conclusion is unavoidable. Cochrane reviews represent an essential target for interrogation because they reveal how measurement inversion can be institutionalized through evidence synthesis itself. The problem is not that Cochrane lacks methodological rigor. It is that methodological rigor is directed toward synthesis after measurement has been assumed. Under representational measurement, this sequence is reversed. Measurement must be established first. Only then can synthesis proceed.

The Cochrane interrogation therefore strengthens the broader thesis. Measurement inversion is not merely an HTA modeling problem. It reaches into the systems that define, collect, summarize and interpret evidence. If Cochrane does not confront representational measurement, then systematic review and meta-analysis risk becoming highly disciplined methods for aggregating non-measures. That may be one of the most important implications of the entire interrogation program.

CONCLUSION

The interrogation of the Cochrane knowledge base points toward a distinctive and potentially far-reaching form of measurement inversion. Unlike organizations directly associated with utilities, QALYs and reference-case simulations, Cochrane occupies a central position in evidence synthesis. Its influence extends across clinical research, guideline development and health technology assessment. Precisely because of this role, the findings have implications beyond Cochrane itself. If evidence synthesis assumes that numerical outcomes entering systematic reviews already possess lawful measurement properties, then measurement inversion becomes embedded not in the construction of evidence but in the process through which evidence acquires legitimacy.

The interrogation suggests that Cochrane exhibits considerable methodological sophistication while remaining only weakly connected to representational measurement. Risk-of-bias assessment, evidence grading, sensitivity analysis and statistical rigor receive extensive attention. Yet these procedures operate after measurement has been assumed. The critical prior question—whether variables entering synthesis constitute lawful measures of defined attributes—largely disappears. Procedural rigor increasingly substitutes for representational legitimacy.

This distinction is important because Cochrane occupies an upstream position within the evidence production process. It does not directly create utilities, QALYs or reference-case models. Instead, it determines how evidence becomes summarized, standardized and legitimized. If outcome measures entering systematic review remain fundamentally ordinal or composite constructions, subsequent synthesis simply amplifies measurement failure. Aggregating non-measures does not create measurement.

The implications extend beyond systematic review methodology. Cochrane may therefore represent an important transmission mechanism within the broader evidence production

memeplex. COSMIN may certify non-measures; Cochrane may aggregate them; HTA agencies subsequently inherit their outputs as accepted evidence. Measurement inversion may therefore arise not from the construction of economic models alone but from the institutional processes through which evidence itself is produced and validated. If so, Cochrane reveals that the challenge extends well beyond HTA to the broader architecture of modern evidence generation.

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] Cumpston M, Li T, Matthew J. Page M et al. "Updated Guidance for Trusted Systematic Reviews: A New Edition of the Cochrane Handbook for Systematic Reviews of Interventions." *Cochrane Database of Systematic Reviews* 2019; 10: ED000142.

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

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

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

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