MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

UNITED STATES: FALSE MEASUREMENT IS NOT GOOD RESEARCH PRACTICE - AN EPISTEMIC AUDIT OF ISPOR

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

LOGIT WORKING PAPER No 10 JANUARY 2026

www.maimonresearch.com

Tucson AZ

FOREWORD

HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

This Logit Working Paper series documents a finding as extraordinary as it is uncomfortable: health technology assessment (HTA), across nations, agencies, journals, and decades, has developed as a global system of non-measurement. It speaks the language of numbers, models, utilities, QALYs, "value for money," thresholds, discounting, incremental ratios, extrapolations, and simulations. It demands arithmetic at every turn, multiplication, division, summation, aggregation, discounting, yet it never once established that the quantities to which these operations are applied are measurable. HTA has built a vast evaluative machinery on foundations that do not exist. The probabilities and normalized logits in the country reports that follow provide the empirical confirmation of this claim. They show, with unsettling consistency, that the global HTA.

The objective of this study is to interrogate the epistemic foundations of the Good Research Practice (GRP) guidance promoted by the International Society for Pharmacoeconomics and Outcomes Research (ISPOR). Rather than treating these guides as neutral methodological aids or consensus statements, the analysis treats them as formal expressions of an underlying belief system about measurement, arithmetic, evidence, and scientific legitimacy in health technology assessment. Using a 24-item diagnostic grounded in representational measurement theory, the study evaluates whether the quantitative objects explicitly endorsed, standardized, and disseminated through ISPOR GRP's, utilities, QALYs, ICERs, composite patient-reported outcomes, and reference-case simulation outputs, satisfy the axioms required for admissible arithmetic, falsification, and the evolution of objective knowledge. The focus is not on individual guideline documents in isolation, but on the coherent conceptual architecture that emerges across ISPOR task force reports, methodological guides, and best-practice statements over time.

The findings are unequivocal and extreme. The ISPOR GRP knowledge base exhibits a near-total inversion of representational measurement theory. Core axiom, measurement preceding arithmetic, the necessity of ratio scales for multiplication, unidimensionality as a prerequisite for quantitative claims, and the inadmissibility of composite constructs such as QALYs are either decisively rejected or relegated to conceptual irrelevance. At the same time, propositions that are mathematically impossible under those axioms are endorsed at near-ceiling levels. Rasch measurement, the only framework capable of legitimizing latent-trait claims derived from patient-reported outcomes, is categorically excluded. The resulting logit profile is not one of partial misunderstanding or disciplinary ambiguity. It is a signature pattern of institutionalized arithmetic without measurement, codified as global professional guidance and exported as "good research practice."

This assessment is significant because ISPOR's GRP guides function as a global standard-setting mechanism. They are cited by national HTA agencies, embedded in journal editorial policies, referenced in payer submissions, and taught in graduate and professional curricula worldwide. By interrogating these guides directly, this study does not challenge peripheral practice; it challenges the central legitimizing authority of contemporary HTA methodology. If the quantitative

constructs endorsed by ISPOR fail the axioms of representational measurement, then alignment with ISPOR guidance does not represent methodological rigor, consensus, or scientific maturity. It represents coordinated adherence to false measurement. Denying the scientific admissibility of ISPOR's GRP guides has consequences far beyond the society itself: it undermines the epistemic authority of HTA decisions that rely on those guides for validation, justification, and international comparability.

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 ISPOR GOOD PRACTICE KNOWLEDGE BASE

For the purposes of this analysis, the ISPOR GRP knowledge base is defined as the shared and repeatedly reinforced body of methodological assumptions, quantitative conventions, and evaluative norms articulated through ISPOR task force reports, GRP guidelines, educational materials, and conference outputs. It is not defined by any single document or committee, but by the stable patterns that recur across ISPOR’s guides on cost-effectiveness analysis, cost-utility analysis, modeling best practices, patient-reported outcomes, value frameworks, and health economic evaluation more broadly.

This GRP knowledge base is characterized by the routine treatment of preference-based utilities as quantitative measures, the uncritical acceptance of QALYs as ratio-scaled outcomes, and the normalization of incremental cost-effectiveness ratios as legitimate decision variables. Composite constructs formed by summation or indexation of ordinal responses are treated as if they possess interval or ratio properties, despite the absence of demonstrated unidimensionality, invariance, or true zero points. These assumptions are not defended explicitly; they are presupposed as methodological givens. Measurement theory is conspicuous by its absence, both in explicit discussion and in implicit constraint.

