# MAIMON RESEARCH LLC

# ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

# REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

# UNITED STATES: THE AMERICAN FOUNDATION FOR PHARMACEUTICAL EDUCATION AND THE ABSENCE OF MEASUREMENT IN ACADEMIC PHARMACY

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

**LOGIT WORKING PAPER No 1238 FEBRUARY 2026**

**www.maimonresearch.com**

**Tucson AZ**

# FOREWORD

## HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

The American Foundation for Pharmaceutical Education (AFPE) is an independent, nonprofit organization that supports the advancement of pharmaceutical sciences education and research in the United States. Its primary role is not to deliver professional pharmacy training or to set curricular standards, but to provide financial and programmatic support for graduate education and academic research, particularly at the PhD and postdoctoral levels. AFPE does this mainly through fellowships, scholarships, and research awards that fund students and early-career scientists pursuing advanced study in areas such as pharmaceutics, pharmacology, medicinal chemistry, pharmacokinetics, and related disciplines. By investing in these trainees, the Foundation aims to strengthen the pipeline of pharmaceutical scientists who contribute to drug discovery, development, and evaluation.

In addition to funding, AFPE plays a networking and developmental role, connecting fellows, academic mentors, and industry or regulatory stakeholders. It encourages collaboration across universities and helps position trainees for careers in academia, industry, and government. Its activities therefore sit "upstream" of healthcare delivery, influencing the scientific workforce that underpins pharmaceutical innovation rather than directly shaping clinical practice or policy. In summary, AFPE functions as a capacity-building organization for pharmaceutical science, supporting the education and development of researchers whose work ultimately informs drug development, therapeutic evaluation, and, indirectly, areas such as outcomes research and health technology assessment.

The objective of this assessment is to determine whether the HTA-related knowledge base associated with the AFPE satisfies the axioms of representational measurement required to support valid therapy impact claims. Using a fixed 24-item canonical statement framework, the study evaluates whether the knowledge base behaves as if core measurement principles such as unidimensionality, dimensional homogeneity, and the requirement that measurement must precede arithmetic are recognized and applied. Each statement is assigned an endorsement probability and transformed into a normalized logit, producing a quantitative profile that contrasts alignment with measurement standards against reliance on conventional HTA constructs such as utilities, QALYs, and model-based outputs. The purpose is evaluative rather than descriptive: to establish whether the claims implied by this knowledge base can be considered measurable, empirically testable, and consistent with the requirements of normal science.

The findings demonstrate a consistent pattern of measurement inversion. Statements reflecting the axioms of representational measurement receive low endorsement (typically probabilities in the range 0.10–0.30, with logits between approximately −0.85 and −2.50), indicating that the foundational requirements for admissible arithmetic are weakly represented or absent. In contrast, statements that assume the validity of utilities, QALYs, composite endpoints, and reference case

outputs are strongly endorsed (generally 0.75–0.90, with logits between +1.10 and +2.20). This polarization indicates that the knowledge base accepts arithmetic operations applied to constructs whose measurement properties have not been established. Although there is nominal support for scientific principles such as falsifiability, this is undermined by the simultaneous endorsement of simulation outputs as evaluable claims. The overall profile shows that the AFPE HTA-related knowledge base cannot support claims that are empirically evaluable, replicable, or falsifiable under the axioms of representational measurement.

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 AMERICAN FOUNDATION FOR PHARMACEUTICAL EDUCATION KNOWLEDGE BASE

The knowledge base associated with the AFPS is distinct from that of professional pharmacy schools or policy-focused HTA organizations. AFPE is an independent, nonprofit body whose primary mission is to support graduate education and research in the pharmaceutical sciences. Its activities are centered on the provision of fellowships, scholarships, and research awards that fund doctoral students, postdoctoral researchers, and early-career scientists. Through this support, AFPE contributes to the development of the scientific workforce involved in drug discovery, development, and evaluation. The knowledge base it represents is therefore upstream of clinical practice and policy, shaping the intellectual formation of researchers rather than directly prescribing methods for therapy assessment.

Within this context, HTA-related concepts are not articulated as a formal methodological framework. Instead, they are embedded indirectly within the broader scientific and educational environment that AFPE supports. Trainees funded by AFPE are typically exposed to disciplines such as pharmacology, pharmaceutics, medicinal chemistry, pharmacokinetics, and increasingly to areas such as outcomes research, epidemiology, and translational science. As these researchers move into applied domains, they encounter and adopt the prevailing conventions of health outcomes assessment, including the use of utility-based measures, QALYs, composite endpoints, and model-based economic evaluation.

