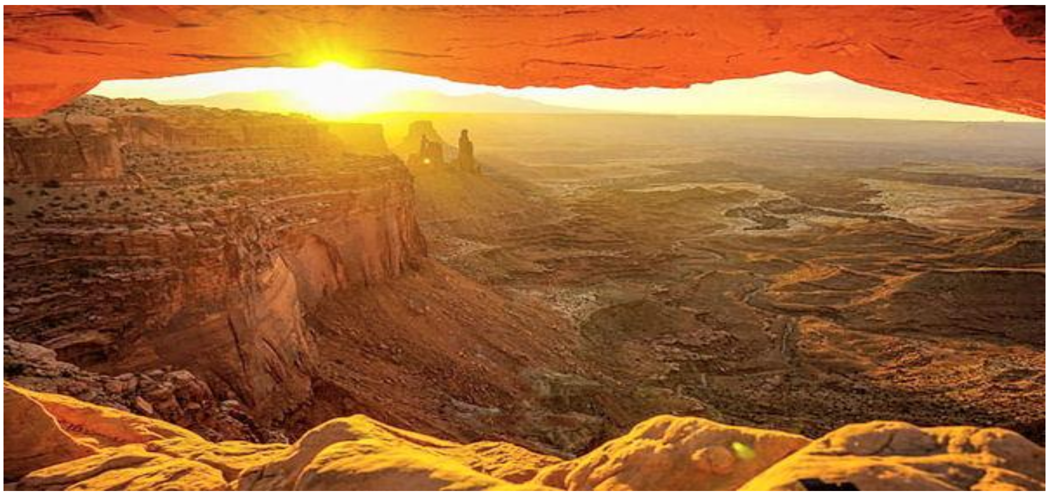
MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION



REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

UNITED KINGDOM: ACADEMIC HTA CENTERS AND THE ABSENCE OF MEASUREMENT

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

LOGIT WORKING PAPER No 29 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 country reports that follow provide the empirical confirmation of this claim. They show, with unsettling consistency, that HTA presents a world of measurement failure.

The objective of this study is to evaluate the extent to which academic health technology assessment (HTA) and health economics centers in the United Kingdom demonstrate possession of the foundational principles required for scientific measurement. Using a standardized 24-item diagnostic grounded in representational measurement theory, the analysis interrogates whether the UK academic knowledge base recognizes and enforces the axioms that determine when numerical operations are permissible. The focus is not on individual publications, authors, or institutions, but on the collective methodological environment reproduced through teaching programs, research outputs, methodological guidance, and professional training pipelines that shape national and international HTA practice.

Rather than treating UK academic HTA as a neutral contributor to policy evaluation, this assessment examines the belief system embedded in what is taught, modeled, published, and normalized as legitimate quantitative evidence. By transforming endorsement probabilities into canonical logits, the analysis identifies the structural orientation of the knowledge base: whether it is anchored in measurement-first principles consistent with normal science, or whether it prioritizes arithmetic frameworks whose dependent variables lack demonstrable measurement status.

The findings reveal a highly stable and internally coherent knowledge system that systematically inverts the scientific ordering of measurement and arithmetic. Core axioms of representational measurement—unidimensionality, scale-type admissibility, and the requirement that measurement precede arithmetic—are driven to the floor of endorsement. In contrast, propositions necessary to sustain cost-utility modeling, utility aggregation, and reference-case simulation are endorsed at or near ceiling levels.

Most striking is the near-total rejection of Rasch measurement as the necessary basis for latent trait quantification. Statements affirming Rasch logit ratio measurement, latent trait possession, and the equivalence between Rasch rules and representational measurement axioms register at the lowest possible logit values. The UK academic HTA knowledge base therefore does not merely underutilize measurement theory; it structurally excludes the only framework capable of transforming subjective observations into invariant quantitative measures. The result is a system that produces numerical outputs at scale while lacking the epistemic conditions required for falsification, replication in the strong sense, or cumulative objective knowledge.

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

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

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

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

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

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

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

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

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

Paul C Langley, Ph.D

Email: langleylapaloma@gmail.com

DISCLAIMER

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

1. INTERROGATING THE LARGE LANGUAGE MODEL

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

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

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

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

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

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

THE ACADEMIC HTA CENTERS KNOWLEDGE BASE

The knowledge base of UK academic HTA and health economics centers can be characterized as a mature, self-reinforcing methodological ecosystem organized around the production and interpretation of numerical outputs rather than the construction of measurable quantities. Within this system, numbers are treated as inherently evidentiary, and the act of calculation itself is taken as a proxy for scientific legitimacy. The foundational question—whether the attributes under analysis possess the structural properties required for measurement—is largely absent from methodological discourse.

