# MAIMON RESEARCH LLC

# ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

# REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

# THE FOOD AND DRUG ADMINISTRATION: THE PRO REGULATORY GUIDANCE AND THE ABSENCE OF RASCH

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

**LOGIT WORKING PAPER No 470 FEBRUARY 2026**

**www.maimonresearch.com**

**Tucson AZ**

# FOREWORD

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

The FDA Patient-Reported Outcome (PRO) Regulatory Guidance, formally issued in 2009 but developed throughout the early 2000s, was concerned with establishing standards for the use of patient-reported outcome instruments to support medical product labeling claims [1]. Its central objective was to ensure that claims based on patient self-reports—such as symptoms, functional status, or health-related quality of life—were supported by scientifically credible evidence. The FDA recognized that patient-reported outcomes could provide direct evidence of treatment benefit, particularly where clinical or laboratory measures failed to capture the patient's experience. However, it also recognized that subjective reports required rigorous development and validation to ensure interpretability, reliability, and regulatory acceptability.

The guidance focused on instrument development, content validity, reliability, construct validity, responsiveness, and interpretability. A key emphasis was content validity, defined as the extent to which an instrument measures the concept of interest in the target patient population. This required qualitative research, including patient interviews, to demonstrate that instrument items reflected relevant and meaningful patient experiences. The FDA also emphasized the need for clear conceptual frameworks, specifying the relationship between items, domains, and the overall construct being assessed. This was intended to avoid ambiguous or multidimensional composite scores that lacked clear interpretability.

Importantly, the guidance addressed the evidentiary standards required for labeling claims. Sponsors were required to predefine endpoints, justify instrument selection, and demonstrate that observed score changes corresponded to meaningful treatment benefit. The concern was not merely statistical significance, but clinical and interpretive credibility. The FDA's framework therefore represented an effort to move subjective outcome claims from informal scoring toward structured, evidence-based measurement, ensuring that PRO-based claims reflected reproducible and interpretable manifestations of treatment impact.

The guidance also reveals a critical omission: it did not engage with the representational theory of measurement. While the FDA emphasized reliability, validity, and responsiveness, it did not address the foundational requirement that measurement must precede arithmetic. Specifically, the guidance did not require demonstration that PRO instruments produced interval or ratio measures rather than ordinal scores. Summation of Likert-type responses was treated as acceptable, despite the fact that such summations generate ordered scores without invariant unit structure. The FDA did not require transformation of subjective observations using Rasch measurement to establish unidimensionality, invariance, and interval scale properties. As a consequence, the resulting PRO endpoints remained scores rather than quantities. This omission had far-reaching implications. By accepting ordinal summations as if they were measures, the FDA effectively legitimized arithmetic operations—such as change scores, group comparisons, and model-based

extrapolations—performed on constructs that lacked lawful measurement properties. This regulatory stance influenced health technology assessment and outcomes research more broadly, reinforcing the use of composite ordinal scores as if they represented measurable quantities. The result was the institutionalization of scoring systems that possess administrative utility but do not satisfy the axioms required for quantitative measurement.

Although more than fifteen years have passed since the release of the FDA's Patient-Reported Outcome (PRO) Regulatory Guidance, the purpose of this study is to apply recently available AI large language model interrogation techniques to assess the extent to which the contemporary FDA knowledge base governing PRO instruments and their application in health technology assessment endorses, excludes, or remains indifferent to the axioms of representational measurement. The analysis employs a structured canonical statement framework to evaluate whether the FDA's operational and methodological position now recognizes the fundamental requirement that measurement must precede arithmetic, and whether subjective observations are transformed using Rasch measurement to establish invariant, unidimensional, interval or ratio scale properties. In doing so, the study determines whether the FDA knowledge base has evolved toward measurement-valid quantitative constructs or continues to rely on ordinal scoring systems that lack the structural properties required for lawful arithmetic and empirically evaluable claims.

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

The modern articulation of this principle can be traced to Stevens' seminal 1946 paper, which introduced the typology of nominal, ordinal, interval, and ratio scales [2]. Stevens made explicit what physicists, engineers, and psychologists already understood: different kinds of numbers permit different kinds of arithmetic. Ordinal scales allow ranking but not addition; interval scales permit addition and subtraction but not multiplication; ratio scales alone support multiplication, division, and the construction of meaningful ratios. Utilities derived from multiattribute preference exercises, such as EQ-5D or HUI, are ordinal preference scores; they do not satisfy the axioms of interval measurement, much less ratio measurement. Yet HTA has, for forty years, treated these utilities as if they were ratio quantities, multiplying them by time to create QALYs and inserting them into models without the slightest recognition that scale properties matter. Stevens' paper should have blocked the development of QALYs and cost-utility analysis entirely. Instead, it was ignored.

