MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

UNITED STATES: THE U.S. ACADEMIC HTA MEMEPLEX - INSTITUTIONALIZED ARITHMETIC WITHOUT MEASUREMENT

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

LOGIT WORKING PAPER No 6 JANUARY 2026

www.maimonresearch.com

Tucson AZ

FOREWORD

HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

This Logit Working Paper series documents a finding as extraordinary as it is uncomfortable: health technology assessment (HTA), across nations, agencies, journals, and decades, has developed as a global system of non-measurement. It speaks the language of numbers, models, utilities, QALYs, "value for money," thresholds, discounting, incremental ratios, extrapolations, and simulations. It demands arithmetic at every turn, multiplication, division, summation, aggregation, discounting, yet it never once established that the quantities to which these operations are applied are measurable. HTA has built a vast evaluative machinery on foundations that do not exist. The probabilities and normalized logits in the reports that follow provide the empirical confirmation of this claim. They show, with unsettling consistency, that the global HTA knowledge base neither possesses nor applies the principles of scientific measurement.

The objective of this study is to interrogate, systematically and explicitly, the measurement knowledge base of a representative sample (n=40) of university-based and similar academic health technology assessment (HTA) research groups in the United States. These institutions occupy a privileged epistemic position within HTA: they train analysts, publish methodological standards, supply evidence reports to public agencies, and routinely construct or validate the cost-effectiveness models that inform pricing, access, and reimbursement decisions. The analysis applies a 24-statement true/false diagnostic grounded in representational measurement theory to assess whether this academic ecosystem possesses, endorses, or operationalizes the axioms that make quantitative claims meaningful. The intent is not to critique individual papers or isolated modeling choices, but to evaluate whether the academic HTA system, taken as a collective knowledge environment, satisfies the minimal conceptual requirements for lawful arithmetic, falsifiable claims, and cumulative scientific learning.

The findings are unequivocal. The U.S. academic HTA knowledge base exhibits a pervasive and systematic rejection of the axioms of representational measurement. Core principles of unidimensionality, the requirement of ratio scales for multiplication, the logical precedence of measurement over arithmetic, and the inadmissibility of composite constructs such as the QALY are weakly endorsed at best and frequently rejected outright. At the same time, false propositions embedded in conventional HTA practice are strongly reinforced, including the treatment of ordinal preference scores as interval or ratio measures, the aggregation of QALYs, and the legitimacy of arithmetic operations performed on non-measures. Rasch measurement, the only framework capable of transforming subjective responses into invariant measures suitable for quantitative inference, is effectively absent from the academic corpus.

What emerges from the logit structure is not a pattern of partial misunderstanding or disciplinary immaturity, but a coherent and internally stable inversion of scientific logic. Arithmetic is treated as epistemically primary, while measurement is relegated to a rhetorical afterthought. The axioms that would constrain or prohibit standard HTA practices are systematically excluded, while the numerical outputs those axioms would invalidate are granted full academic legitimacy. This inversion is not accidental. It reflects the operation of a tightly coupled memeplex, a self-reinforcing belief system in which utilities, QALYs, ICERs, and reference-case simulations mutually sustain one another and are protected from foundational critique by professional norms, curricula, publication standards, and career incentives.

Within this memeplex, falsification is functionally neutralized. Claims are not exposed to empirical risk; they are insulated through modeling assumptions, sensitivity analyses, and appeals to convention. As a result, university-based HTA groups do not function as sites of conjecture and refutation, nor as engines for the evolution of objective knowledge. Instead, they act as institutional amplifiers of false measurement, reproducing and normalizing mathematically inadmissible claims under the imprimatur of academic rigor. The unanimity observed across academic centers is therefore not evidence of correctness or consensus achieved through scientific selection. It is evidence of a closed epistemic system—one that has replaced measurement with arithmetic, science with simulation, and knowledge growth with numerical storytelling.