Because AFPE does not explicitly define HTA standards, its knowledge base is best understood as an inherited and transmitted set of assumptions rather than a deliberately constructed framework. These assumptions reflect the dominant paradigm within pharmaceutical and health outcomes research, where numerical representation is often treated as equivalent to measurement. Constructs such as quality of life, treatment benefit, and patient preference are routinely quantified using multiattribute instruments or aggregated scores, and the resulting values are used in arithmetic operations without prior demonstration that they meet the axioms of representational measurement. The distinction between ordinal, interval, and ratio scales is not consistently enforced, and dimensional homogeneity is not treated as a prerequisite for combining variables.

The knowledge base also reflects a strong orientation toward scientific inquiry, including hypothesis testing, statistical modeling, and experimental design. AFPE’s emphasis on research training ensures that these elements are well represented. However, the presence of statistical and analytical techniques does not guarantee that the variables to which they are applied are measurable. The result is a hybrid structure in which the methods of science are present, but the measurement conditions required to support those methods are only partially satisfied.

Importantly, AFPE’s role in funding and mentoring early-career researchers means that it influences the transmission of these assumptions across the academic pipeline. Researchers trained within this environment are likely to carry forward the conventions they encounter, including the

use of utilities and QALYs in evaluating therapies. In this way, the knowledge base contributes to the stability and persistence of the broader HTA framework, even though it does not explicitly define it. Thus, while it plays a critical role in developing scientific capacity, it does not provide a foundation for generating therapy impact claims that meet the standards of empirical evaluation and normal science.

## 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: AMERICAN FOUNDATION FOR PHARMACEUTICAL EDUCATION

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.

## THE AMERICAN FOUNDATION FOR PHARMACEUTICAL EDUCATION AND THE ABSENCE OF REPRESENTATIONAL MEASUREMENT

The interrogation of the HTA-related knowledge base associated with the American Foundation for Pharmaceutical Education presents a profile that is structurally consistent with those observed across academic pharmacy and HTA-aligned environments, but with a distinctive inflection. Unlike professional education bodies or policy-oriented organizations, AFPE is primarily concerned with the support of pharmaceutical sciences research and graduate training. Its role is not to define HTA frameworks directly, but to shape the intellectual environment in which future researchers are trained. That makes its knowledge base particularly important, because it operates upstream of both teaching and policy. The question is therefore whether that upstream environment corrects or reproduces the measurement deficiencies observed elsewhere. The answer, as indicated by the logit profile, is that it reproduces them (Table 1).

## TABLE 1: ITEM STATEMENT, RESPONSE, ENDORSEMENT AND NORMALIZED LOGITS AMERIAN FOUNDATION FOR PHARMACEUTICAL EDUCATION

| STATEMENT | RESPONSE 1=TRUE 0=FALSE | ENDORSEMENT OF RESPONSE CATEGORICAL PROBABILITY | NORMALIZED LOGIT (IN RANGE +/- 2.50) |
|---|---|---|---|
| INTERVAL MEASURES LACK A TRUE ZERO | 1 | 0.30 | -0.85 |
| 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 | 0.75 | +1.10 |
| RATIO MEASURES CAN HAVE NEGATIVE VALUES | 0 | 0.90 | +2.20 |
| EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES | 0 | 0.80 | +1.40 |
| THE QALY IS A RATIO MEASURE | 0 | 0.80 | +1.40 |
| 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 | 0.80 | +1.40 |
| MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC | 1 | 0.15 | -1.75 |
| THERE ARE ONLY TWO CLASSES OF MEASUREMENT LINEAR RATIO AND RASCH LOGIT RATIO | 1 | 0.05 | -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 | 0.80 | +1.40 |
| 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 | 0.80 | +1.40 |

| | | | |
|---|---|---|---|
| NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED | 1 | 0.75 | +1.10 |
| REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS | 0 | 0.80 | +1.40 |
| THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO | 1 | 0.75 | +1.10 |
| 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 | 0.40 | -0.40 |
| THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT | 1 | 0.30 | -1.60 |
| THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT | 1 | 0.05 | -2.50 |

The defining characteristic of the AFPE knowledge base is a structured divergence between the axioms of representational measurement and the constructs accepted in HTA practice. This divergence is visible immediately in the distribution of probabilities and logits. Statements reflecting the foundational requirements of measurement are weakly endorsed, while those reflecting the assumptions of pharmacoeconomic and outcomes research are strongly endorsed. The resulting profile is one of measurement inversion, not partial misalignment.

The core propositions of measurement theory receive consistently low probabilities. The statement that measurement must precede arithmetic is endorsed at only 0.15, with a logit of −1.75. The proposition that multiplication requires a ratio measure is also at 0.15 and −1.75. The requirement that the axioms of representational measurement must be met before arithmetic is again at 0.15 and −1.75. These values indicate that the knowledge base does not operate under the constraint that measurement is a prerequisite for numerical manipulation. Arithmetic is effectively detached from its foundations.