At the center of this knowledge base lies the normalization of composite constructs. Health outcomes are routinely represented through utilities, preference weights, and aggregated indices derived from multiattribute health state descriptions. These constructs are treated as if they were quantitative attributes capable of supporting arithmetic operations, despite lacking demonstrated unidimensionality, invariant units, or meaningful zero points. The academy does not require scale-type validation prior to analysis; instead, it substitutes psychometric convention, statistical fit, and consensus practice for measurement proof.

Latent attributes play a central rhetorical role within this ecosystem but are never formally constructed. Concepts such as health-related quality of life, wellbeing, burden, and functioning are invoked as if they were quantities, yet they are not operationalized through measurement models capable of producing invariant units. The absence of Rasch measurement is decisive. Without Rasch transformation, subjective responses remain ordinal, regardless of how sophisticated subsequent statistical manipulation may appear. The knowledge base nevertheless treats summation, averaging, regression, and sensitivity analysis as sufficient to confer quantitative meaning.

This permissive environment allows arithmetic to proceed independently of measurement admissibility. Differences in scores are interpreted as differences in magnitude, changes over time as improvement, and between-group contrasts as treatment effects, even though the underlying numbers do not support interval or ratio interpretation. The distinction between ordering and measuring is blurred or ignored entirely. As a result, statistical coherence replaces measurement validity as the governing standard.

The UK academic ecosystem also maintains close alignment with downstream HTA institutions. Preference-based instruments, utility algorithms, and model-based projections are accepted as legitimate analytic inputs without interrogation of their measurement properties. In doing so, the academy supplies the numerical artifacts required for cost-utility analysis and reference-case modeling while insulating those artifacts from foundational challenge. Measurement theory is not explicitly rejected; it is rendered irrelevant through omission.

What most clearly defines this knowledge base is its patterned silence. Representational measurement theory is not taught as a gatekeeping framework. Scale-type admissibility is not treated as a threshold requirement for claims. Rasch measurement is tolerated only at the margins, never elevated to governing status. These omissions allow the ecosystem to remain internally consistent while remaining epistemically fragile.

The result is a literature that appears methodologically sophisticated yet rests on unexamined assumptions about quantification. It supports elaborate modeling structures, international comparability, and administrative closure, but cannot support falsifiable claims about therapy impact grounded in measurable quantities. The UK academic HTA knowledge base therefore functions not as a corrective force within the global HTA system, but as one of its principal stabilizing mechanisms—reproducing arithmetic without measurement while presenting the result as scientific evaluation.

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: *ACADEMIC HTA CENTERS*

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 *ACADEMIC HTA CENTERS*

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	0.15	-1.,75
MULTIPLICATION REQUIRES A RATIO MEASURE	1	0.10	-2.20
TIME TRADE-OFF PREFERENCES ARE UNIDIMENSIONAL	0	0.90	+2.20

RATIO MEASURES CAN HAVE NEGATIVE VALUES	0	0.90	+2.20
EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES	0	0.90	+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	0.10	-2.20
SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES	0	0.90	+2.20
MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC	1	0.10	-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	0.05	-2.50
SUMMATION OF LIKERT QUESTION SCORES CREATES A RATIO MEASURE	0	0.90	+2.20
THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE	0	0.85	+1.75
CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.10	-2.20
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	0.90	+2.20
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	0.05	-2.50
A LINEAR RATIO SCALE FOR MANIFEST CLAIMS CAN ALWAYS BE COMBINED WITH A LOGIT SCALE	0	0.55	+0.20
THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT	1	0.15	-1.75

THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.05	-2.50

UK ACADEMIC HTA CENTERS: THE INTELLECTUAL ENGINE OF THE ABSENCE OF MEASUREMENT

The UK academic HTA and health economics ecosystem presents itself as the intellectual engine room of "rigorous" evaluation: methods leadership, guideline production, model innovation, and training pipelines that feed agencies, consultancies, journals, and global HTA programs. The 24-item diagnostic profile, however, identifies a very different reality. What is being reproduced in the UK academy is not measurement competence but a stable methodological culture in which arithmetic is treated as inherently authoritative, while the measurement conditions that make arithmetic meaningful are either ignored, denied, or relegated to polite footnotes. The pattern is not a set of small disagreements at the margins. It is a structural inversion: the system treats the outputs of preferred methods as evidence, and then retrofits "validity" language after the fact, rather than beginning with the gating question of whether the quantitative objects under manipulation are measures at all.