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

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

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

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

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

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

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

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

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

Paul C Langley, Ph.D

Email: langleylapaloma@gmail.com

DISCLAIMER

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

1. INTERROGATING THE LARGE LANGUAGE MODEL

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

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

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

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

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

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

THE UNIVERSITY RESEARCH KNOWLEDGE BASE

For the purposes of this analysis, the *academic HTA knowledge base* is defined as the shared and recurrent body of concepts, assumptions, methods, and evaluative norms produced and reinforced by university-based health technology assessment and health economics research groups in the United States. It is not identified with any single institution, guideline, or publication. Rather, it is inferred from stable patterns of practice that recur across teaching materials, methodological papers, evidence assessments, simulation models, advisory outputs, and professional training activities undertaken by these groups over time.

The sampling frame is intentionally representative rather than exhaustive. Approximately forty university-based or academically anchored HTA and outcomes research centers were treated as illustrative of the broader U.S. academic ecosystem. These include academic evidence-assessment centers participating in federally sponsored review programs, pharmacy-school and public-health-school health economics units, university-affiliated cost-effectiveness modeling groups, and academic teams that regularly supply analytic support to payers, government agencies, or organizations producing reference-case HTA evaluations. What unites these entities is not organizational form, funding source, or disciplinary label, but function: they generate, teach, review, or legitimate quantitative claims about therapy impact.

The knowledge base is inferred from recurring methodological commitments rather than from explicit statements of measurement philosophy. These commitments include routine use of cost-utility analysis, widespread acceptance of QALYs and ICERs as admissible quantitative constructs, reliance on reference-case simulation modeling, summation or indexation of patient-reported outcome instruments without Rasch transformation, and the systematic treatment of sensitivity analysis as a substitute for empirical falsification. Equally important are the absences: the near-total exclusion of representational measurement theory from curricula, the lack of engagement with scale-type constraints in methodological discourse, and the effective invisibility of Rasch measurement despite pervasive reliance on latent-trait claims.

In this sense, the academic HTA knowledge base is defined behaviorally and structurally, not rhetorically. It reflects what university-based HTA groups do repeatedly, defend implicitly, and transmit to students, journals, reviewers, and policy audiences as *best practice*. The 24-statement diagnostic is therefore applied not as a survey of individual beliefs or intentions, but as a probe of the conceptual boundaries within which academic HTA work is produced, evaluated, and deemed legitimate. The resulting profile captures the epistemic architecture of U.S. academic HTA as it actually operates and as the findings demonstrate, that architecture is fundamentally incompatible with the requirements of scientific measurement.

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: US UNIVERSITY RESEARCH HTA GROUPS

Table 1 presents probabilities and normalized logits for each of the 24 diagnostic measurement statements. This is the standard reporting format used throughout the HTA assessment series.

It is essential to understand how to interpret these results.

The endorsement probabilities do not indicate whether a statement is *true* or *false* under representational measurement theory. Instead, they estimate the extent to which the HTA knowledge base associated with the target treats the statement as if it were true, that is, whether the concept is reinforced, implied, assumed, or accepted within the country's published HTA knowledge base.

The logits provide a continuous, symmetric scale, ranging from +2.50 to –2.50, that quantifies the degree of this endorsement. the logits, of course link to the probabilities (p) as the logit is the natural logarithm of the odds ratio; logit = ln[p/1-p].

- Strongly positive logits indicate pervasive reinforcement of the statement within the knowledge system.
- Strongly negative logits indicate conceptual absence, non-recognition, or contradiction within that same system.
- Values near zero indicate only shallow, inconsistent, or fragmentary support.

Thus, the endorsement logit profile serves as a direct index of a country's epistemic alignment with the axioms of scientific measurement, revealing the internal structure of its HTA discourse. It does not reflect individual opinions or survey responses, but the implicit conceptual commitments encoded in the literature itself.