This detachment is reinforced by weak endorsement of scale properties. The requirement that measures must be unidimensional is assigned a probability of 0.25 and a logit of −1.10. The recognition that interval measures lack a true zero is at 0.30 and −0.85. These values suggest partial awareness but no operational commitment. The knowledge base does not enforce the distinction between ordinal, interval, and ratio scales as a condition for analysis. As a result, composite measures and derived scores can be treated as if they were admissible objects of arithmetic.

In contrast, the positive logits reveal strong endorsement of conventional HTA constructs. The proposition that EQ-5D preference algorithms create interval measures is assigned a probability of

0.80 and a logit of +1.40. The statement that the QALY is a ratio measure is also at 0.80 and +1.40. The same values apply to the dimensional homogeneity of the QALY and the aggregation of QALYs. Summation of subjective responses and Likert scores as ratio measures are endorsed at the same level. This clustering of positive logits indicates that the fundamental assumptions of the reference case framework are accepted without qualification.

The contrast between these two sets of results defines the inversion. The knowledge base assigns −1.75 to the requirement that arithmetic be grounded in measurement, while assigning +1.40 to the assumption that QALYs support valid arithmetic. It assigns −2.50 to the necessity of Rasch-based transformation for latent constructs, while assigning +1.40 to the validity of summated subjective scores. This is not a marginal inconsistency. It is a systematic reversal of the correct sequence of reasoning.

The exclusion of Rasch measurement is particularly important in the AFPE context. As an organization that supports graduate training and research in pharmaceutical sciences, AFPE is positioned to influence how latent constructs are conceptualized and measured. Yet the statements that define Rasch measurement as necessary for transforming subjective responses receive the lowest possible probabilities of 0.05, with logits of −2.50. This indicates that Rasch-based measurement is not part of the effective knowledge base. Instead, latent constructs are represented through composite scores and preference weights, which are then treated as if they were measures.

This substitution has direct implications. Without Rasch transformation, subjective responses cannot be converted into invariant units. The resulting scores remain ordinal, yet they are used in arithmetic operations. The QALY is constructed by multiplying time, correctly recognized as a ratio measure at 0.95 and +2.50, by a utility score that lacks ratio properties. The resulting product is treated as a ratio measure. The knowledge base therefore simultaneously recognizes the scale properties of time and ignores the requirement that only ratio quantities may be multiplied. Arithmetic is again given priority over measurement.

The treatment of falsifiability further illustrates the structure of the knowledge base. The statement that non-falsifiable claims should be rejected is endorsed at 0.75, with a logit of +1.10. This suggests alignment with the principles of normal science. However, the statement that reference case simulations generate falsifiable claims is endorsed at 0.80 and +1.40. This creates a contradiction. Simulation models, being closed systems, do not produce empirically testable claims. Yet they are treated as if they do. The knowledge base therefore maintains the language of falsifiability while endorsing practices that violate it.

The AFPE profile differs from that of teaching institutions in one important respect: it is slightly less extreme in its positive logits. Where pharmacy schools often show values of +1.75 or +2.20 for QALY-related propositions, AFPE shows +1.40. This suggests a somewhat weaker but still substantial endorsement of HTA conventions. However, this moderation does not change the overall structure. The negative logits for measurement principles remain strong, and the positive logits for HTA assumptions remain clearly dominant. The system is therefore not transitional, but stable.

The significance of this stability lies in AFPE's position within the academic ecosystem. By supporting research and graduate training, it influences the intellectual formation of future scholars. If the knowledge base at this level does not incorporate the axioms of representational measurement, then those axioms are unlikely to be introduced at later stages. The result is a self-reinforcing cycle in which flawed assumptions are reproduced through successive generations of researchers.

This has consequences for the evolution of HTA as a discipline. Scientific progress depends on the ability to generate claims that can be tested and potentially falsified. This requires measurement. Without measurement, arithmetic operations are meaningless, and the resulting claims cannot be evaluated. The AFPE knowledge base, by endorsing arithmetic without measurement, undermines this process. It supports the production of numerical outputs, but not the accumulation of objective knowledge.

The implications are therefore unavoidable. A knowledge base that assigns −1.75 to the proposition that measurement must precede arithmetic and +1.40 to the proposition that the QALY is a ratio measure cannot support evaluable claims. A knowledge base that assigns −2.50 to Rasch measurement and +1.40 to summated scores cannot measure latent constructs. A knowledge base that affirms falsifiability while endorsing simulation outputs as falsifiable cannot sustain a coherent scientific framework.