The foundational theory that establishes *when* and *whether* a set of numbers can be interpreted as measurements came with the publication of Krantz, Luce, Suppes, and Tversky's *Foundations of Measurement* (1971) [3] . Representational Measurement Theory (RMT) formalized the axioms under which empirical attributes can be mapped to numbers in a way that preserves structure. Measurement, in this framework, is not an act of assigning numbers for convenience, it is the

discovery of a lawful relationship between empirical relations and numerical relations. The axioms of additive conjoint measurement, homogeneity, order, and invariance specify exactly when interval scales exist. RMT demonstrated once and for all that measurement is not optional and not a matter of taste: either the axioms hold and measurement is possible, or the axioms fail and measurement is impossible. Every major construct in HTA, utilities, QALYs, DALYs, ICERs, incremental ratios, preference weights, health-state indices, fails these axioms. They lack unidimensionality; they violate independence; they depend on aggregation of heterogeneous attributes; they collapse under the requirements of additive conjoint measurement. Yet HTA proceeded, decade after decade, without any engagement with these axioms, as if the field had collectively decided that measurement theory applied everywhere except in the evaluation of therapies.

Whereas representational measurement theory articulates the axioms for interval measurement, Georg Rasch's 1960 model provides the only scientific method for transforming ordered categorical responses into interval measures for latent traits [4]. Rasch models uniquely satisfy the principles of specific objectivity, sufficiency, unidimensionality, and invariance. For any construct such as pain, fatigue, depression, mobility, or need, Rasch analysis is the only legitimate means of producing an interval scale from ordinal item responses. Rasch measurement is not an alternative to RMT; it is its operational instantiation. The equivalence of Rasch's axioms and the axioms of representational measurement was demonstrated by Wright, Andrich and others as early as the 1970s. In the latent-trait domain, the very domain where HTA claims to operate; Rasch is the only game in town [5].

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

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

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

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

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

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

**Paul C Langley, Ph.D**

**Email: langleylapaloma@gmail.com**

# 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 FDA PRO KNOWLEDGE BASE

The FDA Patient-Reported Outcome (PRO) knowledge base is defined by the agency's regulatory guidance, supporting methodological documents, qualification program materials, and its broader framework governing clinical outcome assessments (COAs). This knowledge base emerged from the FDA's recognition that patient-reported symptoms, functional status, and treatment experience represent important dimensions of therapeutic evaluation that cannot be fully captured through laboratory tests or clinician observation. The agency's 2009 PRO Guidance established the evidentiary and procedural standards required for PRO instruments to support labeling claims, emphasizing the need for clear conceptual frameworks, defined target populations, appropriate item content, and evidence of reliability, validity, and responsiveness. The guidance positioned PRO instruments as tools capable of capturing treatment benefit directly from the patient's perspective, provided they were developed and evaluated according to systematic methodological principles.

Central to this knowledge base is the concept of the PRO instrument as a structured questionnaire composed of multiple items intended to capture a defined construct such as symptom severity, physical functioning, or health-related quality of life. Instrument development procedures emphasize qualitative research to establish content validity, including patient interviews and cognitive testing to ensure item relevance and clarity. Quantitative evaluation focuses on psychometric properties such as internal consistency, test-retest reliability, construct validity, and sensitivity to change. These properties are typically assessed using statistical methods drawn from classical test theory, including correlation analysis, factor analysis, and responsiveness indices. Instruments that demonstrate acceptable performance across these criteria may be considered suitable for use in clinical trials and regulatory submissions.

The FDA knowledge base also incorporates regulatory pathways for formal instrument qualification. Through the Clinical Outcome Assessment Qualification Program, sponsors may submit instruments for review and qualification for defined contexts of use. Qualification signifies that an instrument has been judged acceptable for measuring a specified concept within a specified population and trial setting. This process reinforces the institutional role of PRO instruments as standardized evaluative tools capable of generating quantitative evidence supporting claims regarding treatment benefit. Once qualified, such instruments may be used across multiple development programs, thereby embedding their scoring algorithms and interpretive frameworks within the regulatory environment.