Equally central to the ISPOR GRP knowledge base is the reliance on reference-case simulation modeling as a surrogate for empirical evaluation. Model outputs are routinely treated as evidence-bearing claims rather than as conditional projections dependent on assumptions. Sensitivity analysis is presented as a form of robustness or validation, effectively substituting internal consistency for falsifiability. In this framework, claims are stabilized through repetition and consensus rather than exposed to empirical risk. The appearance of scientific discipline is preserved while the core requirement of testability is abandoned.

The handling of patient-reported outcomes is particularly revealing. ISPOR guides consistently endorses the use of PRO instruments while rejecting, implicitly and explicitly, the only measurement model capable of transforming ordinal responses into invariant quantitative scales: Rasch measurement. Instead, classical test theory scoring, summation, and thresholding are treated as sufficient. This allows subjective responses to be incorporated into economic models without ever being measured in the scientific sense. Patient experience becomes numerically exploitable without being quantitatively meaningful.

In this sense, the ISPOR GRP knowledge base is defined as much by its silences as by its prescriptions. There is no sustained engagement with representational measurement theory, no acknowledgment of scale-type constraints on arithmetic, and no recognition that falsification requires empirically measurable quantities. What emerges instead is a closed methodological ecosystem in which arithmetic operations are licensed by convention, consensus, and institutional authority rather than by measurement validity.

The 24-item diagnostic therefore captures the epistemic architecture of ISPOR as it actually operates. It shows a knowledge base that is internally coherent, globally influential, and fundamentally incompatible with the requirements of scientific measurement. Far from being a neutral arbiter of good practice, ISPOR functions as a central node in a self-reinforcing memeplex that stabilizes and propagates false measurement across health technology assessment worldwide.

.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: ISPOR GOOD RESEARCH PRACTICE

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

It is essential to understand how to interpret these results.

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

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

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

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

TABLE 1: ITEM STATEMENT, RESPONSE, ENDORSEMENT AND NORMALIZED LOGITS ISPOR GOOD RESEARH PRACTICE

STATEMENT	RESPONSE 1=TRUE 0=FALSE	ENDORSEMENT OF RESPONSE CATEGORICAL PROBABILITY	NORMALIZED LOGIT (IN RANGE +/- 2.50)
INTERVAL MEASURES LACK A TRUE ZERO	1	0.15	-1.75
MEASURES MUST BE UNIDIMENSIONAL	1	0.15	-1.75
MULTIPLICATION REQUIRES A RATIO MEASURE	1	0.10	-2.20
TIME TRADE-OFF PREFERENCES ARE UNIDIMENSIONAL	0	0.85	+1.75

RATIO MEASURES CAN HAVE NEGATIVE VALUES	0	0.90	+2.20
EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES	0	0.90	+2.20
THE QALY IS A RATIO MEASURE	0	0.95	+2.50
TIME IS A RATIO MEASURE	1	0.95	+2.50
MEASUREMENT PRECEDES ARITHMETIC	1	0.10	-2.20
SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES	0	0.85	+1.75
MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC	1	0.10	-2.20
THERE ARE ONLY TWO CLASSES OF MEASUREMENT LINEAR RATIO AND RASCH LOGIT RATIO	1	0.05	-2.50
TRANSFORMING SUBJECTIVE RESPONSES TO INTERVAL MEASUREMENT IS ONLY POSSIBLE WITH RASH RULES	1	0.05	-2.50
SUMMATION OF LIKERT QUESTION SCORES CREATES A RATIO MEASURE	0	0.90	+2.20
THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE	0	0.85	+1.75
CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.15	-1.75
QALYS CAN BE AGGREGATED	0	0.95	+2.50
NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.70	+0.85
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.90	+2.20
THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.65	+0.60
THE RASCH LOGIT RATIO SCALE IS THE ONLY BASIS FOR ASSESSING THERAPY IMPACT FOR LATENT TRAITS	1	0.05	-2.50
A LINEAR RATIO SCALE FOR MANIFEST CLAIMS CAN ALWAYS BE COMBINED WITH A LOGIT SCALE	0	0.65	+0.60
THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT	1	0.15	-1.75