TABLE 1: ITEM STATEMENT, RESPONSE, ENDORSEMENT AND NORMALIZED LOGITS US UNIVERSITY RESEARCH HTA GROUPS

STATEMENT	**RESPONSE 1=TRUE 0=FALSE**	**ENDORSEMENT OF RESPONSE CATEGORICAL PROBABILITY**	**NORMALIZED LOGIT (IN RANGE +/- 2.50)**
INTERVAL MEASURES LACK A TRUE ZERO	1	0.20	-1.40
MEASURES MUST BE UNIDIMENSIONAL	1	0.25	-1.10
MULTIPLICATION REQUIRES A RATIO MEASURE	1	0.15	-1.75

TIME TRADE-OFF PREFERENCES ARE UNIDIMENSIONAL	0	0.85	+1.75
RATIO MEASURES CAN HAVE NEGATIVE VALUES	0	0.850	+1.75
EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES	0	0.90	+2.20
THE QALY IS A RATIO MEASURE	0	0.90	+2.20
TIME IS A RATIO MEASURE	1	0.95	+2.50
MEASUREMENT PRECEDES ARITHMETIC	1	0.15	-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	0.15	-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	0.10	-2.20
SUMMATION OF LIKERT QUESTION SCORES CREATES A RATIO MEASURE	0	0.90	+2.20
THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE	0	0.80	+1.40
CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.20	-1.40
QALYS CAN BE AGGREGATED	0	0.95	+2.50
NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.60	+0.40
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.85	+1.75
THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.65	+0.60
THE RASCH LOGIT RATIO SCALE IS THE ONLY BASIS FOR ASSESSING THERAPY IMPACT FOR LATENT TRAITS	1	0.10	-2.20
A LINEAR RATIO SCALE FOR MANIFEST CLAIMS CAN ALWAYS BE COMBINED WITH A LOGIT SCALE	0	0.60	+0.40

THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT	1	0.25	-1.10
THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.05	-2.50

US ACADEMIC HTA RESEARCH CENTERS: THE SYSTEMATIC INSTITUTIONALIZATION OF FALSE MEASUREMENT

The purpose of this assessment is not to identify methodological disagreement or intellectual diversity within United States academic health technology assessment and health economics research centers. It is to determine whether these centers, taken collectively, satisfy the minimum conditions required for participation in normal science. The answer provided by the 24-item diagnostic is unequivocal. US academic HTA research centers do not merely fail to enforce representational measurement axioms; they have collectively constructed an evaluative culture in which those axioms are actively displaced, marginalized, or treated as irrelevant to practice. The resulting belief system is coherent, stable, and deeply entrenched. It is also scientifically indefensible.

The defining feature of the diagnostic profile is not confusion but inversion. Propositions that would constrain arithmetic by requiring valid measurement are consistently rejected, while propositions that permit the unrestricted manipulation of numbers derived from ordinal, multidimensional, or non-invariant sources are endorsed at or near ceiling levels. Measurement is not treated as a prerequisite for calculation. It is treated as an optional philosophical add-on that can be ignored without consequence. Arithmetic, once performed, is assumed to confer meaning retroactively. This is the precise opposite of scientific reasoning.

This inversion is immediately visible in the treatment of the most elementary principle of quantitative science: measurement precedes arithmetic. With endorsement at p = 0.15 (logit −1.75), this proposition is rejected by the academic HTA corpus. Yet the same corpus overwhelmingly endorses the aggregation of QALYs at p = 0.95 (logit +2.50), the ratio status of QALYs at p = 0.90 (+2.20), and the interval status of EQ-5D algorithms at p = 0.90 (+2.20). These commitments cannot coexist coherently. Aggregation, multiplication, and ratio arithmetic presuppose measurement properties that the system simultaneously denies. The contradiction is resolved not by revising arithmetic, but by expelling measurement theory from the domain of admissible critique.

The incremental cost-effectiveness ratio illustrates this pathology in its purest form. Academic HTA centers continue to teach, publish, and defend ICERs while rejecting the proposition that multiplication requires a ratio measure at p = 0.15 (−1.75). This is not a subtle inconsistency. It is a categorical denial of the condition under which cost can be divided by effect. The ICER survives not because it satisfies measurement axioms, but because those axioms are treated as optional. This is arithmetic without permission, institutionalized as best practice.

The treatment of subjective outcomes reveals an even deeper failure. Summation of Likert responses is endorsed as creating ratio measures at p = 0.90 (+2.20), while the proposition that subjective responses require Rasch transformation to achieve interval measurement is rejected at p = 0.10 (−2.20). This pairing is devastating. It demonstrates that academic HTA centers have chosen summation as a substitute for measurement and have done so knowingly. Ordinal categories are treated as quantities because treating them as such is operationally convenient. The discipline has normalized the belief that numbers become measures through repetition, not through axiomatic justification.