Operationally, PRO instruments function through scoring algorithms that aggregate responses across multiple questionnaire items to produce numerical scores representing the patient's reported status. These scores may be analyzed as endpoints in clinical trials, compared between treatment groups, and used to support claims regarding symptom improvement or functional benefit. The

regulatory framework does not mandate a single class of PRO instrument but provides general methodological expectations applicable across disease areas. As a result, the FDA knowledge base encompasses a wide range of instruments differing in structure, content, and scoring conventions, but sharing a common role as numerical representations of patient-reported outcomes.

Over time, the FDA has expanded its framework beyond PROs to include other clinical outcome assessments, such as clinician-reported and observer-reported outcomes, within an integrated evidentiary structure. Nevertheless, PRO instruments remain a central component of therapeutic evaluation, particularly in conditions where patient experience represents a primary indicator of treatment impact. The knowledge base therefore reflects an institutionalized methodological framework in which structured subjective observations are converted into numerical scores and incorporated into regulatory decision making, clinical evidence evaluation, and the broader assessment of therapeutic effectiveness.

## 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: FOOD AND DRUG ADMINISTRATION

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

It is essential to understand how to interpret these results.

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

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

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

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

**TABLE 1: ITEM STATEMENT, RESPONSE, ENDORSEMENT AND NORMALIZED LOGITS FDA PRO GUIDANCE**

| **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 | 0.20 | -1.25 |
| MULTIPLICATION REQUIRES A RATIO MEASURE | 1 | 0.15 | -1.60 |
| TIME TRADE-OFF PREFERENCES ARE UNIDIMENSIONAL | 0 | 0.55 | +0.20 |
| RATIO MEASURES CAN HAVE NEGATIVE VALUES | 0 | 0.20 | -1.25 |
| EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES | 0 | 0.65 | +0.85 |
| THE QALY IS A RATIO MEASURE | 0 | 0.65 | +0.85 |
| TIME IS A RATIO MEASURE | 1 | 0.90 | +2.10 |
| MEASUREMENT PRECEDES ARITHMETIC | 1 | 0.25 | -1.10 |
| SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES | 0 | 0.65 | +0.85 |
| MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC | 1 | 0.20 | -1.25 |
| THERE ARE ONLY TWO CLASSES OF MEASUREMENT LINEAR RATIO AND RASCH LOGIT RATIO | 1 | 0.10 | -1.95 |
| TRANSFORMING SUBJECTIVE RESPONSES TO INTERVAL MEASUREMENT IS ONLY POSSIBLE WITH RASH RULES | 1 | 0.10 | -1.95 |
| SUMMATION OF LIKERT QUESTION SCORES CREATES A RATIO MEASURE | 0 | 0.65 | +0.85 |
| THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE | 0 | 0.60 | +0.45 |
| CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT | 1 | 0.20 | -1.25 |
| QALYS CAN BE AGGREGATED | 0 | 0.65 | +0.85 |
| NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED | 1 | 0.80 | +1.55 |
| REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS | 0 | 0.60 | +0.45 |
| THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO | 1 | 0.70 | +1.05 |
| THE RASCH LOGIT RATIO SCALE IS THE ONLY BASIS FOR ASSESSING THERAPY IMPACT FOR LATENT TRAITS | 1 | 0.10 | -1.95 |

| | | | |
|---|---|---|---|
| A LINEAR RATIO SCALE FOR MANIFEST CLAIMS CAN ALWAYS BE COMBINED WITH A LOGIT SCALE | 0 | 0.40 | -0.85 |
| THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT | 1 | 0.30 | -0.85 |
| THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT | 1 | 0.10 | -1.95 |

## FDA PRO GUIDANCE: IGNORING REPRESENTATIONAL MEASUREMENT

The probability–logit profile of the FDA Patient-Reported Outcome (PRO) Guidance knowledge base reveals a structured evaluative framework that has achieved procedural sophistication without securing the measurement foundations necessary for quantitative science. The pattern of categorical probabilities and normalized logits demonstrates not a random or inconsistent engagement with measurement principles, but a coherent operational system in which arithmetic operations are routinely performed on constructs whose measurement properties are not established. This condition represents a structural inversion of representational measurement requirements, in which numerical manipulation precedes and substitutes for measurement validation.