For AFPE, the conclusion is not that its mission is misplaced. Supporting pharmaceutical sciences research and training is essential. The issue is that the HTA-related component of its knowledge base does not meet the standards of representational measurement. As a result, it contributes to the persistence of a framework that cannot generate evaluable claims.

The only viable response is transition. This requires a reorientation toward single-attribute claims grounded in admissible measurement. For manifest attributes, this means linear ratio measures. For latent constructs, it requires Rasch-based logit ratio scales. These approaches provide the conditions necessary for arithmetic operations, empirical evaluation, and replication.

The interrogation shows that the HTA-related knowledge base associated with the AFPE exhibits a clear pattern of measurement inversion. The probabilities and logits demonstrate strong endorsement of conventional HTA constructs and weak endorsement of measurement axioms. The result is a system that produces numerical outputs but cannot generate evaluable claims. If HTA is to function as a scientific discipline, this framework must be replaced with one grounded in representational measurement.

What remains inescapable is not simply that the axioms of representational measurement are absent, but that their absence has persisted in plain sight for almost eighty years. Since the formal articulation of scale types and admissible transformations, the conditions required for measurement have been known, stable, and uncontested within the measurement literature. Yet within HTA and academic pharmacy, these conditions have been set aside without acknowledgement, replaced by a permissive arithmetic applied to constructs that do not meet even the minimum criteria for measurement. This is not an oversight that can be attributed to technical complexity or recent innovation. It is a sustained failure of disciplinary self-scrutiny. The

explanation is therefore institutional, not methodological. A framework has been adopted, taught, and reproduced that delivers numerical outputs with the appearance of rigor, while avoiding the prior obligation to establish what is being measured. Once embedded in curricula, journals, and professional expectations, this framework becomes self-validating. Students are trained within it, researchers publish within it, and decision-makers rely upon it, with no point at which the foundational question is revisited. That this could persist for decades speaks not to the strength of the framework, but to the absence of a requirement that it justify itself. The implication is direct. What has been maintained is not a scientific tradition, but a convention that has insulated itself from the standards that define science. Until those standards are reinstated as non-negotiable, the production of numerical claims will continue, but the possibility of measurement, and therefore of knowledge will remain out of reach.

The failure of the American Foundation for Pharmaceutical Education to recognize the axioms of representational measurement is shared by the American Association of Colleges of Pharmacy. That these two peer educational bodies, founded in 1946 and 1900, respectively should exhibit the same structural deficiency carries significant implications for the evolution of pharmacy education over the past century. The principles of measurement theory have been established and available throughout this period, yet they have not been incorporated into the intellectual framework that underpins pharmacoeconomics and outcomes assessment. Instead, a system has developed in which numerical constructs are treated as measures without satisfying the conditions required for measurement. This suggests not a temporary oversight, but a sustained institutional failure embedded within curricula, research training, and professional expectations. The consequence is that generations of students and researchers have been trained within a framework that produces numerical outputs but cannot support empirically evaluable claims. If pharmacy education is to align with the standards of normal science, this omission must be addressed directly. Without reinstating the axioms of representational measurement as a non-negotiable foundation, the discipline risks continuing to reproduce claims that have the appearance of evidence but lack the properties required to constitute it.

# III. 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 three distance education programs to support the transition to a new paradigm in HTA. These comprise 12 module senior level program that details the standards for measurement, the failure of current HTA standards and the basis for protocol supported claims assessment for ratio measures of manifest attributes and Rasch logic ratio logit measures for latent attributes. The two other programs are only 5 modules but are designed to complement the 12-module program, for measurement axioms and Rasch attribute possession.

**MAIMON RESEARCH LLC**

**DISTANCE EDUCATION PROGRAMS IN THE THEORY OF MEASUREMENT**

Three programs are available: two short 5-module programs and a 12-module program that is structured as a senior level course on the transition from the current HTA belief system to a new paradigm for HTA

The two short programs are (i) **NUMERICAL STORYTELLING: SYSTEMATIC MEASUREMENT FAILURE IN HEALTH TECHNOLOGY ASSESSMENT** and (ii) A **NEW START IN MEASUREMENT FOR HEALTH TECHNOLOGY ASSESSMENT.** They are designed to complement the 12-module course program. They can be accessed through the **DISTANCE EDUCATION** section of the website with URL https://maimonresearch.com/distance-education-programs/

The senior level course **HEALTH TECHNOLOGY ASSESSMENT REBUILT: EVIDENCE AND VALUE** is accessed through the EVIDENCE AND VALUE section of the website or URL link https://maimonresearch.com/evidence-and-value/.

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

## 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