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

GOOD RESEARCH PRACTICE WITHOUT MEASUREMENT: A LOGIT-BASED ASSESSMENT OF ISPOR GUIDANCE

The various GRP guides present themselves as the methodological conscience of health technology assessment. They are framed as neutral, technical documents whose purpose is to improve rigor, transparency, and consistency across studies. Their authority derives not from statute but from collective authorship: long lists of senior academics, journal editors, consultants, and HTA insiders whose reputations confer legitimacy on what is being recommended. This is precisely what makes the GRP corpus so revealing. When examined as a whole, the guidelines do not function as correctives to bad practice. They function as a memeplex: a self-reinforcing belief system that normalizes arithmetic without measurement and stabilizes false quantification as "good research practice." Normal science would reject them on measurement grounds.

A memeplex, in Dawkins' sense, is not a single false idea but a mutually supporting cluster of ideas that propagate together because they protect one another from challenge. ISPOR's GRP documents are a textbook example. They embed, repeat, and cross-cite the same assumptions: that utilities are quantitative measures; that QALYs are admissible ratio constructs; that summated ordinal scores can be treated as interval or ratio scales; that reference-case simulation outputs constitute evidence; and that sensitivity analysis substitutes for falsification. Each GRP guide reinforces the others, and each is signed off by overlapping author networks. The result is not methodological pluralism but epistemic closure.

What is striking is not that individual authors make errors. It is that across decades of GRP production, no guideline ever pauses to ask the prior question: are the quantities being manipulated actually measures? Representational measurement theory is absent. Scale-type requirements are ignored. The axioms governing addition, multiplication, and aggregation are never treated as constraints. Arithmetic is taken as given; measurement is treated as optional background. This inversion is not accidental. It is the enabling condition for the entire HTA enterprise as currently constituted.

Collective authorship plays a central role in enforcing this inversion. GRP documents are produced by committees, task forces, and working groups whose membership overlaps heavily with journal editorial boards, HTA agencies, consultancy firms, and academic training programs. Authors rotate across guidelines, journals, and policy documents, creating a dense citation and endorsement network. This has two effects. First, it creates the appearance of consensus: when dozens of senior figures sign off on a guideline, disagreement appears eccentric or unserious. Second, it suppresses internal critique: challenging a GRP assumption is not challenging an idea, but an entire professional community. The cost of dissent is exclusion.

This explains why the same false propositions recur unchanged. Consider utilities. Across GRP documents, utilities are treated as quantitative entities suitable for arithmetic operations, despite being derived from ordinal preference elicitation exercises. No GRP guideline asks whether these preferences satisfy unidimensionality, invariance, or interval properties. No guideline addresses the absence of a true zero. Instead, utilities are simply assumed to be "values" that can be multiplied by time. The QALY inherits this assumption wholesale. Once utilities are treated as measures, the QALY becomes inevitable. Once the QALY is accepted, aggregation across individuals becomes permissible. Each step depends on the previous one, and none is interrogated.

The same pattern appears with patient-reported outcomes. GRP documents routinely discuss instrument selection, responsiveness, and minimal important differences, but never measurement. Summated Likert scores are treated as quantitative endpoints by default. Rasch measurement, the only framework capable of transforming ordinal responses into invariant measures, is conspicuously absent. Its absence is not due to ignorance. Rasch has been available, well-documented, and widely applied in other fields for decades. Its exclusion is functional. Adopting Rasch would force HTA to confront unidimensionality, item invariance, and scale validity. It would expose most existing PRO instruments as non-measures. The GRP memeplex cannot survive that exposure.

Reference-case modeling is the keystone of the system. GRP guidelines devote enormous attention to model structure, parameter uncertainty, scenario analysis, and transparency. What they never address is falsifiability. Models are treated as if they generate evidence about the future, rather than conditional projections contingent on assumptions. Sensitivity analysis is elevated to a surrogate for empirical testing. If results are "robust" across scenarios, they are treated as credible. This is not falsification; it is internal consistency checking. Yet the GRP corpus repeatedly conflates the two, allowing models to acquire epistemic authority without ever being exposed to the possibility of being wrong.