The first and most fundamental observation concerns the absence of endorsement for the axioms that define measurement itself. Statements asserting that measures must be unidimensional, that multiplication requires a ratio measure, and that measurement must precede arithmetic all register negative logits between −0.85 and −1.60. These values indicate that the FDA PRO knowledge base does not structurally recognize these principles as operational constraints. This absence is decisive because measurement is not defined by the existence of numbers, but by the existence of invariant unit structure that permits lawful arithmetic operations. Without unidimensionality, there is no single attribute being quantified. Without ratio scale properties, multiplication and division lack empirical meaning. Without measurement preceding arithmetic, numerical operations become formal manipulations disconnected from empirical quantities.

This inversion is not an abstract philosophical concern. It is embedded directly within the operational logic of PRO instrument use. The FDA guidance framework allows subjective responses to questionnaire items to be aggregated into numerical scores and analyzed as quantitative endpoints in clinical trials. These scores are treated as if they represented measurable quantities capable of supporting arithmetic comparison, statistical analysis, and regulatory interpretation. Yet the logit profile demonstrates that the knowledge base does not recognize the representational measurement axioms required to justify such treatment. The negative logit of −1.10 for the statement that measurement must precede arithmetic captures this inversion

precisely. Arithmetic is performed first, and measurement validity is assumed rather than demonstrated.

This structural inversion is further reinforced by the strong negative logits associated with Rasch measurement principles. Statements asserting that transforming subjective responses into interval measures is only possible with Rasch rules, that Rasch logit ratio scales provide the only basis for assessing therapy impact for latent traits, and that Rasch measurement rules are identical to the axioms of representational measurement all collapse to −1.95. These values indicate effective exclusion of Rasch measurement as an operational framework. This exclusion is particularly consequential because Rasch measurement provides the only lawful method for transforming ordinal subjective observations into invariant interval or ratio-compatible measures. Without such transformation, summated questionnaire scores remain ordinal rankings rather than quantitative measures.

The absence of Rasch measurement within the FDA PRO knowledge base means that subjective responses are aggregated using scoring algorithms that preserve ordinal structure but do not establish quantitative unit structure. The resulting scores represent ordered categories rather than measured quantities. Differences between scores do not represent invariant differences in the underlying attribute. They represent differences in scoring outcomes determined by item selection and response patterns. Yet these scores are used as endpoints in clinical trials and interpreted as indicators of treatment impact.

The logit profile demonstrates that this substitution of scoring for measurement is not accidental. Statements rejecting the proposition that summations of subjective instrument responses constitute ratio measures receive positive logits of +0.85, indicating that the knowledge base partially recognizes that summated scores lack ratio scale properties. However, this recognition does not prevent the operational use of such scores in arithmetic analysis. The framework acknowledges the absence of ratio properties but proceeds with arithmetic operations regardless. This contradiction defines the structural condition of the knowledge base: awareness without operational correction.

The asymmetry between manifest and latent constructs further clarifies the nature of this inversion. Time is correctly recognized as a ratio measure, receiving a strong positive logit of +2.10. This demonstrates that the knowledge base possesses the conceptual capacity to recognize lawful measurement structures when dealing with manifest attributes. Time possesses a true zero and invariant unit structure, permitting multiplication, division, and ratio comparison. The positive logit confirms that this principle is fully internalized within the knowledge base.

However, this recognition does not extend to latent constructs such as patient-reported symptoms or functional status. Statements asserting that latent traits require Rasch transformation for measurement collapse to strong negative logits. This asymmetry reveals that the knowledge base applies measurement discipline selectively. Manifest attributes are recognized as requiring ratio scale structure, while latent attributes are treated as if scoring procedures alone suffice. This asymmetry cannot be justified scientifically. Arithmetic admissibility is determined by scale properties, not by the manifest or latent nature of the attribute. Latent constructs require measurement discipline precisely because they are not directly observable. The failure to apply

Rasch transformation means that latent trait scores remain ordinal classifications rather than quantitative measures.

The logit profile also reveals systematic endorsement of composite constructs that violate dimensional homogeneity. Statements rejecting the dimensional homogeneity of the QALY and rejecting its status as a ratio measure receive positive logits, indicating partial recognition of its structural limitations. Yet statements asserting that QALYs can be aggregated also receive positive logits. This pattern demonstrates that the knowledge base recognizes certain structural limitations while continuing to operate within the composite utility framework. The contradiction reflects institutional stabilization rather than measurement validation. Composite constructs persist because they serve administrative and evaluative functions, not because they satisfy measurement axioms.