Unidimensionality, the defining requirement for any meaningful scale, is treated with similar disregard. The belief that measures must be unidimensional is weakly endorsed at p = 0.25 (−1.10), while the belief that time trade-off preferences are unidimensional is strongly endorsed at p = 0.85 (+1.75). This contradiction reveals how unidimensionality is operationalized in practice: it is declared, not demonstrated. Multiattribute constructs are labeled as single quantities because arithmetic requires them to be so. Factor analysis and internal consistency coefficients are used rhetorically to mask the absence of true dimensional homogeneity.

The Rasch block of the diagnostic provides the clearest evidence that this is not accidental ignorance but structural exclusion. Every Rasch-related proposition collapses to near-floor endorsement. The existence of only two admissible classes of measurement, linear ratio for manifest attributes and Rasch logit ratio for latent traits, is rejected at p = 0.10 (−2.20). The identity of Rasch rules with representational measurement axioms is rejected at the same level. The necessity of Rasch for latent-trait impact assessment is likewise rejected. This pattern does not mean that Rasch papers never appear in the literature. It means that Rasch is not allowed to function as a governing constraint. It is tolerated as a niche technique, admired at the margins, but never permitted to invalidate dominant instrument families or analytic practices.

This is precisely how a memeplex survives. As Dawkins described, successful memeplexes permit local variation while protecting their replicators. In US academic HTA, the replicators are summed rating scales, multiattribute indices, utility algorithms, QALYs, ICERs, and reference-case simulation models. Rasch threatens these replicators because it exposes their lack of invariant units and meaningful zero points. The academic ecosystem therefore neutralizes Rasch by rendering it optional. The logit profile captures this boundary with mathematical clarity.

The same logic governs falsification. Academic HTA centers endorse the abstract principle that non-falsifiable claims should be rejected at only moderate levels, p = 0.60 (+0.40), while strongly endorsing the belief that reference-case simulations generate falsifiable claims at p = 0.85 (+1.75). This is a category error elevated to doctrine. Simulation outputs are conditional projections derived from assumptions, not empirical tests of claims against reality. Sensitivity analysis explores internal model behavior; it does not expose hypotheses to refutation. By redefining falsifiability as robustness to scenario variation, academic HTA replaces scientific risk with model stability. Claims become unfalsifiable by construction while retaining the appearance of rigor.

The consequence of this belief system is epistemic stagnation. Without measurement, there can be no cumulative knowledge. Without invariant quantities, replication becomes repetition with different scoring rules. Disagreement is resolved through consensus, guideline alignment, or negotiated thresholds rather than empirical refutation. What evolves is not objective knowledge, but institutional confidence in numerical storytelling.

It is essential to emphasize that this is not a failure of individuals. It is a systemic outcome of training, publication incentives, funding structures, and professional identity. Academic HTA centers produce analysts who are fluent in modeling but illiterate in measurement. They can manipulate utilities and ICERs with great sophistication while being unable to defend the arithmetic they perform. When challenged, they retreat to precedent. The diagnostic shows that precedent is the problem.

The implications are severe. These centers do not merely study HTA; they constitute its epistemic backbone. They train students, advise agencies, review manuscripts, and define what counts as methodological competence. When they reject measurement axioms and normalize false arithmetic, that posture propagates throughout the entire HTA ecosystem. Agencies, payers, and journals are downstream consumers of an upstream failure.

The remedy is not incremental refinement. It is categorical change. If US academic HTA research centers wish to claim scientific legitimacy, they must accept that only two classes of quantitative claims are admissible. Manifest attributes must be measured on linear ratio scales. Latent traits must be measured on Rasch logit ratio scales with demonstrated invariance. Composite indices, utilities, QALYs, ICERs, and reference-case projections must be reclassified as descriptive constructs, not evidence. Until that transition occurs, US academic HTA will remain what the diagnostic reveals: a sophisticated, confident, and deeply entrenched system of arithmetic without measurement.