The profile is immediately revealed by the collapse of the measurement-first axioms. "Measurement precedes arithmetic" is endorsed at p = 0.10 (−2.20), and the companion proposition, "Meeting the axioms of representational measurement is required for arithmetic," is also p = 0.10 (−2.20). These are not specialist claims. They are the threshold rules that distinguish scientific quantification from symbol pushing. When a knowledge base drives these to the floor, it is declaring—whether explicitly or by patterned practice—that the central task is to compute and to compare, not to justify the admissibility of the computations. In this environment, the "quality" of an assessment becomes synonymous with adherence to established modeling conventions, not adherence to measurement law.

The strongest signal of this inversion is the system's selective literacy. The academy has no difficulty with ratio measurement when the attribute is plainly manifest: "Time is a ratio measure" reaches p = 0.95 (+2.50). That is correct and unsurprising. Yet the same ecosystem simultaneously sustains the arithmetic that requires ratio properties while refusing the ratio gatekeeper. "Multiplication requires a ratio measure" sits at p = 0.10 (−2.20). This is catastrophic because it is the exact condition required to legitimate the defining arithmetic of the QALY framework: multiplying time by a preference weight. The UK academic center does not merely make a mistake here; it institutionalizes an exemption. It insists on the ratio status of time while quietly discarding the ratio requirement at the moment it becomes inconvenient. The result is a discipline that understands measurement perfectly well in one domain and suspends it in another, which is the signature of a belief system rather than a scientific program.

The exemption is maintained by near-ceiling endorsement of the false propositions that keep the QALY machine operating. "The QALY is a ratio measure" sits at p = 0.90 (+2.20). "QALYs can be aggregated" is at p = 0.95 (+2.50). "EQ-5D-3L preference algorithms create interval measures" is at p = 0.90 (+2.20). "Ratio measures can have negative values" is at p = 0.90 (+2.20), which is effectively an admission that the system wishes to keep calling something "ratio" while tolerating exactly what a true zero forbids. These are not peripheral misunderstandings. They are the load-bearing beams of cost-utility analysis. If any one of them were treated as a genuine measurement question, the framework would not merely be weakened; it would be rendered inadmissible as a basis for arithmetic claims about "value" and "cost-effectiveness." The UK academic knowledge base therefore protects these propositions as doctrine, not as testable commitments.

Unidimensionality is treated with the same strategic neglect. "Measures must be unidimensional" sits at p = 0.15 (−1.75). Yet "Time trade-off preferences are unidimensional" is endorsed in the opposite direction at p = 0.90 (+2.20), even though the TTO apparatus is entangled with multiattribute state descriptions, framing effects, and response heuristics that do not magically collapse into a single measurable continuum simply because the field wishes them to. The contradiction is the point. The academy does not enforce unidimensionality as a requirement; it invokes unidimensionality only when it is needed to legitimate the downstream arithmetic. The result is not "imperfect measurement practice." It is a methodological culture in which definitional convenience substitutes for demonstration.

The diagnostic also makes clear where the UK academy places its epistemic boundary: Rasch measurement and the concept of latent trait possession. The statements that would force the field to confront latent attributes as measurable quantities collapse to the absolute floor. "There are only two classes of measurement—linear ratio for manifest attributes and Rasch logit ratio for latent traits"—p = 0.05 (−2.50). "Transforming subjective responses to interval measurement is only possible with Rasch rules"—p = 0.05 (−2.50). "The Rasch logit ratio scale is the only basis for assessing therapy impact for latent traits"—p = 0.05 (−2.50). "Rasch rules are identical to the axioms of representational measurement"—p = 0.05 (−2.50). These are not mild disagreements. They show that Rasch is not permitted to become sovereign. The ecosystem may occasionally tolerate Rasch as a technical option, but it refuses the implication that Rasch is the gatekeeper for any serious claim about measuring patient experience, functioning, symptom burden, or need fulfillment. This refusal is decisive because it preserves the dominant instrument families: summed Likert totals, composite indices, preference algorithms, and "mapped utilities." Rasch would not merely revise those practices; it would invalidate the bulk of them as non-measurement.