The positive logit of +1.55 for the statement that non-falsifiable claims should be rejected provides an important contrast. This indicates that the knowledge base recognizes the importance of falsifiability as a principle of scientific inference. However, this recognition does not extend to the operational constructs used within the PRO framework. Summated ordinal scores and composite utility indices cannot generate falsifiable quantitative claims because they do not represent invariant measured quantities. Differences in scores may reflect changes in item response patterns rather than changes in underlying attribute magnitude. Without invariant measurement structure, empirical falsification is not possible.

The positive logit of +1.05 for the statement defining the logit as the natural logarithm of the odds ratio demonstrates partial conceptual familiarity with the mathematical structure of Rasch measurement. However, this familiarity does not translate into operational adoption. Rasch transformation remains absent from the evaluative framework. This separation between conceptual awareness and operational implementation reflects institutional inertia rather than conceptual ignorance. The knowledge base recognizes elements of measurement theory but does not integrate them into its evaluative architecture.

The consequences of this structural inversion extend beyond methodological consistency to the scientific status of PRO-based claims. Measurement enables quantitative comparison, replication, and cumulative knowledge development. Without measurement, numerical outputs cannot function as empirical quantities. They function as administrative numbers generated by scoring algorithms. These numbers possess internal consistency within the scoring framework but lack invariant correspondence with empirical attributes.

The FDA PRO Guidance knowledge base therefore operates within a coherent administrative framework that produces numerical outputs without establishing measurement validity. This distinction between administrative coherence and measurement validity is critical. Administrative coherence ensures consistency of procedures, reproducibility of scoring, and transparency of interpretation. Measurement validity ensures that numerical outputs correspond to invariant empirical quantities. The logit profile demonstrates that the knowledge base has achieved the former without securing the latter.

This condition reflects the historical evolution of PRO methodology. Classical test theory and psychometric validation procedures emphasize reliability, internal consistency, and responsiveness. These criteria ensure that instruments produce consistent and interpretable scores. However, they do not establish measurement structure. Reliability does not create measurement. Validity in the psychometric sense does not guarantee representational validity. Without Rasch transformation, ordinal responses remain ordinal regardless of statistical refinement.

The persistence of ordinal scoring frameworks within the FDA PRO knowledge base reflects institutional stabilization of methodological conventions. Once scoring algorithms and validation procedures are established, they acquire regulatory legitimacy. Their outputs are incorporated into clinical trials, regulatory submissions, and labeling claims. This institutional embedding reinforces their continued use. The resulting framework becomes self-stabilizing, independent of measurement validity.

The logit profile demonstrates that this stabilization occurs despite partial recognition of measurement principles. The knowledge base acknowledges certain structural requirements while excluding the transformation necessary to satisfy them. This selective recognition reflects the historical trajectory of psychometric methodology, which prioritized statistical reliability over measurement structure.

The implications for the evolution of objective knowledge are profound. Scientific progress depends on measurement because measurement enables falsification and replication. Without measurement, numerical claims cannot be empirically evaluated. They cannot be confirmed or refuted in the quantitative sense. They can only be recalculated within the scoring framework. This condition limits the capacity of PRO-based claims to contribute to cumulative scientific knowledge.

The FDA PRO Guidance knowledge base therefore represents a structurally coherent regulatory framework that has achieved procedural rigor without establishing measurement validity. The logit evidence demonstrates systematic exclusion of the axioms required for representational measurement. Arithmetic operations are performed on constructs lacking ratio scale properties. Subjective responses are scored rather than measured. Latent traits are represented numerically without invariant unit structure.

This condition does not reflect technical incompetence. It reflects conceptual omission. Measurement theory defines the conditions under which numbers represent empirical quantities. Without satisfying these conditions, numerical outputs remain formal representations rather than quantitative measures. The logit profile confirms that the FDA PRO knowledge base operates within this structural condition. Its numerical outputs possess regulatory authority and administrative utility. They support labeling claims, clinical trial evaluation, and regulatory decision making. Yet their measurement foundations remain absent.

This distinction defines the present epistemic status of PRO-based evaluation. Numerical sophistication has been achieved without measurement discipline. Arithmetic operations are performed on ordinal scores lacking invariant unit structure. Composite constructs are treated as quantitative measures without satisfying representational axioms.