THE MEASUREMENT MEMEPLEX

The unanimity of belief that is seen across research centers is not an empirical accident, nor is it the result of independent convergence on a correct framework. It is the predictable outcome of a discipline governed by a memeplex rather than by falsification. Once this is understood, the absence of internal debate ceases to be puzzling. It becomes inevitable.

The concept of a memeplex, articulated most clearly by Richard Dawkins, explains why belief systems can persist even when they are demonstrably false. A memeplex is not a single idea competing on evidentiary merit; it is a mutually reinforcing bundle of beliefs, practices, incentives, and professional identities that evolve to protect themselves against disruption. Health technology assessment, as practiced in academic research centers, fits this description precisely. Utilities, QALYs, ICERs, reference-case models, and summated ordinal instruments do not survive because they are valid. They survive because they **cohere**, institutionally and professionally, into a closed system that rewards internal consistency over external truth.

This explains why there is no serious internal debate about measurement. Debate would require a shared willingness to expose core claims to falsification. Yet the diagnostics show that the proposition that non-falsifiable claims should be rejected is either weakly endorsed or rendered meaningless by the simultaneous endorsement of simulation outputs as falsifiable evidence. In practice, falsification has been redefined out of existence. A claim is considered "tested" if it survives sensitivity analysis, not if it risks refutation by the world. Once falsification is displaced in this way, there is no evolutionary pressure on ideas. They do not live or die by empirical failure. They persist by convention.

The memeplex is reinforced structurally through training, publication, and career progression. Graduate students are not taught competing measurement frameworks. They are taught how to use utilities, compute QALYs, populate models, and interpret cost-effectiveness planes. These are presented not as hypotheses, but as professional competencies. A student who questions whether a utility score is a measure does not initiate a scientific debate; they signal non-membership. The cost of dissent is not intellectual rebuttal, but marginalization. This is how memeplexes enforce conformity without argument.

Academic journals complete the loop. Manuscripts are reviewed by peers who share the same foundational assumptions. Review criteria focus on technical refinement within the framework, not on whether the framework itself is admissible. A paper that questions the ratio properties of utilities or the aggregability of QALYs is not controversial; it is unintelligible within the prevailing belief system. It violates what Kuhn would have called normal science, but here normal science has been severed from measurement.

The result is a discipline that does not evolve objective knowledge. Evolution of knowledge requires conjecture and refutation. It requires claims that can fail. HTA research centers do not produce such claims. They produce numerical artifacts whose authority derives from repetition, not from survival under empirical challenge. Models generate outputs that cannot be wrong in the Popperian sense because they are insulated from reality by assumptions. When results are inconvenient, assumptions are changed. Nothing is learned; the narrative is adjusted.

This is why unanimity persists even as the numbers grow more elaborate. Increasing model complexity does not increase epistemic risk. It reduces it. Complexity absorbs criticism by making it procedural rather than foundational. Disagreement is channeled into parameter choice, discount rates, time horizons, or willingness-to-pay thresholds, none of which threaten the underlying arithmetic. The memeplex thrives on such disputes because they create the appearance of scientific activity while leaving first principles untouched.

The exclusion of representational measurement theory and Rasch measurement is therefore not an oversight. It is a defensive adaptation. Accepting measurement axioms would force the discipline to abandon its central outputs. QALYs would collapse. ICERs would lose meaning. Simulation models would be reclassified as illustrative tools rather than evidence. Careers built on these constructs would face retroactive invalidation. No memeplex voluntarily commits such suicide.

The absence of belief in falsification is thus not philosophical naivety; it is functional necessity. A system built on arithmetic without measurement cannot permit falsification because falsification would expose that there is nothing there to falsify. There are no quantities, only conventions. There are no measures, only scores. There is no cumulative knowledge, only accretion of models.

Seen in this light, the unanimity across U.S. academic HTA and health economics centers is the strongest possible evidence that the field is not practicing normal science. In a healthy scientific discipline, foundational disagreement is common and productive. Here it is absent. Everyone agrees because everyone is speaking the same inherited numerical language, and because speaking another language carries professional cost without institutional reward.