The cumulative effect of these practices is the normalization of false measurement. Arithmetic is performed on non-measures, and the outputs are treated as evidence. Because this is done collectively, repeatedly, and authoritatively, it becomes invisible. Students are trained to accept ICERs, QALYs, and modeled value thresholds as the natural objects of HTA. Journals publish them without question. Agencies demand them as inputs. The memeplex reproduces itself through education, publication, and policy.

What is especially damaging is that this system actively blocks the evolution of objective knowledge. In normal science, claims are formulated in ways that allow them to be falsified. Measurements are refined. Instruments improve. Errors are exposed and corrected. In HTA under the GRP regime, this process cannot occur. There is no invariant quantity to reproduce, no measurement to refine, no hypothesis to falsify. Disagreements are resolved through consensus processes, methodological checklists, or stakeholder negotiation. Knowledge does not evolve; it accretes.

ISPOR often presents its GRP guidelines as living documents that evolve over time. In practice, what evolves is technique, not epistemology. Models become more complex. Sensitivity analyses become more elaborate. Reporting standards become more detailed. The underlying measurement

failures remain untouched. Indeed, complexity serves as camouflage. As models become harder to understand, foundational questions recede further from view. The memeplex becomes more resilient.

The collective nature of authorship is critical here. No single author is responsible for the failure. Responsibility is diffused across committees and years. This diffusion provides insulation. When critique arises, it can be dismissed as philosophical, impractical, or outside scope. The GRP documents define their own scope, and that scope excludes measurement axioms by design. What is excluded cannot be debated.

This also explains why ISPOR has not been seriously challenged before. Individual critics could be marginalized. Journals aligned with the GRP worldview would not publish fundamental critiques. Training programs would not teach them. Without a way to systematically interrogate the belief system as a whole, the memeplex remained invisible. Large language models change this dynamic. They can ingest and synthesize decades of text, identify recurring assumptions, and expose structural patterns that are not obvious from any single document. What appears here as a 24-item diagnostic is not a survey of opinions, but a mirror held up to the collective discourse.

The implications are severe. ISPOR's GRP guidelines do not merely fail to promote good research practice. They actively redefine "good practice" in ways that exclude science. By normalizing arithmetic without measurement, they license pricing decisions, access restrictions, and resource allocations based on quantities that cannot support the operations performed on them. Patients experience real consequences from numbers that have no empirical meaning.

This is not a call for incremental reform. Adding a paragraph on limitations or expanding sensitivity analysis does nothing to address the core problem. The problem is axiomatic. Measurement must precede arithmetic. Unidimensionality must be demonstrated, not assumed. Ratio operations require ratio scales. Latent traits require Rasch measurement if they are to be quantified at all. Until these principles are adopted, no amount of methodological polish can rescue HTA from pseudoscience and irrelevance.

The GRP memeplex persists because it is socially efficient. It allows large numbers of professionals to coordinate around shared practices without confronting uncomfortable truths. It creates careers, journals, consultancies, and institutions. Challenging it threatens not just ideas, but livelihoods and identities. That is precisely why the challenge is necessary.

ISPOR stands at a crossroads. It can continue to curate and propagate a belief system that confuses arithmetic with measurement and consensus with knowledge. Or it can confront the implications of representational measurement theory and rebuild HTA on scientific foundations. The first path offers comfort and continuity. The second offers intellectual honesty and the possibility of progress. The GRP corpus, as it stands, has chosen the first. The diagnostic makes that choice visible.

Calling this out is not hostility; it is the minimum requirement of science. False measurement is not good research practice. When a professional society institutionalizes it through collective authorship and guideline production, the failure is no longer individual or accidental. It is systemic.

And systems, unlike individuals, can only change when their underlying belief structures are exposed and dismantled.

MAPPING AS THE FINAL ADMISSION: ISPOR'S ABANDONMENT OF MEASUREMENT

It would be possible to deconstruct every ISPOR Good Research Practice (GRP) guideline individually. Such an exercise, however, would offer diminishing analytical returns. The same epistemic structure appears throughout: the normalization of ordinal scoring as measurement, the detachment of arithmetic from scale-type constraints, and the substitution of model stability for falsification. Rather than rehearse this pattern repeatedly, this paper concentrates on a single GRP recommendation where the underlying belief system becomes fully visible.