This is why the endorsement of "Summation of Likert question scores creates a ratio measure" at p = 0.90 (+2.20) and "Summations of subjective instrument responses are ratio measures" at p = 0.90 (+2.20) is so damning. These statements represent the system's substitute for measurement. Instead of constructing latent measures, it treats arithmetic on ordinal categories as if it manufactured scale properties by repetition and professional consensus. The UK academy is not simply "using imperfect endpoints." It is promoting the belief that scoring is measurement, that a total score is a quantity, and that the subsequent statistical performance of the score is evidence that measurement has occurred. That belief is exactly how the upstream literature supplies the downstream agencies with the pseudo-quantities needed for QALYs and threshold-based closure.

Latent trait possession—what it means to have more or less of an attribute, measured in invariant units—is correspondingly weak. "The outcome of interest for latent traits is the possession of that trait" is p = 0.15 (−1.75). That low endorsement is not accidental. The possession concept forces a discipline to define the attribute, build a measure, test invariance, and report meaningful unit differences. It forces measurement discipline. A culture that prefers to speak in the language of "scores," "changes," "improvements," and "responsiveness" can avoid possession entirely, because those terms allow publication and policy influence without committing to a measurable quantity. The UK academic ecosystem, on this diagnostic, has chosen avoidance.

The profile also exposes how the academy performs a rhetorical tribute to science while operationally bypassing it. "Non-falsifiable claims should be rejected" sits at p = 0.60 (+0.40), which is high enough to maintain a posture of Popper-friendliness. Yet the very mechanism that structures HTA conclusions—long-horizon reference-case simulation—is treated as falsifiable: "Reference case simulations generate falsifiable claims" sits at p = 0.90 (+2.20). This is the laundering step. A simulation is not a hypothesis exposed to empirical risk; it is a conditional projection. It cannot be "verified" in the strong sense, and it is rarely structured to be refuted prospectively within decision-relevant timeframes. By endorsing simulations as falsifiable, the academy preserves the rhetoric of scientific legitimacy while retaining a device whose practical function is administrative closure. "Robustness" becomes stability across scenarios rather than survival against reality.

The consequence is that UK academic HTA does not operate as a correction mechanism for the global memeplex; it operates as its high-status replication engine. The academy supplies trained personnel, publishes methods doctrine, generates "good practice" artifacts, and exports a template that appears sophisticated precisely because it is numerically dense. But density is not legitimacy. The diagnostic shows that the densest part of the apparatus—the arithmetic of utilities, QALYs, aggregation, and thresholds—sits on the most aggressively rejected measurement requirements. In normal science, the constraints would dominate and the conveniences would be punished. Here, the conveniences dominate and the constraints are driven to the floor.

That is why the UK case matters beyond the UK. If the academic centers at the origin point had enforced measurement as a gatekeeper, the global adoption of QALYs and reference-case closure would have faced immediate epistemic resistance. Instead, the UK academy became the priesthood of the new scholasticism: it endowed the framework with mathematical ceremony, methodological liturgy, and a professionalized language of "validation" that carefully avoids the core issue of measurement admissibility. The system did not need everyone to understand representational measurement theory; it needed a critical mass of professionals who could produce plausible numbers in front of audiences who would not ask what those numbers were allowed to mean.

The appropriate interpretation of the UK academic profile is therefore not "room for improvement." It is that the academy, as a knowledge base, has built and defended a methodological identity in which measurement is optional and arithmetic is sovereign. That ordering cannot be repaired by better modeling checklists, more sensitivity analysis, or more elaborate uncertainty quantification. Those are refinements within the memeplex. The defect is prior: the dependent variables are not measures, and the rules that determine whether arithmetic is permitted are rejected at the floor. Until the UK academic ecosystem treats measurement status—unidimensionality, invariance, interval/ratio structure, and in the latent case Rasch logit ratio scaling—as a gating condition for publication, teaching, and assessment, it will remain what the diagnostic reveals: a highly efficient system for producing and exporting numerical storytelling with the prestige of "science" but without the discipline of science.

If the UK academy wished to pivot toward normal science, the implications are stark. It would have to abandon the doctrine that summed ordinal responses can be treated as quantities, treat utilities and QALYs as non-admissible composites rather than "measures," and re-found outcomes assessment on the only two admissible measurement forms: linear ratio measures for manifest attributes and Rasch logit ratio measures for latent traits. That would not be a minor reform; it would be a reclassification of what the field has been doing for forty years. The diagnostic suggests precisely why this has not happened: the present knowledge base is not merely comfortable with the absence of measurement; it is organized to preserve it.