Until a center explicitly breaks with the memeplex by reinstating falsification, enforcing measurement as a prerequisite for arithmetic, and accepting that objective knowledge must evolve through error elimination, unanimity will persist. Not because the framework is correct, but because it is evolutionarily stable in the absence of scientific selection pressure.

3. NEXT STEPS: TRANSITION TO SINGLE-CLAIM MEASUREMENT

The results of LLM interrogation leave no middle path. The measurement cat is out of the bag, and any system that continues using QALYs, utilities, DALYs, or simulation modelling invites scientific ridicule.

DISOWN THE PRESENT BELIEF SYSTEM

The first step toward scientific rehabilitation is an unambiguous renunciation of the non-measurement architecture that has underpinned HTA decision-making for decades. The logic is not rhetorical but structural: if the axioms of representational measurement are violated at the foundation, then no amount of statistical sophistication, modelling embellishment, or "best practice guidelines" can rescue the outputs from incoherence. QALYs, ordinal utilities, DALYs, and reference-case simulations are not merely suboptimal, they are incompatible with any conception of measurement. They lack a legitimate scale type, violate the requirements for meaningful arithmetic, and cannot be integrated into a numerically coherent comparison across interventions. A belief system built on such constructs cannot be amended or partially retained; it must be disowned.

The QALY is the clearest illustration of this impossibility. It is constructed by multiplying ordinal preferences by time, a procedure that lacks dimensional justification and produces outputs that cannot be interpreted as measures of anything. Yet this fiction has persisted because it supplies administrators with a single number, something they can rank, apply a threshold, or negotiate against. The same is true for DALYs, whose lineage in burden-of-disease accounting does nothing to endow them with legitimate measurement properties. Reference-case simulation modelling compounds the error: it takes non-measures as inputs, adds speculation about future clinical and economic pathways, and then outputs a figure that is treated as if it were evidence. The entire apparatus survives only because reviewers, policymakers, and faculty have never been trained in measurement, and thus have lacked the conceptual tools to recognize that these constructs are scientifically impossible.

Disowning the belief system is therefore not an admission of past failure but an unavoidable act of disciplinary self-correction. A field cannot progress while clinging to artefacts that cannot, even in principle, support falsifiable claims. NICE as the exemplar must say so explicitly, not as a symbolic gesture but as the precondition for rebuilding a scientifically credible evaluative architecture.

RECONSTRUCT HTA FROM MEASUREMENT UP

Once the non-measurement framework has been dismantled, reconstruction must begin from the only defensible starting point: measurement theory. There is no shortcut, no incremental reform, and no "middle way" in which QALYs or utilities are patched, modified, or reweighted. The fundamental lesson of representational measurement theory is simple: numbers have meaning only when the empirical structure of the attribute supports a specific scale type. If NICE, assuming it still exists, wants to produce claims that can be evaluated, replicated, and falsified, then it must adopt scale types capable of sustaining the arithmetic it wishes to perform.

For manifest attributes, events that are directly observable, such as hospital days avoided, therapy switching, medication possession, or relapse counts, the appropriate structure is a linear ratio scale. Such scales have a true zero, constant unit intervals, and permit the full suite of permissible arithmetic operations. They allow NICE to make legitimate statements about proportional differences and resource utilization grounded in evidence rather than interpretation. Crucially, ratio scales for manifest outcomes are already ubiquitous in health system data; they require no modelling conjecture and no constructed preferences.

For latent attributes, experiential or subjective constructs such as symptom burden, need-fulfilment, or patient-reported outcomes, the only valid transformation model is the Rasch model. Rasch provides logit-based ratio scales generated through conjoint simultaneous measurement of person ability and item difficulty. Without Rasch, subjective outcomes collapse to ordinal scores that cannot be meaningfully compared or used alongside manifest ratio measures. With Rasch, we acquire disease specific instruments that satisfy unidimensionality, invariance, and interval structure, enabling legitimate claims about latent change.