The focus is the ISPOR guidance on mapping, examined through the editorial endorsement published in *Value in Health* [5]. Mapping is not a peripheral methodological issue. It is the clearest expression of the HTA memeplex in operation. It explicitly acknowledges that measurable utility data are absent and then proceeds to recommend statistical reconstruction as an acceptable substitute. In doing so, mapping transforms the absence of measurement from a scientific limit into a technical inconvenience. For that reason, the mapping guidance provides a uniquely powerful lens through which to deconstruct not merely one GRP document, but the intellectual logic sustaining the entire ISPOR framework.

The ISPOR mapping editorial is not a marginal methodological commentary. It is a rare moment of epistemic clarity, not because it defends its position rigorously, but because it exposes the belief system that underpins modern HTA in its most unguarded form. Mapping is presented not as a provisional workaround, but as an accepted and necessary component of cost-utility analysis. That alone makes it a uniquely powerful object for deconstruction. Mapping exists for one reason only: clinical studies frequently do not include preference-based instruments required to generate utilities. Under any legitimate theory of measurement, the absence of measurement would terminate the analysis. Instead, ISPOR proposes substitution. Utilities may be estimated after the fact, using statistical relationships derived from external data. This move is not methodological sophistication; it is a categorical violation of representational measurement. Measurement cannot be reconstructed retrospectively through regression. A quantity that was not measured cannot later be inferred into existence. Mapping therefore does not extend measurement; it replaces it with prediction. That substitution marks the precise point at which science gives way to ritual.

The editorial's definition of health utility exposes the depth of the problem. Utilities are described as values on a cardinal numeric scale, anchored at one for full health, zero for death, and extending into negative territory for states worse than dead. This description is offered without any attempt to justify the existence of equal intervals, a true zero, or invariance across persons and contexts. These properties are not minor technicalities. They are the axioms that permit arithmetic. Without them, multiplication by time is meaningless. The editorial simply asserts cardinality and proceeds. This is not ignorance; it is doctrinal assumption. By defining utilities as ratio-like quantities without establishing the empirical operations that could produce such scales, the guidance reveals its foundational posture: arithmetic is treated as self-justifying. Numbers are assumed to measure because decisions demand numbers. The logical direction of science is reversed.

The mapping framework then reframes the entire problem as one of statistical craftsmanship. The editorial devotes extensive attention to model choice, distributional form, goodness-of-fit, predictive accuracy, and uncertainty characterization. Yet nowhere is there any engagement with scale type. Nowhere is the question asked whether the dependent variable possesses the properties required to support the arithmetic that follows. Statistical association is treated as equivalent to measurement. This is the critical substitution. Regression can predict outcomes, but it cannot generate measurement units. A function that predicts ordinal responses does not transform them into interval or ratio scales. No improvement in model performance can overcome that barrier. By redefining measurement as prediction, ISPOR quietly abandons representational measurement theory altogether, replacing it with a purely instrumental conception of numbers.

The editorial also concedes, without recognizing the implication, that different mapping approaches generate different utility values, which in turn generate different cost-effectiveness results. From a scientific standpoint, this admission should have been catastrophic. If multiple methods yield different numerical representations of the same health state, then no invariant quantity is being measured. Variation is not noise around a true value; it is evidence that no true value exists. Instead of confronting this implication, the guidance reframes divergence as a technical challenge requiring careful method selection. The possibility that the enterprise itself is invalid is never entertained. The instability of results is normalized. This is precisely how a memeplex protects itself: anomalies are absorbed as methodological complexity rather than recognized as falsification.

The language of "good research practice" performs essential ideological work in this process. The task force does not ask whether mapping is permissible under measurement theory. It assumes permissibility and restricts debate to how mapping should be conducted. This is the defining characteristic of a mature memeplex. Once a belief becomes non-negotiable, intellectual variation is permitted only within boundaries that preserve the replicator. Here, the replicator is the QALY. Mapping exists to ensure the survival of the QALY when empirical reality fails to supply the required inputs. Rather than questioning the framework, ISPOR expands it. Rather than restricting claims, it invents procedures to maintain them. The appearance of methodological progress masks a deepening epistemic retreat.