WHY ARE THE HTA PROFESSIONALS IN THE UK UNANIMOUSLY IN FAVOR OF THE ABSENCE OF MEASUREMENT

The unanimity with which UK health technology assessment professionals have accepted—and continue to defend; the absence of measurement is one of the most striking features of the HTA enterprise. It is not plausibly explained by ignorance alone. The principles of measurement were not obscure when the HTA framework was constructed. Stevens' typology of scales was published in 1946, representational measurement theory was formalized decades before NICE was created, and the logic that arithmetic requires admissible scale properties was well established in the natural and social sciences. Yet none of this entered HTA practice in any operational sense. The question, therefore, is not why measurement was misunderstood, but why its absence became so universally tolerated.

The first reason is administrative necessity. HTA in the UK was never designed as a scientific discovery enterprise. It was designed as a rationing instrument. Policymakers required a decision framework that could be applied rapidly, consistently, and with limited empirical data. Measurement-based science, grounded in falsifiable claims and invariant quantities, does not provide closure. Claims remain provisional, open to challenge as new evidence emerges, and subject to revision over time. For administrators charged with allocating finite budgets, this is intolerable. What was needed was not truth in the scientific sense, but decisiveness. The reference case provided precisely that: a standardized template that could generate a numerical conclusion even when the underlying quantities were imaginary.

Second, the HTA framework offered professional insulation. By embedding decisions within a technical apparatus of models, utilities, and thresholds, responsibility could be displaced from judgment to process. Once a model had been run "according to guidance," outcomes were no longer personal or political choices; they were the results of a sanctioned method. Measurement theory would have reintroduced accountability by forcing explicit justification of scale properties, construct definition, and falsifiability. That would have required professionals to defend not merely how they modeled, but whether they were entitled to calculate at all. The absence of measurement thus became protective. It allowed authority without epistemic exposure.

Third, consensus formed early and rapidly. Once the QALY was institutionalized and NICE adopted it as its central decision metric, dissent became professionally irrational. Challenging measurement assumptions did not offer incremental improvement; it threatened total invalidation. This created a powerful selection environment. Researchers who accepted the framework could publish, advise, and advance. Those who questioned it were marginalized as "philosophical," "theoretical," or "unhelpful to decision making." Over time, unanimity emerged not because everyone agreed intellectually, but because disagreement carried no professional payoff.

Fourth, UK HTA developed within an academic culture historically comfortable with scholastic reasoning. The reference case resembles medieval scholasticism far more than experimental science: internal consistency is prized, assumptions are debated within a closed system, and conclusions are judged by coherence rather than empirical refutation. Measurement theory, by contrast, is disruptive. It does not negotiate; it excludes. Either a quantity exists or it does not. Either arithmetic is permitted or it is not. Such gatekeeping is alien to a culture that evolved to manage disagreement through consensus documents and methodological guidance rather than through falsification.

Fifth, the appearance of quantification proved sufficient. Numbers carry authority regardless of their origin. Utilities, QALYs, and ICERs look quantitative, behave numerically, and can be plotted, averaged, and compared. For most audiences—including ministers, managers, and clinicians—the presence of numbers signals rigor. The deeper question of whether those numbers correspond to measurable attributes is invisible unless one is trained to ask it. In this sense, HTA professionals did not need to believe in measurement; they needed others to believe that measurement had occurred.

Finally, the system became self-validating. Because all parties used the same framework, results appeared consistent. That consistency was mistaken for scientific reliability. Yet replication of method is not replication of knowledge when the dependent variable is not a measure. The HTA community confused stability of procedure with truth of outcome. Over time, the absence of measurement ceased to be seen as a problem at all; it became the silent precondition of the entire enterprise.

In combination, these forces explain the apparent unanimity. The UK HTA profession did not converge on the absence of measurement because it was correct, but because it was functional. It delivered administrative closure, protected professional authority, rewarded conformity, and sustained the illusion of scientific objectivity. Measurement, by contrast, would have forced uncertainty, accountability, and ongoing empirical challenge. The absence of measurement was not an oversight. It was the price of a system designed to decide, not to

3. THE TRANSITION TO MEASUREMENT IN HEALTH TECHNOLOGY ASSESSMENT

THE IMPERATIVE OF CHANGE

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

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

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

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

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

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

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

MEANINGFUL THERAPY IMPACT CLAIMS

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

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

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

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

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

THE PATH TO MEANINGFUL MEASUREMENT

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

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

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

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

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

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

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

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

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

TRANSITION REQUIRES TRAINING

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

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

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

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

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

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

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

A NEW START IN MEASUREMENT FOR HEALTH TECHNOLOGY ASSESSMENT

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

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

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/

DESIGNED FOR CLOSURE

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

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

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

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

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

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

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

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

ACKNOWLEDGEMENT

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

REFERENCES

---

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

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

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

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