Reconstruction means reinstating the basic rule that every claim must have the appropriate measurement architecture. This is not an aesthetic preference but the necessary foundation for a science of evaluation. HTA becomes coherent only when claims rest on instruments that conform to the axioms of measurement, not on the administrative desire for a “single number.” The transition is radical only because the prior framework ignored measurement entirely.

MOVE TO PROTOCOL-BASED SINGLE CLAIMS

A measurement-valid HTA system cannot rely on summary constructs or composite evaluations. It must instead adopt a single-claim architecture in which each value claim stands alone, meeting the requirements of falsifiability, replication, and transparent reporting. This follows directly from the logic of science: a claim must be empirically testable, reproducible in the same target population, and supported by an agreed protocol that specifies exactly how evidence will be generated. Multi-outcome cost-effectiveness analysis cannot meet these standards because it integrates non-measures into speculative models and converts them into an imaginary “value for money” figure that cannot be falsified. Single claims, by contrast, are grounded in measurement.

Each claim begins with a precisely defined target population, typically patients initiated on a therapy within a defined window. This eliminates the ambiguity inherent in modelling lifetime populations or hypothetical cohorts. The endpoint must be measurement-valid; a linear ratio measure for manifest attributes or a Rasch logit ratio measure for latent ones. The protocol must articulate the evidence generation plan prospectively: how data will be collected, over what timeframe, using what analytic criteria, and under what conditions replication will be evaluated.

A single-claim architecture aligns HTA with the logic of clinical science. Claims are constructed in advance, not retrospectively assembled from model outputs. They are specific, narrow, and auditable. They permit comparability across therapies because each claim is defined in measurement terms rather than through the aggregation of unrelated dimensions. Importantly, single claims also eliminate the bureaucratic temptation to collapse multiple endpoints into an artificial summary. Instead, each outcome is assessed on its own merits, with its own ruler.

This shift does more than improve methodological defensibility; it transforms the institutional culture of evaluation. NICE, again as the exemplar, would no longer operate as a quasi-modelling agency but as a measurement-based adjudicator of empirically testable propositions. The result is a transparent, reproducible, and scientifically legitimate HTA system.

ADOPT THE MAIMON RESEARCH DISTANCE EDUCATION PROGRAM

Reconstruction requires education, and at present there is no conventional textbook, curriculum, or HTA training program that teaches measurement theory, Rasch, and protocol-based single-claim architecture in a scientifically coherent manner. The existing academic infrastructure remains trapped in the old belief system, recycling utilities, QALYs, and reference-case models as if these constructs were measures. Replacing that architecture therefore requires retraining. systematic, structured, and accessible to agencies, universities, and policy staff. The Maimon Research Distance Education Program is currently the only platform that provides this.

A NEW START IN MEASUREMENT FOR HEALTH TECHNOLOGY ASSESSMENT

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

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

Each program consists of five modules (approx. 5,500 words each), with extensive questions and answers. Each program is priced at US$65.00. Invitations to participate in these programs will be distributed in the first instance to 8,700 HTA professionals in 40 countries.

More detail on program content and access, including registration and on-line payment, is provided with this link: https://maimonresearch.com/distance-education-programs/

The program builds HTA from measurement upward. It teaches representational measurement theory as the foundation for any evaluative claim. It trains participants in Rasch modelling, including item calibration, person–item maps, logit transformations, and the construction of valid, unidimensional latent-trait measures. It provides protocol templates that define how claims are constructed, evaluated, and replicated. It supplies checklists to ensure scale-type coherence, target population definition, and the exclusion of non-measures. It also addresses the institutional, pedagogical, and administrative barriers that have historically prevented HTA from adopting measurement standards.

Most importantly, the program replaces the HTA belief system with a scientific one. It does not attempt to “improve” QALYs or “modernize” utilities. It demonstrates why those constructs are impossible and shows how to build a new system from first principles that produces claims that can be defended in court, in peer review, and in public policy. The program equips faculty and decision-makers with the conceptual tools they were never given, tools that allow them to recognize the difference between a measure and a number masquerading as one. Adopting the program is therefore not supplementary; it is the enabling step. Without a trained workforce, we cannot transition to single-claim measurement.

ACKNOWLEDGEMENT

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

REFERENCES

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

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

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

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