The recovery of measurement validity would require structural reconstruction of the evaluative framework. Subjective responses would need to be transformed using Rasch measurement to establish invariant logit ratio scales. Arithmetic operations would need to be restricted to constructs satisfying ratio scale requirements. Only within such a framework could PRO-based claims achieve the status of quantitative measurement.

The logit profile demonstrates that this reconstruction has not occurred. The FDA PRO Guidance knowledge base continues to operate within an evaluative architecture that institutionalizes scoring without measurement. Its numerical outputs remain administratively coherent but measurement-invalid.

This condition defines the structural limitation of the current PRO framework. It represents not a temporary methodological gap but an operational inversion of representational measurement requirements. Until measurement axioms become operational constraints rather than conceptual abstractions, numerical outputs generated within the framework will remain administrative scores rather than empirical quantities.

## THE CONSEQUENCES OF IGNORING RASCH

Ignoring Rasch measurement has had consequences that extend far beyond methodological preference. It has actively encouraged false claims, institutionalized epistemic error, and compromised the duty of care owed to patients, clinicians, and health systems. What began as a technical omission has matured into a systemic failure with real-world harms.

The most immediate consequence is the normalization of false quantitative claims. When ordinal responses are summed and treated as measures, numerical change is mistaken for empirical change. Improvements are declared where none can be demonstrated, differences are asserted without invariant units, and comparative claims proliferate without any lawful basis for comparison. Rasch would have stopped this at the gate. By requiring unidimensionality, invariance, and equal intervals, it would have forced developers to discard instruments that do not measure anything. Its absence allowed virtually any questionnaire to become a “scale,” and any score to become “evidence.”

This directly enabled the expansion of illusory treatment effects. PRO-based endpoints, ungrounded in measurement, routinely support claims of benefit that cannot be falsified. Once embedded in regulatory labels, these claims are recycled in HTA submissions, economic models, and promotional narratives. The appearance of precision masks the absence of substance. Rasch would have made many of these claims impossible; not by being conservative, but by being honest.

The damage is magnified in health technology assessment, where these non-measures are subjected to arithmetic operations that presuppose ratio properties. Utilities derived from ordinal responses are multiplied by time, discounted, aggregated, and compared across interventions. Each step compounds the original error. Without Rasch, there is no way to establish whether the underlying construct is even measurable, let alone suitable for arithmetic. The result is a cascade of numerical artifacts presented as decision-relevant quantities.

This has profound implications for pricing and access decisions. When false measures drive cost-effectiveness claims, resources are allocated on the basis of imaginary quantities. Treatments may be rejected not because they fail to help patients, but because they fail against a denominator that has no empirical meaning. Conversely, interventions may be rewarded for "value" that exists only on paper. Rasch would not guarantee better decisions; but it would guarantee that decisions are made on quantities that exist.

The failure also undermines the evolution of objective knowledge. Measurement is the prerequisite for learning. Without invariant measures, there is no replication, no accumulation of insight, and no correction of error. Each assessment becomes an isolated event, justified internally and insulated from refutation. Ignoring Rasch ensured that outcome claims could persist indefinitely, regardless of whether they corresponded to reality. The system learned how to decide, but forgot how to know.

Most troubling is the impact on duty of care. Institutions that authorize, publish, and rely on non-measures present themselves as evidence-based while withholding the information necessary to evaluate truth. Patients are told that benefits are quantified when they are not. Clinicians are asked to trust numbers that cannot be interpreted. Policymakers are assured of rigor where none exists. Ignorance here is not benign; it is consequential. A duty of care that tolerates false measurement is no duty at all.

Finally, ignoring Rasch entrenched a culture of epistemic complacency. Once false measurement became routine, questioning it threatened careers, institutions, and identities. Silence was rewarded; critique was marginalized. This cultural inertia explains why the problem persists despite decades of availability of a solution.

Ignoring Rasch did not merely delay progress; it licensed error, amplified false claims, distorted decisions, and eroded ethical responsibility. The damage is not hypothetical. It is written into pricing decisions, access restrictions, and a generation of irrecoverable lost knowledge. From the perspective of HTA. The FDA's position amplified the attraction of the false measurement memeplex and the global promotion of ordinal measurement for subjective observations.

# 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 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] U.S. Food and Drug Administration. *Guidance for Industry: Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims.* Silver Spring, MD: U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research, and Center for Devices and Radiological Health, December 2009.

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

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

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

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