From the perspective articulated in the *Logit Working Papers*, mapping violates both admissible pathways to measurement. It is not a linear ratio measure of a manifest attribute, such as time, events, or resource counts. Nor is it a Rasch logit ratio measure of a latent trait derived from invariant item functioning. It occupies an impossible middle ground: ordinal responses are statistically projected onto another ordinal system, which is then algorithmically converted into a pseudo-continuous output treated as ratio-scaled. At no stage are the axioms of measurement satisfied. Mapping therefore does not represent weak measurement or imperfect measurement. It represents non-measurement masquerading as precision. It institutionalizes the belief that if arithmetic can be performed, then measurement must have occurred.

This has profound implications for falsification and the evolution of objective knowledge. Because mapped utilities are model-dependent constructions, they cannot be empirically refuted. Sensitivity analysis does not test claims against reality; it explores alternative assumptions within a closed system. Stability across scenarios becomes a surrogate for truth. Replication becomes

repetition using similar conventions. Knowledge does not evolve through confrontation with error; it accumulates through consensus around shared practices. This is not Popperian science. It is scholasticism with software. The editorial's confident assertion that mapping supports decision making disguises the fact that it eliminates the possibility of learning from being wrong.

What makes this editorial so consequential is not its technical detail but its position in the HTA supply chain. ISPOR does not merely advise analysts; it defines legitimacy. When ISPOR labels mapping a good research practice, it authorizes journals to publish it, agencies to accept it, manufacturers to submit it, and reviewers to enforce it. The result is a self-reinforcing ecosystem in which false measurement is not merely tolerated but required. Quality-of-life journals generate ordinal instruments. Mapping converts those instruments into utilities. Economic models convert utilities into QALYs. Pricing bodies convert QALYs into thresholds. At no point is measurement established. Yet each step appears rational because the previous step has been sanctified.

This is why the editorial is more than sufficient for aggressive deconstruction. It demonstrates, in its own words, that ISPOR has replaced the scientific requirement that measurement precede arithmetic with the administrative requirement that arithmetic precede decision making. Mapping is not an unfortunate compromise. It is a doctrinal statement: when measurement fails, invent a procedure that allows calculation to continue. From the standpoint of representational measurement, this is not merely incorrect. It is incoherent. It denies the conditions under which numbers can represent empirical attributes at all.

Mapping does not represent methodological progress in health technology assessment; it represents an admission of failure. The ISPOR editorial does not propose mapping because it improves measurement, but because measurement has already failed. Faced with the impossibility of constructing valid utility measures from patient-reported outcomes, the HTA community resorts to statistical substitution: transferring numbers from one non-measure to another and treating the result as if it were quantitative evidence. Mapping is therefore not a bridge between instruments; it is a mechanism for preserving arithmetic in the absence of measurement. It allows the QALY framework to survive by bypassing the axioms that would otherwise invalidate it. In doing so, the editorial reveals the core logic of the HTA memeplex: when measurement cannot be established, it must be simulated; when simulation cannot be justified, it must be standardized; and when standards fail, they must be declared "good research practice." Mapping is not science advancing, it is belief replication under epistemic stress.

3. THE TRANSITION TO MEASUREMENT IN HEALTH TECHNOLOGY ASSESSMENT

THE IMPERATIVE OF CHANGE

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

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

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

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

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

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

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

MEANINGFUL THERAPY IMPACT CLAIMS

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

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

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

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

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

THE PATH TO MEANINGFUL MEASUREMENT

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

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

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

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

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

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

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

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

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

TRANSITION REQUIRES TRAINING

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

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

Training must therefore begin with scale types and their permissible operations. Linear ratio measurement applies to manifest attributes that possess a true zero and invariant units, such as

time, counts, and resource use. Latent attributes, by contrast, cannot be observed directly and cannot be measured through summation or weighting. They require formal construction through a measurement model capable of producing invariant units. This distinction is the conceptual fulcrum of reform, because it determines which claims are admissible and which are not.

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

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

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

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

A NEW START IN MEASUREMENT FOR HEALTH TECHNOLOGY ASSESSMENT

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

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

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

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

DESIGNED FOR CLOSURE

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

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

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

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

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

them to be posed. The absence of measurement standards was not an oversight; it was functionally essential.

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

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

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

ACKNOWLEDGEMENT

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

REFERENCES

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

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

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

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

[5] Wolowacz S. New ISPOR Recommendations -,Mapping Methods for Estimation of Health State Utility. V*alueHealth.* 2017; 20: 28-29